ADASIGNSGD: TURNING 1-BIT WORKER VOTES INTO ADAPTIVE UPDATES

Anonymous authors
Paper under double-blind review

ABSTRACT

Momentum-sign methods reduce synchronization traffic in distributed AI training by transmitting each worker's local momentum with one bit per coordinate. However, these methods discard local momentum magnitudes and often fail to match the optimization quality of full-precision methods. We introduce ADASIGNSGD, which uses the majority sign to determine the update direction and the vote margin to set the update magnitude, taking larger steps on coordinates with stronger worker agreement. In C4 language-model pretraining, ADASIGNSGD outperforms all evaluated 1-bit worker-sign baselines in validation perplexity across 13 model and training-budget settings from 60M to 1B parameters. Its gap to dense AdamW narrows with model scale. At 1B, ADASIGNSGD closely matches AdamW in validation perplexity while delivering a $1.46\times$ complete-step speedup over fused AdamW on 32 H200 GPUs. ADASIGNSGD uses one model-sized momentum state while incurring only 14.1% of FP32 dense-DDP model-sized traffic at 128 workers. On 128 P100 GPUs, ADASIGNSGD delivers a $4.39\times$ complete-step speedup over AdamW for the 350M model. Together, these results show that one-bit worker votes can be turned into adaptive updates that combine strong optimization quality with a smaller memory footprint and faster distributed training steps at scale.

1 INTRODUCTION

Replicated data-parallel training is simple and widely used, but synchronizing model-sized tensors at every step can become a significant scalability bottleneck as model size and the number of data-parallel workers grow (Li et al., 2020; Jiang et al., 2024). This bottleneck is particularly difficult for adaptive optimization. Methods such as Adam (Kingma & Ba, 2015) use coordinate-wise information to modulate their updates. By contrast, signSGD communicates coordinate-wise stochastic-gradient signs (Bernstein et al., 2018), while its distributed majority-vote variant aggregates worker signs into a single binary update direction (Bernstein et al., 2019). These methods reduce each worker's message to one bit per coordinate, but discard both local magnitudes and the strength of cross-worker agreement. Methods designed to preserve adaptive optimization under compression often pay for it elsewhere. SIGNUM-MV+EF (Karimireddy et al., 2019) and Birder+EC (Peng et al., 2023) maintain persistent error-feedback residuals, while Birder's hierarchy additionally performs a full-precision intra-node Reduce-Scatter at every step. 1-Bit Adam and 0/1 Adam retain multiple optimizer states and use, respectively, an initial full-precision Adam phase or scheduled full-precision variance updates (Tang et al., 2021; Lu et al., 2023). These mechanisms reduce communication, but often shift rather than eliminate systems cost into persistent state, dense communication, or both.

The majority sign contains no magnitude, but the pattern of signs across workers contains more information than its winner. With eight workers, $5{:}3$ and $8{:}0$ positive-to-negative votes select the same direction under majority vote, although the latter provides much stronger evidence for the positive direction. Winner-only aggregation discards this distinction; the exact count retains it. The count must still be converted into a principled coordinate-wise scale and constructed without reintroducing a dense communication path. This leads to our central question: *can one-bit worker votes enable coordinate-wise adaptivity without dense communication or additional persistent model-sized state?*

Our approach. We answer this question with ADASIGNSGD and VOTEREDUCE. For each coordinate, ADASIGNSGD retains the exact number of workers with nonnegative local momentum. The

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.

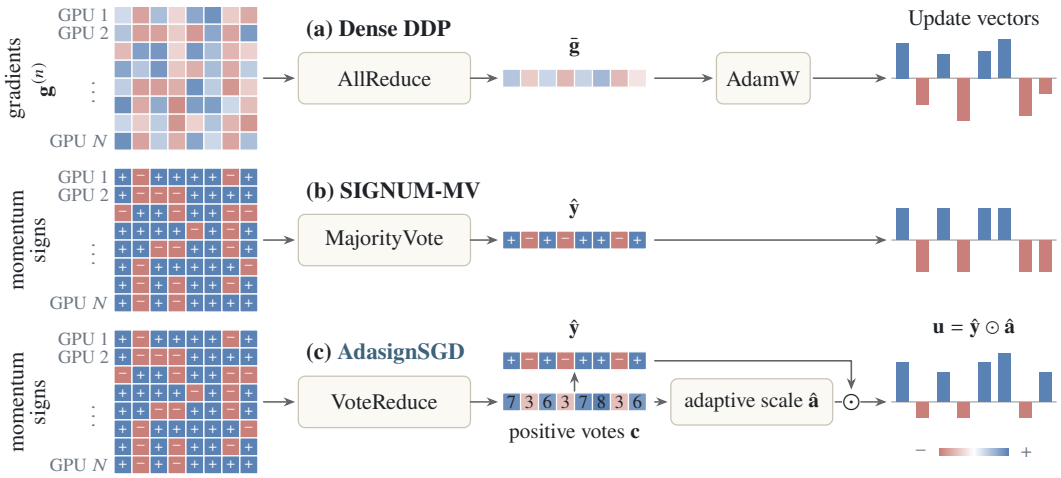

Figure 1: **ADASIGNSGD derives adaptive updates from worker momentum signs.** **(a)** Dense DDP averages the N worker gradients $\mathbf{g}^{(n)}$ to obtain $\bar{\mathbf{g}}$ before applying AdamW. **(b)** SIGNUM-MV aggregates momentum signs into the majority direction $\widehat{\mathbf{y}}$ with fixed update magnitude. **(c)** VOTEREDUCE recovers the per-coordinate counts $\mathbf{c}$ of nonnegative worker momenta. The counts determine $\widehat{\mathbf{y}}$ and the adaptive scale $\widehat{\mathbf{a}}$, yielding update $\mathbf{u} = \widehat{\mathbf{y}} \odot \widehat{\mathbf{a}}$, where $\odot$ denotes elementwise multiplication. Blue/red encode positive/negative values or signs.

count determines the majority direction and the fraction of workers supporting it. Under an idealized homogeneous Gaussian model, this support probability is $\Phi(S)$, where $S = |\mu|/\sigma$ is the momentum signal-to-noise ratio. ADASIGNSGD applies the inverse standard-normal CDF to estimate S and maps it to the bounded RMS-normalized scale $S/\sqrt{1+S^2}$. The resulting update uses majority vote for direction and worker agreement for scale. VOTEREDUCE routes packed sign shards to aggregating workers and redistributes only the resulting exact counts, avoiding full-value worker momenta and a model-sized residual.

Algorithm 1 ADASIGNSGD: complete vote-derived optimizer step

Require: N synchronized workers and $n \in \{1, \dots, N\}$; shared parameters $\boldsymbol{\theta}_t \in \mathbb{R}^d$; local gradient $\mathbf{g}_t^{(n)}$ and EMA momentum $\mathbf{m}_{t-1}^{(n)}$, both in $\mathbb{R}^d$; $\beta \in [0, 1)$ EMA coefficient, $\eta_t \geq 0$ learning rate, $\lambda \geq 0$ decoupled weight decay, $\epsilon_p \in (0, 1/4)$ probability clamp, and $\varepsilon \geq 0$ scale stabilizer

Ensure: Shared updated parameters $\boldsymbol{\theta}_{t+1}$ and updated local EMA momentum $\mathbf{m}_t^{(n)}$

Φ is the standard-normal CDF; $\mathrm{sgn}_+(x) = +1$ for $x \geq 0$ and -1 otherwise.

$\mathrm{clip}(x, \ell, u) = \min\{\max\{x, \ell\}, u\}$; c: nonnegative-vote count.

Initialize: $\mathbf{m}_0^{(n)} \leftarrow \mathbf{0}$ on every worker.

Precompute once: Build lookup table LUT_N (vote margin $\delta = |2c - N|$):

for $\delta = 0, \dots, N$ **do**
  $p \leftarrow \frac{1}{2} + \delta/(2N)$ ▷ fraction supporting the majority
  $S \leftarrow \Phi^{-1}\big(\mathrm{clip}(p, \frac{1}{2} + \epsilon_p, 1 - \epsilon_p)\big)$
  $\mathrm{LUT}_N[\delta] \leftarrow S/\sqrt{(1 + S^2 + \varepsilon)}$ ▷ adaptive scale
end for

At optimization step t**:**

$\mathbf{m}_t^{(n)} \leftarrow \beta \mathbf{m}_{t-1}^{(n)} + (1 - \beta)\mathbf{g}_t^{(n)}$
$\mathbf{b}_t^{(n)} \leftarrow \mathbb{I}[\mathbf{m}_t^{(n)} \geq 0]$ ▷ element-wise; one bit per coordinate
$\mathbf{c}_t \leftarrow \mathrm{VoteReduce}(\mathbf{b}_t^{(n)}) \equiv \sum_{n=1}^{N} \mathbf{b}_t^{(n)}$ ▷ exact nonnegative-vote counts
$\widehat{\mathbf{y}}_t \leftarrow \mathrm{sgn}_+(2\mathbf{c}_t - N)$, $\boldsymbol{\delta}_t \leftarrow |2\mathbf{c}_t - N|$ ▷ direction and vote margin
$\widehat{\mathbf{a}}_t \leftarrow \mathrm{LUT}_N[\boldsymbol{\delta}_t]$ ▷ element-wise scale lookup
$\mathbf{u}_t \leftarrow \widehat{\mathbf{y}}_t \odot \widehat{\mathbf{a}}_t$
$\boldsymbol{\theta}_{t+1} \leftarrow (1 - \eta_t \lambda)\boldsymbol{\theta}_t - \eta_t \mathbf{u}_t$ ▷ decoupled weight decay

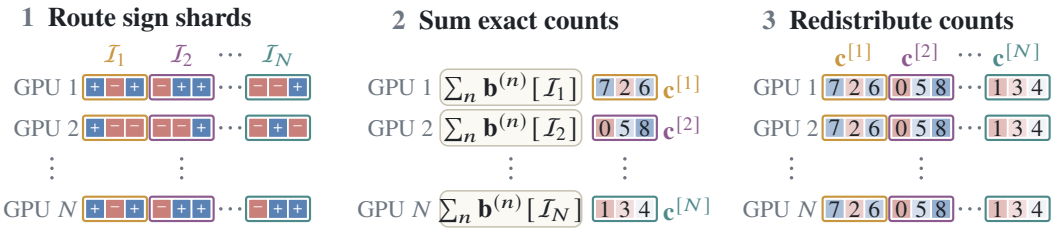

Figure 2: **VOTEREDUCE computes exact vote counts from packed one-bit signs.** The d coordinates are partitioned into shards $\{\mathcal{I}_s\}_{s=1}^{N}$. Each worker n routes $\mathbf{b}^{(n)}[\mathcal{I}_s]$ in packed form to owner s, which computes the per-coordinate counts $\mathbf{c}^{[s]} = \sum_{n=1}^{N} \mathbf{b}^{(n)}[\mathcal{I}_s]$. The count shards are redistributed to all workers, yielding the full vector $\mathbf{c}$.

The paper makes three contributions. First, we derive a nonlinear vote-count-to-scale map and establish finite-worker recovery and optimization guarantees under explicitly stated conditions. Second, we introduce an exact packed-sign collective and characterize its traffic, memory, communication-path, and complete-step costs. Third, we validate the Gaussian prediction relating momentum-sign agreement to SNR using fixed-parameter checkpoints and distributed training trajectories, then evaluate C4 pretraining under method-specific endpoint tuning. Across all 13 model–budget settings from 60M to 1B, ADASIGNSGD achieves the lowest validation perplexity among the evaluated one-bit worker-sign methods; at 1B, it reaches 16.95 versus AdamW's 17.08. The systems evidence spans two complementary regimes. On communication-constrained P100, ADASIGNSGD achieves $3.58\times$–$4.39\times$ complete-step speedups over AdamW at 128 GPUs across 60M–350M models, with matched model, per-worker workload, precision, and timing settings. On H200, intra-node NVLink and four 400-Gb/s NDR InfiniBand rails using GPUDirect RDMA provide a fast dense-communication path (NVIDIA, n.d.; 2023). AdamW further overlaps gradient reductions with backpropagation (Li et al., 2020), leaving less exposed communication for compression to remove. Even in this setting, ADASIGNSGD achieves a $1.46\times$ complete-step speedup for 1B training on 32 GPUs over fused, overlapped AdamW.

2 RELATED WORK

Gradient compression. Early work on quantized gradient communication includes 1-Bit SGD for speech (Seide et al., 2014), QSGD's stochastic multilevel quantizer with convergence guarantees (Alistarh et al., 2017), and TernGrad's stochastic ternary gradients (Wen et al., 2017). Deep Gradient Compression combines sparsification with momentum correction, local clipping, momentum-factor masking, and warmup training (Lin et al., 2018). PowerSGD compresses matrix-shaped gradients through low-rank power iteration and is designed for efficient all-reduce aggregation (Vogels et al., 2019). DeepReduce instead reduces the systems overhead of sparse tensors by separately encoding their values and indices, making it complementary to gradient sparsifiers (Xu et al., 2021).

Deterministic sign aggregation. SignSGD studies optimization with coordinate-wise gradient signs, while its distributed majority-vote variant aggregates worker signs into a single binary direction (Bernstein et al., 2018; 2019). Balles and Hennig connect Adam-style scaling and Gaussian sign reliability to gradient signal-to-noise ratio (Balles & Hennig, 2018). Error feedback accumulates compression residuals and reinjects them into subsequent updates; SIGNUM-MV+EF combines this mechanism with majority-vote SIGNUM (Karimireddy et al., 2019). Distributed Lion communicates binary local Lion updates with majority aggregation (Liu et al., 2024).

Randomized one-bit quantization and error feedback. Birder stochastically quantizes a bounded adaptive update without an uncompressed warmup; its full formulation applies both worker-local and post-aggregation error feedback, which we denote Birder+EC in our experiments (Peng et al., 2023). Its hierarchy combines full-precision intra-node Reduce-Scatter, one-bit inter-node aggregation, and intra-node All-Gather. Error feedback reinjects compression residuals and can restore convergence guarantees for biased sign and sparse compressors (Karimireddy et al., 2019; Stich et al., 2018). LoCo uses a moving average of historical compression errors to support low-bit gradient communication with optimizers such as Adam (Xie et al., 2025).

Adaptive optimization with dense communication phases. 1-Bit Adam first runs full-precision Adam, then freezes the stabilized variance state and applies error-compensated one-bit momentum communication (Tang et al., 2021). 1-Bit LAMB uses a full-precision LAMB warmup before its compressed stage and preserves adaptive layer-wise scaling (Li et al., 2022). In contrast, 0/1 Adam uses adaptive variance freezing, one-bit communication, and local steps within a unified schedule, while still performing full-precision all-reduce on designated variance-update steps (Lu et al., 2023).

3 Motivation: Sign Agreement Reveals Momentum SNR

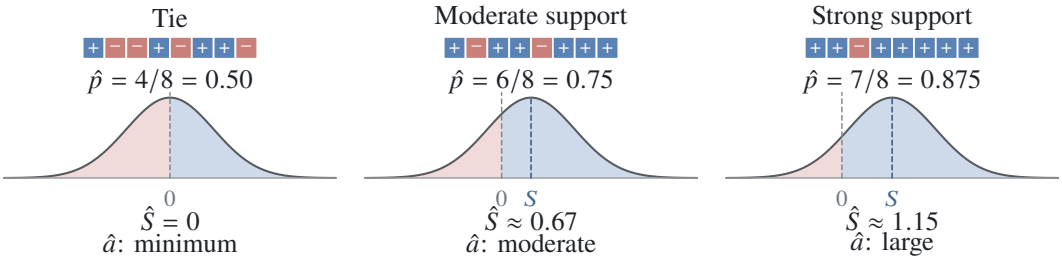

Figure 3: **From vote counts to momentum SNR and adaptive scale.** For $N = 8$, the counts $c = 4, 6, 7$ produce increasing selected-sign support $\widehat{p} = \max(c, N - c)/N$. Under the idealized Gaussian model, $p = \Phi(S)$, so $\widehat{S} = \Phi^{-1}(\widehat{p})$ yields increasing adaptive scales.

EMA momentum is substantially less heavy-tailed than the raw stochastic gradient, making a Gaussian approximation useful for interpreting its sign. For any scalar x, define $\mathrm{sgn}_+(x) = +1$ if $x \geq 0$ and -1 otherwise. At fixed model parameters and for one coordinate, let the scalar momentum m have mean μ and standard deviation $\sigma > 0$, and define its signal-to-noise ratio and directional sign-agreement probability as $S := |\mu|/\sigma$ and $p := \Pr[\mathrm{sgn}_+(m) = \mathrm{sgn}_+(\mu)]$, respectively. If $m \sim \mathcal{N}(\mu, \sigma^2)$, then $p = \Phi(S)$, where Φ is the standard-normal CDF. Hence the agreement probability identifies the momentum SNR through $S = \Phi^{-1}(p)$: a sign carries no magnitude, but sign agreement carries normalized signal-strength information. Figure 3 illustrates this relationship using eight worker votes.

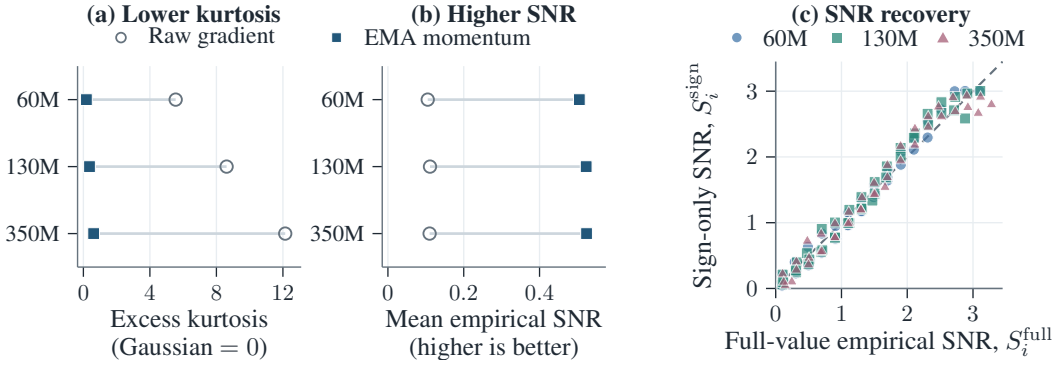

Figure 4: **Momentum-sign agreement estimates coordinate-wise momentum SNR. (a)** EMA momentum has lower excess kurtosis than the corresponding raw gradient across all three model sizes. **(b)** EMA momentum has higher mean empirical SNR. **(c)** The sign-only estimate S_i^{sign} closely tracks the full-value empirical SNR S_i^{full} across model sizes and training stages. Full per-checkpoint results and probability-space calibration are reported in Appendix C.

We test this prediction on 60M, 130M, and 350M LLaMA-style models trained on English C4 (Touvron et al., 2023; Raffel et al., 2020). At frozen model parameters, we run a diagnostic EMA over fresh mini-batches. After a 32-update burn-in, we record $K = 128$ raw-gradient and momentum samples using $\beta = 0.9$. For coordinate i, let $m_{k,i}$ be momentum sample k, and let $\widehat{\mu}_i$ and $\widehat{\sigma}_i$ denote the sample mean and standard deviation of $\{m_{k,i}\}_{k=1}^{K}$. We compare the full-value empirical SNR

with the SNR recovered from empirical sign agreement:

$$S_i^{\text{full}} = \frac{|\widehat{\mu}_i|}{\widehat{\sigma}_i}, \qquad \widehat{p}_i = \frac{1}{K}\sum_{k=1}^{K} \mathbb{I}\big[\operatorname{sgn}_+(m_{k,i}) = \operatorname{sgn}_+(\widehat{\mu}_i)\big], \tag{1}$$

$$S_i^{\text{sign}} = \Phi^{-1}\big(\operatorname{clip}\big(\widehat{p}_i, \tfrac{1}{2} + \epsilon_p, 1 - \epsilon_p\big)\big). \tag{2}$$

Here $\operatorname{clip}(x, \ell, u)$ truncates x to $[\ell, u]$, and $\epsilon_p \in (0, 1/4)$ bounds the inverse CDF. We compute the same statistics for raw gradients by replacing $m_{k,i}$ with the corresponding gradient sample.

First, EMA reduces excess kurtosis by 94.9%–96.7% and increases mean empirical SNR by $4.7\times$–$4.8\times$ relative to raw gradients, placing momentum in a substantially more Gaussian and higher-SNR regime. Second, the parameter-free inverse-CDF map recovers full-value coordinate-wise SNR from sign agreement: across the three model sizes, S_i^{sign} tracks S_i^{full} with model-level mean absolute errors of only 0.101–0.122. After the bounded map used by ADASIGNSGD, the corresponding scale error is only 0.067–0.075. Both relationships persist across all sampled training stages. A single sign is magnitude-free; its agreement rate carries SNR information.

4 ADASIGNSGD: VOTE-DERIVED ADAPTIVE UPDATES

Let t index optimization steps, $n \in \{1, \dots, N\}$ index synchronized workers, and $i \in \{1, \dots, d\}$ index coordinates of the shared parameters $\boldsymbol{\theta}_t \in \mathbb{R}^d$. Worker n receives local gradient $\mathbf{g}_t^{(n)} \in \mathbb{R}^d$ and maintains one local EMA momentum $\mathbf{m}_{t-1}^{(n)} \in \mathbb{R}^d$. Algorithm 1 gives the complete optimizer step and one-time LUT precomputation. Figure 2 summarizes the two-phase VOTEREDUCE collective; Appendix E.2 gives its full specification.

4.1 FROM EXACT COUNTS TO ADAPTIVE UPDATES

For coordinate i, $b_{t,i}^{(n)}$ is worker n's one-bit vote, and $c_{t,i} := \sum_{n=1}^{N} b_{t,i}^{(n)} \in \{0, \dots, N\}$ is the exact number of nonnegative votes at step t. The same count determines the majority direction $\widehat{y}_{t,i} = \operatorname{sgn}_+(2c_{t,i} - N)$ and its margin from a tie $\delta_{t,i} = |2c_{t,i} - N|$. The margin indexes the precomputed LUT_N, which returns the adaptive scale $\widehat{a}_{t,i}$, and their product gives $u_{t,i} = \widehat{y}_{t,i}\widehat{a}_{t,i}$. The inverse CDF is therefore evaluated only during LUT construction; each optimization step requires only the exact count, an element-wise lookup, and multiplication. The resulting $\mathbf{u}_t$ is the vote-derived adaptive update applied to the shared parameters with learning rate η_t and decoupled weight decay.

4.2 FINITE-WORKER RECOVERY UNDER THE GAUSSIAN MODEL

Fix one coordinate and step and suppress t, i. We use the idealized homogeneous worker model

$$m^{(n)} = \mu + \sigma\xi^{(n)}, \qquad \xi^{(n)} \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0,1), \qquad n = 1, \dots, N, \qquad \sigma > 0. \tag{3}$$

The nonnegative-sign probability q and directional agreement probability p satisfy

$$q := \Pr[m^{(n)} \geq 0] = \Phi(\mu/\sigma), \qquad p := \max\{q, 1-q\} = \Phi(S). \tag{4}$$

The scalar count corresponding to Algorithm 1 is $c := \sum_{n=1}^{N} \mathbb{I}[m^{(n)} \geq 0] \sim \text{Binomial}(N, q)$. Its exact value determines both the majority direction and the empirical support used to select the adaptive scale.

Define the bounded reliability scale $a(S) := S/\sqrt{1+S^2}$. The corresponding signed population target is

$$u^\star := \operatorname{sgn}_+(\mu)\, a(S) = \frac{\mu}{\sqrt{\mu^2 + \sigma^2}} = \frac{\mathbb{E}[m^{(n)}]}{\sqrt{\mathbb{E}[(m^{(n)})^2]}}. \tag{5}$$

Thus agreement identifies a bounded cross-worker RMS-normalized target: weak signal produces a small update, whereas decisive agreement approaches unit scale.

To isolate finite-worker estimation, set the scale stabilizer $\varepsilon = 0$. Define the clipped population SNR

$$S_{\epsilon_p} = \Phi^{-1}\big(\operatorname{clip}\big(\Phi(S), \tfrac{1}{2} + \epsilon_p, 1 - \epsilon_p\big)\big), \qquad K_{\epsilon_p} = \frac{1}{\varphi(\Phi^{-1}(1-\epsilon_p))}, \tag{6}$$

where φ is the standard-normal density. Write $\widehat{u}$ for the scalar update returned by Algorithm 1 from the random count c.

Theorem 1 (Finite-worker approximation)**.** *Under Equation 3,*

$$\mathbb{E}[|\widehat{u} - u^\star|] \le \left|a(S_{\epsilon_p}) - a(S)\right| + \frac{K_{\epsilon_p}}{2\sqrt{N}} + 2a(S_{\epsilon_p}) \exp\left[-2N\left(\Phi(S) - \tfrac{1}{2}\right)^2\right]. \tag{7}$$

Consequently, for every fixed $S > 0$ satisfying $\Phi(S) \in [1/2 + \epsilon_p, 1 - \epsilon_p]$,

$$\widehat{u} \xrightarrow[N\to\infty]{L^1} u^\star = \frac{\mu}{\sqrt{\mu^2 + \sigma^2}}. \tag{8}$$

The three terms in Equation 7 respectively bound probability clipping, finite-worker SNR-estimation error, and majority-direction error. Thus the exact vote count consistently estimates the SNR-derived bounded target in the unclipped regime.

5 Experiments

5.1 Training Trajectories Test Gaussian Worker Predictions

We evaluate Gaussian worker predictions on 18 C4 pretraining trajectories: 60M, 130M, and 350M LLaMA-style models, each at $N \in \{4, 8, 16, 32, 64, 128\}$ workers. Every 500 training steps, we hold the model parameters and worker momenta fixed and independently sample 128 global C4 batches. Each batch yields candidate next-step momenta without advancing training. We divide the 128 samples into four groups of 32. For each worker and selected coordinate, we fit a Gaussian using three groups and test its CDF and sign predictions on the fourth, using each group for testing once. Recorded training momenta provide the full-value reference for vote direction and RMS-normalized scale. Results are grouped into warmup, early, middle, and late stages (Appendix D).

Conditional fit and cross-worker dependence. Across all 72 prespecified model–N–stage cells, the conditional CDF gap to the matched Gaussian reference is at most $D_{\text{cond}} = 0.0156$. Applying $p = \Phi(S)$ from Equation 4 to each worker, predicted sign accuracy differs from observed accuracy by at most 0.339 percentage points. Moreover, the observed sign-count variance across the 128 draws matches the conditionally independent prediction retaining worker-specific sign probabilities to within 0.211% (Table 1).

Operational consequence. Exact vote ties are uninformative: the prescribed tie-break agrees with the sign of the same workers' full-value mean only 49.46–51.34% of the time (Appendix Table 25). At the fixed normalized margin $\delta/N = 0.25$, direction agreement increases monotonically with N, from 82.62–82.75% at $N = 8$ to 99.92–100.00% at $N = 128$ (Table 1). These trajectories support the Gaussian relation between momentum SNR and worker-sign probabilities, and show that vote counts carry information about the direction and RMS-normalized scale of the full-value update.

Table 1: **Training trajectories validate the mechanism.** The three gap columns report the worst case over all worker counts and all stages. The variance gap is the relative deviation from the conditionally independent, worker-specific count-variance prediction. Direction agreement is measured against the sign of the full-value mean of the same workers at the same step, at $\delta/N = 0.25$.

	Gaussian worker model		Conditional independence	Direction agreement at $\delta/N = 0.25$ (%)				
Model	CDF gap (%)	Accuracy gap (pp)	Variance gap (%)	$N = 8$	$N = 16$	$N = 32$	$N = 64$	$N = 128$
60M	0.974	0.156	**0.211**	82.75	90.77	97.25	99.26	100.00
130M	0.895	0.131	**0.205**	82.70	90.93	97.22	99.75	100.00
350M	1.558	0.339	**0.183**	82.62	90.67	96.83	99.67	99.92

5.2 Language-Model Pretraining

Optimizer comparisons are unusually sensitive to protocol artifacts: a shared hyperparameter recipe can favor one update rule over another, and rankings at intermediate checkpoints can reverse during

Table 2: **Validation perplexity on C4-en under the controlled endpoint protocol.** Each entry is the lowest endpoint perplexity in the corresponding method-specific sweep; intermediate checkpoints are never used for selection. Bold marks the best result in each setting among 1-bit worker-sign methods, and blue highlights ADASIGNSGD.

	60M				130M				350M				1B
Optimizer	1×	2×	4×	8×	1×	2×	4×	8×	1×	2×	4×	8×	1×
Dense references													
AdamW	34.56	32.17	30.41	29.04	26.68	25.07	23.90	23.18	20.26	19.26	18.56	17.98	17.08
ADASIGNSGD (uncompressed)	35.94	33.97	31.67	31.70	28.04	26.69	25.77	25.10	21.61	20.55	19.77	19.40	17.85
Communication-efficient methods with dense phases													
Birder	39.15	35.87	33.88	31.37	30.02	27.54	25.49	24.32	21.84	20.51	19.60	19.21	17.46
Birder + EC	38.37	34.90	32.74	30.52	29.31	26.71	24.67	23.48	20.83	19.64	18.79	18.56	18.01
1-Bit Adam	42.96	35.19	32.93	30.30	31.06	26.51	24.67	24.12	28.58	25.18	22.72	20.92	27.03
0/1 Adam	42.28	36.93	33.48	31.09	32.31	29.64	26.54	24.37	22.83	21.04	18.50	17.50	19.52
1-bit worker-sign methods													
SIGNUM-MV	72.73	66.49	70.98	74.03	69.83	70.47	97.41	98.59	107.03	139.77	98.42	97.05	128.14
SIGNUM-MV+EF	62.96	47.76	39.76	34.83	44.48	37.72	32.70	29.23	28.66	24.96	22.32	20.70	29.49
D-Lion-MaVo	72.14	62.87	59.07	68.38	99.23	128.06	160.96	161.63	124.60	115.29	117.17	111.22	117.36
ADASIGNSGD (ours)	**37.43**	**34.86**	**32.64**	**30.98**	**28.32**	**26.61**	**25.33**	**24.32**	**21.11**	**20.11**	**19.28**	**18.73**	**16.95**

learning-rate decay (Wen et al., 2026). We therefore treat this study as a controlled optimizer benchmark rather than a single-recipe ablation. Following the evaluation principles of Wen et al. (2026), we use three safeguards: method-specific tuning, evaluation across both model and token-budget scales, and model selection only at prescribed training endpoints.

Controlled setup. We pretrain LLaMA-style models (Touvron et al., 2023) from scratch on English C4 (Raffel et al., 2020), using sequence length 256, 8 synchronized workers, and global batch 512. Following LLM scaling practice (Hoffmann et al., 2022), $1\times$ denotes the reference update/token budget in Appendix B, while $2\times$, $4\times$, and $8\times$ multiply that horizon at fixed architecture and batch geometry. Within each model–budget setting, every method uses the same model definition and synchronized initialization, pretokenized corpus and stored example order, the same collection of eight local microbatches at every update, numerical precision, scheduler family, training endpoint, and fixed validation prefix of approximately one million next-token targets. Thus candidates consume identical tokens in identical order and are compared after identical numbers of updates; only the optimizer rule and explicitly swept optimizer hyperparameters differ.

Method-specific endpoint sweeps. Every method receives its own coordinate sweep over the applicable learning rate, momentum coefficients, warmup, weight decay, clipping, and method-specific controls. No method is selected at its most favorable intermediate checkpoint. Appendices B and J disclose the complete grids, endpoint losses, and $1{,}108$ distinct run identifiers. The baselines span dense references, communication-efficient optimizers with dense phases, and methods based on one-bit worker signs. MV denotes majority vote, EF error feedback, EC Birder's residual compensation (Peng et al., 2023), and MaVo the majority-vote D-Lion variant (Liu et al., 2024). ADASIGNSGD (uncompressed) directly uses the full-value cross-worker RMS-normalized update.

Results. The primary comparison is among methods whose worker messages are one-bit signs. Under this method-specific endpoint protocol, ADASIGNSGD ranks first in all 13 evaluated model–budget settings. SIGNUM-MV+EF is the strongest competing method in this class in every cell; ADASIGNSGD reduces its perplexity by 9.5%–42.5%, and it also beats D-Lion-MaVo in every cell. The comparison to dense AdamW also improves systematically with scale. At $1\times$, ADASIGNSGD's relative perplexity gap narrows from 8.3% at 60M to 6.2% at 130M and 4.2% at 350M, then reverses at 1B, where ADASIGNSGD reaches 16.95 versus AdamW's 17.08. Among methods that permit dense phases, ADASIGNSGD is lower than Birder in 12 cells and tied in one. Birder+EC is lower in seven cells, but by at most 0.84; ADASIGNSGD wins the remaining six, including the 1B comparison. We report this class separately because dense phases and residual compensation have a different memory and communication profile; Section 5.3 quantifies that systems trade-off.

5.3 SYSTEMS EFFICIENCY OF EXACT VOTE COUNTS

VOTEREDUCE reconstructs exact counts while preserving low-bit communication, but traffic reduction alone does not determine end-to-end systems efficiency: optimizer-state memory and the extent to which communication overlaps with backpropagation also shape the realized cost. We therefore evaluate three separate costs: persistent and peak memory, the exposed communication path, and complete forward-to-update step time, using the last as the end-to-end criterion.

Question-specific controls. Quality and systems scaling require different controls because varying the worker count N cannot keep both the global batch and per-worker token workload fixed. We therefore fix global batch for quality, but hold per-worker token shape fixed for systems through weak scaling (Gustafson, 1988; Hoefler & Belli, 2015); strong scaling would instead confound collective costs with shrinking local compute and activations. The configurations thus differ by design. Within each hardware block, all reported methods use the same model, per-worker token shape, parameter-aware bucketization, and timing boundary.

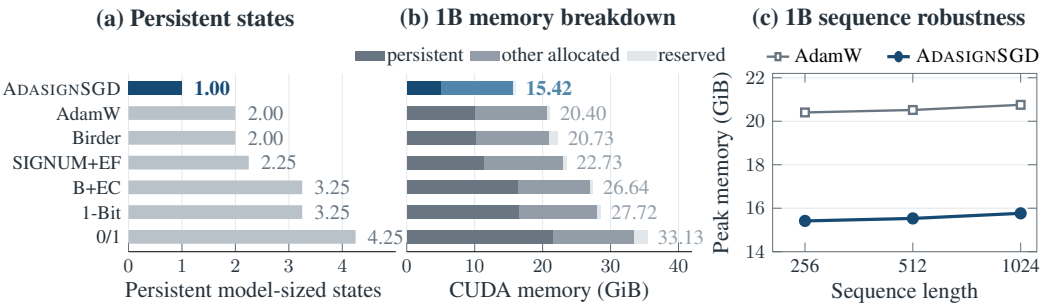

Figure 5: **Memory efficiency.** **(a)** Persistent model-sized optimizer states at $N = 4$. **(b)** CUDA memory for 1B training on 4×H200; dark and medium segments show persistent and other allocated memory, pale extensions show reserved memory, and labels give peak allocated GiB. **(c)** Complete-step peak allocated memory versus sequence length. SIGNUM+EF, B+EC, 1-Bit, and 0/1 abbreviate SIGNUM-MV+EF, Birder+EC, 1-Bit Adam, and 0/1 Adam, respectively. SIGNUM-MV and D-Lion-MaVo are omitted because they share the same one-state minimum and nearly identical measured memory footprints. Lower is better.

Memory. ADASIGNSGD keeps one worker-local model-sized momentum state and no model-sized residual, matching the minimum footprint of sign-only methods while recovering an adaptive scale. At 1B, its allocator-tracked persistent optimizer allocation is 4.988 GiB/GPU, versus 9.977 for AdamW and Birder, 11.226 for SIGNUM-MV+EF, 16.215 for Birder+EC and 1-Bit Adam, and 21.203 for 0/1 Adam. These measurements agree with analytical state counts within 0.26%: ADASIGNSGD uses 50.0–76.5% less persistent optimizer memory than these multi-state baselines. Complete-step peaks additionally include the shared parameter, gradient, and activation footprint plus method-specific communication buffers and temporaries. At 1B, ADASIGNSGD peaks at 15.42 GiB, versus 20.40 for AdamW, 22.73 for SIGNUM-MV+EF, 26.64 for Birder+EC, 27.72 for 1-Bit Adam, and 33.13 for 0/1 Adam, reductions of 24.4%, 44.4%, and 53.5% relative to AdamW, 1-Bit Adam, and 0/1 Adam, respectively. Sequence length changes activation and workspace memory but not parameter, gradient, or persistent optimizer-state sizes; the reduction relative to AdamW remains 24.0–24.4% from length 256 to 1024 (Figure 5).

Communication path. VOTEREDUCE preserves low-bit communication without approximating the reduced statistic: packed one-bit signs are routed to shard owners, which redistribute each exact count using either the minimum $\lceil \log_2(N+1) \rceil$ bits (q-bit) or one byte (`uint8`). At $N = 32$, these encodings require only 10.94% and 14.06% of AdamW's FP32 dense-DDP model-sized traffic, respectively. These savings translate into lower measured communication-path time. At $N = 128$ on P100, `uint8` requires only 19.7–20.2% of AdamW's communication-path time across 60M–350M, corresponding to 4.95–5.08× speedups, and is consistently faster than 1-Bit Adam, Birder, and Birder+EC. On H200, both optimized encodings achieve 1.29–2.82× speedups over dense DDP across all 16 model–world-size blocks. At $N = 32$, their speedups range from 1.88× to 2.18× across 60M–1B (Figure 6). For 1B at $N = 32$, `uint8` reduces communication-path time from 114.815 ms to 52.776 ms, a 2.18× speedup. The optimized and reference implementations agree

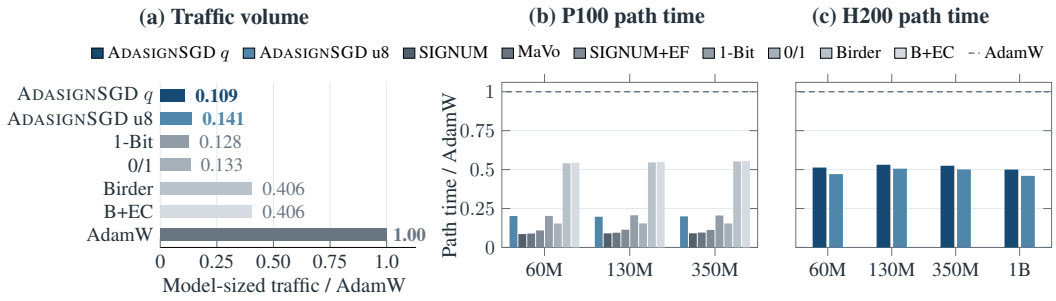

Figure 6: **Communication-path efficiency.** **(a)** Model-sized traffic at $N = 32$ normalized to AdamW. **(b,c)** Communication-path time normalized within each model to the corresponding AdamW dense-DDP path, for matched-stack P100 at $N = 128$ and optimized H200 at $N = 32$, respectively. Here q/`u8` denote minimum-width/`uint8` exact-count encodings, while SIGNUM and MaVo abbreviate SIGNUM-MV and D-Lion-MaVo. Lower is better.

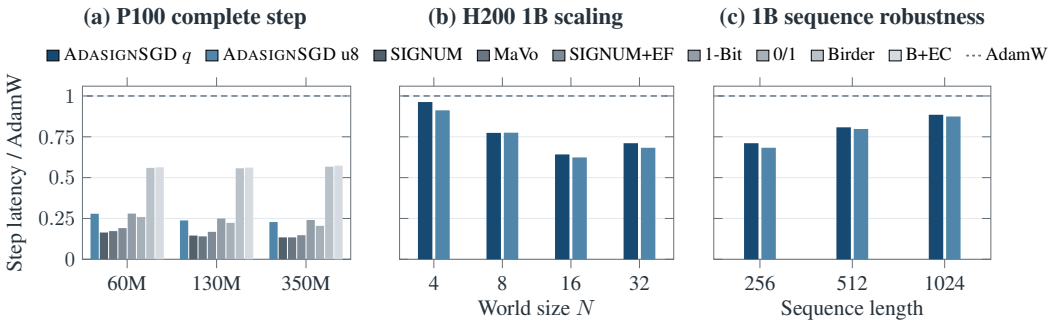

Figure 7: **Normalized complete-step latency.** **(a)** Matched-stack P100 comparison at $N = 128$. **(b)** H200 weak scaling for the 1B model. **(c)** H200 sequence-length robustness for the 1B model at $N = 32$. Values are normalized to AdamW within each hardware-model block; lower is better.

bitwise on identical inputs, while reversing benchmark order changes all three 1B/$N = 32$ means by less than 0.1%.

Complete step. VoteReduce turns low-bit exact-count communication into substantial end-to-end training acceleration. On P100 at $N = 128$, the `uint8` path requires only 27.9%, 23.8%, and 22.8% of AdamW's complete-step latency for the 60M, 130M, and 350M models, corresponding to $3.58\times$, $4.20\times$, and $4.39\times$ speedups, respectively. On H200, both encodings accelerate 1B training at every $N \in \{4, 8, 16, 32\}$, achieving 1.04–$1.60\times$ speedups over fused, overlapped AdamW. At $N = 32$, `uint8` lowers the complete step from 151.166 ms to 103.216 ms, a $1.46\times$ speedup. These gains persist as sequence length increases fourfold from 256 to 1024, where the q-bit and `uint8` paths retain $1.13\times$ and $1.14\times$ speedups, respectively (Figure 7a–c).

6 Conclusion

We introduced AdaSignSGD, which converts exact cross-worker sign counts into bounded adaptive updates, and VoteReduce, which constructs these counts efficiently from packed signs. Under the homogeneous Gaussian model, sign probabilities identify a cross-worker RMS-normalized scale; training-trajectory diagnostics separately show agreement with the full-value RMS-normalized update formed from the same workers (Appendix D). Across all 13 model–budget settings on C4, AdaSignSGD achieves the lowest perplexity among the evaluated 1-bit worker-sign methods and closely matches AdamW at 1B. These quality gains do not recover a dense systems footprint: AdaSignSGD uses one model-sized optimizer state, `uint8` VoteReduce retains only 14.06% of FP32 dense-DDP traffic at $N = 32$, and delivers a $1.46\times$ complete-step speedup for 1B training on H200 against fused, overlapped AdamW. Together, AdaSignSGD turns the information preserved in distributed one-bit votes into adaptive updates with both strong pretraining quality and systems efficiency.

AI USE STATEMENT

Generative AI tools assisted with drafting and polishing parts of the manuscript; retrieving and identifying related work; and research execution, including feedback on experimental protocols, code and implementation, and the interpretation and presentation of results. They also assisted with LaTeX formatting, migration to the template, and figure and table refinement. AI tools were not used to generate synthetic datasets or reported experimental measurements. The authors checked AI-assisted suggestions against the cited sources, underlying derivations, source code, and experiment records; reviewed the resulting text, equations, figures, tables, and citations; and take full responsibility for the final content and artifacts of this work.

REPRODUCIBILITY STATEMENT

The paper documents the assumptions, algorithms, and protocols underlying its principal claims. Appendix A states the theoretical assumptions and gives complete proofs; Algorithm 1 and Appendix E specify the update rule and exact distributed vote-count reduction, including tie handling and count encodings. Appendix B fixes the model and data pipeline, example order, initialization, batching, precision, evaluation, and endpoint-selection protocol. Appendix B.5 lists the selected configuration for each result in Table 2; Appendix J reports the completed candidates, including non-selected ones, with their configurations, endpoint losses, and 1,108 distinct Weights & Biases run identifiers. Appendix D specifies the checkpoint and coordinate sampling, conditional probes, and aggregation used for the trajectory diagnostics. Appendices F–I document the systems hardware and software, matched workloads, timing and memory boundaries, repetitions, and correctness checks. For double-blind review, identifying components of the tracking links are redacted while the run identifiers and reported measurements are retained. Upon acceptance, we will release the code, configuration files, run manifests, and corresponding tracking records.

REFERENCES

Dan Alistarh, Demjan Grubic, Jerry Li, Ryota Tomioka, and Milan Vojnovic. QSGD: Communication-efficient SGD via gradient quantization and encoding. In *Advances in Neural Information Processing Systems*, volume 30, pp. 1709–1720. Curran Associates, Inc., 2017. URL https://proceedings.neurips.cc/paper_files/paper/2017/hash/6c340f25839e6acdc73414517203f5f0-Abstract.html.

Lukas Balles and Philipp Hennig. Dissecting Adam: The sign, magnitude and variance of stochastic gradients. In *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pp. 404–413. PMLR, 2018. URL https://proceedings.mlr.press/v80/balles18a.html.

Jeremy Bernstein, Yu-Xiang Wang, Kamyar Azizzadenesheli, and Animashree Anandkumar. signSGD: Compressed optimisation for non-convex problems. In *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pp. 560–569. PMLR, 2018. URL https://proceedings.mlr.press/v80/bernstein18a.html.

Jeremy Bernstein, Jiawei Zhao, Kamyar Azizzadenesheli, and Anima Anandkumar. signSGD with majority vote is communication efficient and fault tolerant. In *International Conference on Learning Representations*, 2019. URL https://openreview.net/forum?id=BJxhijAcY7.

John L. Gustafson. Reevaluating Amdahl's law. *Communications of the ACM*, 31(5):532–533, 1988. doi: 10.1145/42411.42415.

Torsten Hoefler and Roberto Belli. Scientific benchmarking of parallel computing systems: Twelve ways to tell the masses when reporting performance results. In *Proceedings of the International Conference for High Performance Computing, Networking, Storage and Analysis*, SC '15, pp. 73:1–73:12. ACM, 2015. ISBN 978-1-4503-3723-6. doi: 10.1145/2807591.2807644.

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, Oriol Vinyals, Jack W. Rae, and Laurent Sifre. Training compute-optimal large language models. In *Advances in Neural Information Processing Systems*, volume 35, pp. 30016–30030. Curran Associates, Inc., 2022. doi: 10.52202/068431-2176. URL https://proceedings.neurips.cc/paper_files/paper/2022/file/c1e2faff6f588870935f114ebe04a3e5-Paper-Conference.pdf.

Ziheng Jiang, Haibin Lin, Yinmin Zhong, Qi Huang, Yangrui Chen, Zhi Zhang, Yanghua Peng, Xiang Li, Cong Xie, Shibiao Nong, Yulu Jia, Sun He, Hongmin Chen, Zhihao Bai, Qi Hou, Shipeng Yan, Ding Zhou, Yiyao Sheng, Zhuo Jiang, Haohan Xu, Haoran Wei, Zhang Zhang, Pengfei Nie, Leqi Zou, Sida Zhao, Liang Xiang, Zherui Liu, Zhe Li, Xiaoying Jia, Jianxi Ye, Xin Jin, and Xin Liu. MegaScale: Scaling large language model training to more than 10,000 GPUs. In *21st USENIX Symposium on Networked Systems Design and Implementation (NSDI 24)*, pp. 745–760. USENIX Association, 2024. ISBN 978-1-939133-39-7. URL https://www.usenix.org/conference/nsdi24/presentation/jiang-ziheng.

Sai Praneeth Karimireddy, Quentin Rebjock, Sebastian U. Stich, and Martin Jaggi. Error feedback fixes SignSGD and other gradient compression schemes. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pp. 3252–3261. PMLR, 2019. URL https://proceedings.mlr.press/v97/karimireddy19a.html.

Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *International Conference on Learning Representations*, 2015.

Conglong Li, Ammar Ahmad Awan, Hanlin Tang, Samyam Rajbhandari, and Yuxiong He. 1-bit LAMB: Communication efficient large-scale large-batch training with LAMB's convergence speed. In *2022 IEEE 29th International Conference on High Performance Computing, Data, and Analytics (HiPC)*, pp. 272–281. IEEE, 2022. doi: 10.1109/HiPC56025.2022.00044. URL https://ieeexplore.ieee.org/document/10106313.

Shen Li, Yanli Zhao, Rohan Varma, Omkar Salpekar, Pieter Noordhuis, Teng Li, Adam Paszke, Jeff Smith, Brian Vaughan, Pritam Damania, and Soumith Chintala. PyTorch Distributed: Experiences on accelerating data parallel training. *Proceedings of the VLDB Endowment*, 13(12):3005–3018, 2020. doi: 10.14778/3415478.3415530. URL https://www.vldb.org/pvldb/vol13/p3005-li.pdf.

Yujun Lin, Song Han, Huizi Mao, Yu Wang, and William J. Dally. Deep gradient compression: Reducing the communication bandwidth for distributed training. In *International Conference on Learning Representations*, 2018. URL https://openreview.net/forum?id=SkhQHMW0W.

Bo Liu, Lemeng Wu, Lizhang Chen, Kaizhao Liang, Jiaxu Zhu, Chen Liang, Raghuraman Krishnamoorthi, and Qiang Liu. Communication efficient distributed training with distributed Lion. In *Advances in Neural Information Processing Systems*, volume 37, pp. 18388–18415. Curran Associates, Inc., 2024. doi: 10.52202/079017-0583. URL https://proceedings.neurips.cc/paper_files/paper/2024/hash/20cea6c1b36ae5f69c48427a68b67fbc-Abstract-Conference.html.

Yucheng Lu, Conglong Li, Minjia Zhang, Christopher De Sa, and Yuxiong He. Maximizing communication efficiency for large-scale training via 0/1 Adam. In *International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=-CefY2EOupj.

NVIDIA. GPUDirect RDMA. https://docs.nvidia.com/cuda/archive/12.1.0/gpudirect-rdma/index.html, 2023. CUDA 12.1.0 documentation.

NVIDIA. NVIDIA H200 GPU. https://www.nvidia.com/en-us/data-center/h200/, n.d. Accessed September 23, 2026.

Hanyang Peng, Shuang Qin, Yue Yu, Jin Wang, Hui Wang, and Ge Li. Birder: Communication-efficient 1-bit adaptive optimizer for practical distributed DNN training. In *Advances in Neural Information Processing Systems*, volume 36, pp.

39529–39552. Curran Associates, Inc., 2023. doi: 10.52202/075280-1717. URL https://proceedings.neurips.cc/paper_files/paper/2023/hash/7c72fcd7b6bffc3864c7152ab5a2dd83-Abstract-Conference.html.

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. URL https://www.jmlr.org/papers/v21/20-074.html.

Frank Seide, Hao Fu, Jasha Droppo, Gang Li, and Dong Yu. 1-bit stochastic gradient descent and its application to data-parallel distributed training of speech DNNs. In *Interspeech 2014*, pp. 1058–1062, 2014. doi: 10.21437/Interspeech.2014-274. URL https://www.isca-archive.org/interspeech_2014/seide14_interspeech.html.

Sebastian U. Stich, Jean-Baptiste Cordonnier, and Martin Jaggi. Sparsified SGD with memory. In *Advances in Neural Information Processing Systems*, volume 31, pp. 4447–4458. Curran Associates, Inc., 2018. URL https://proceedings.neurips.cc/paper_files/paper/2018/hash/b440509a0106086a67bc2ea9df0a1dab-Abstract.html.

Hanlin Tang, Shaoduo Gan, Ammar Ahmad Awan, Samyam Rajbhandari, Conglong Li, Xiangru Lian, Ji Liu, Ce Zhang, and Yuxiong He. 1-bit Adam: Communication efficient large-scale training with Adam's convergence speed. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 10118–10129. PMLR, 2021. URL https://proceedings.mlr.press/v139/tang21a.html.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. LLaMA: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*, 2023. URL https://arxiv.org/abs/2302.13971.

Thijs Vogels, Sai Praneeth Karimireddy, and Martin Jaggi. PowerSGD: Practical low-rank gradient compression for distributed optimization. In *Advances in Neural Information Processing Systems*, volume 32, pp. 14236–14245. Curran Associates, Inc., 2019. URL https://proceedings.neurips.cc/paper_files/paper/2019/hash/d9fbed9da256e344c1fa46bb46c34c5f-Abstract.html.

Kaiyue Wen, David Hall, Tengyu Ma, and Percy Liang. Fantastic pretraining optimizers and where to find them. In *International Conference on Learning Representations*, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/hash/ea4f0a6ad7088c9b93f22d74c0c9e8c2-Abstract-Conference.html.

Wei Wen, Cong Xu, Feng Yan, Chunpeng Wu, Yandan Wang, Yiran Chen, and Hai Li. TernGrad: Ternary gradients to reduce communication in distributed deep learning. In *Advances in Neural Information Processing Systems*, volume 30, pp. 1509–1519. Curran Associates, Inc., 2017. URL https://proceedings.neurips.cc/paper_files/paper/2017/hash/89fcd07f20b6785b92134bd6c1d0fa42-Abstract.html.

Xingyu Xie, Zhijie Lin, Kim-Chuan Toh, and Pan Zhou. LoCo: Low-bit communication adaptor for large-scale model training. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 47(6):4285–4298, 2025. doi: 10.1109/TPAMI.2025.3544764. URL https://ieeexplore.ieee.org/document/10904294.

Hang Xu, Kelly Kostopoulou, Aritra Dutta, Xin Li, Alexandros Ntoulas, and Panos Kalnis. DeepReduce: A sparse-tensor communication framework for federated deep learning. In *Advances in Neural Information Processing Systems*, volume 34, pp. 21150–21163. Curran Associates, Inc., 2021. URL https://proceedings.neurips.cc/paper_files/paper/2021/hash/b0ab42fcb7133122b38521d13da7120b-Abstract.html.

A ONE-BIT RECOVERY OF RMS-NORMALIZED UPDATES

Under a homogeneous Gaussian model, finite-worker error bounds establish RMS approximation and consistency in the unclipped range. Sections A.2 and A.3 treat heterogeneous consensus and conditional optimization error. Scalar results apply to one coordinate and step, without coordinate independence. Define

$$S := \frac{|\mu|}{\sigma}, \qquad y := \operatorname{sign}(\mu), \qquad a(S) := \frac{S}{\sqrt{1+S^2}},$$

where $\operatorname{sign}(0) = 0$, whereas the encoded sign satisfies $\operatorname{sgn}_+(0) = +1$. The population RMS-normalized target is

$$u^\star := \frac{\mu}{\sqrt{\mu^2+\sigma^2}} = y\, a(S).$$

Assumption 1 (Homogeneous Gaussian worker model)**.** *For a fixed coordinate and optimization step, the worker momenta satisfy*

$$m^{(n)} = \mu + \sigma\xi^{(n)}, \qquad \xi^{(n)} \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0,1), \qquad n = 1,\dots,N,$$

with $\mu \in \mathbb{R}$, $\sigma > 0$, and integer $N \geq 1$. Worker n communicates $b^{(n)} := \mathbb{I}[m^{(n)} \geq 0] \in \{0,1\}$.

Let Φ and φ denote the standard-normal CDF and density. Write

$$c := \sum_{n=1}^{N} b^{(n)}, \qquad \widehat{q} := \frac{c}{N}, \qquad q := \Pr[m^{(n)} \geq 0] = \Phi(\mu/\sigma),$$

and define

$$p := \max\{q, 1-q\} = \Phi(S), \qquad \widehat{p} := \max\{\widehat{q}, 1-\widehat{q}\}, \qquad \widehat{y} := 2\mathbb{I}[2c \geq N] - 1.$$

Exact ties select $\widehat{y} = +1$; $y = 0$ denotes a zero population target. The Bernoulli extension permits atoms at zero, unlike the Gaussian model. Fix $\epsilon_p \in (0, 1/4)$, and let

$$\operatorname{clip}\big(z, \tfrac{1}{2}+\epsilon_p, 1-\epsilon_p\big) := \min\big\{\max\big\{z, \tfrac{1}{2}+\epsilon_p\big\}, 1-\epsilon_p\big\}.$$

The clipped population and empirical SNRs are

$$S_{\epsilon_p} := \Phi^{-1}\big(\operatorname{clip}\big(p, \tfrac{1}{2}+\epsilon_p, 1-\epsilon_p\big)\big), \qquad \widehat{S} := \Phi^{-1}\big(\operatorname{clip}\big(\widehat{p}, \tfrac{1}{2}+\epsilon_p, 1-\epsilon_p\big)\big).$$

We use the shorthand

$$s^-_{\epsilon_p} := \Phi^{-1}(1/2+\epsilon_p), \qquad s^+_{\epsilon_p} := \Phi^{-1}(1-\epsilon_p), \qquad K_{\epsilon_p} := \frac{1}{\varphi(s^+_{\epsilon_p})}.$$

Lemma 1 (Finite-sample vote-to-SNR concentration)**.** *Under Assumption 1, for every $\zeta > 0$,*

$$\Pr\Big(|\widehat{S} - S_{\epsilon_p}| \geq K_{\epsilon_p}\zeta\Big) \leq 2e^{-2N\zeta^2}.$$

Consequently, for $\rho \in (0,1)$, with probability at least $1-\rho$,

$$|\widehat{S} - S_{\epsilon_p}| \leq K_{\epsilon_p}\sqrt{\frac{\log(2/\rho)}{2N}},$$

and

$$\mathbb{E}|\widehat{S} - S_{\epsilon_p}| \leq \frac{K_{\epsilon_p}}{2\sqrt{N}}, \qquad \mathbb{E}|\widehat{S} - S_{\epsilon_p}|^2 \leq \frac{K^2_{\epsilon_p}}{4N}.$$

Proof. The support and clipping maps are 1-Lipschitz, and

$$\sup_{u \in [1/2+\epsilon_p,\, 1-\epsilon_p]} \big|(\Phi^{-1})'(u)\big| = K_{\epsilon_p}.$$

Hence

$$|\widehat{S} - S_{\epsilon_p}| \leq K_{\epsilon_p}|\widehat{q} - q|.$$

The tail bound follows from Hoeffding's inequality. The moment bounds follow from

$$\mathbb{E}|\widehat{q} - q| \leq \sqrt{\operatorname{Var}(\widehat{q})} \leq \frac{1}{2\sqrt{N}}, \qquad \mathbb{E}|\widehat{q} - q|^2 = \frac{q(1-q)}{N} \leq \frac{1}{4N}.$$

□

Remark 1 (Clipping bias). *The projection identity*

$$S_{\epsilon_p} = \Pi_{[s^-_{\epsilon_p},\, s^+_{\epsilon_p}]}(S),$$

with Π_I denoting projection onto I, gives

$$|S_{\epsilon_p} - S| = \begin{cases} s^-_{\epsilon_p} - S, & S < s^-_{\epsilon_p}, \\ 0, & S \in [s^-_{\epsilon_p}, s^+_{\epsilon_p}], \\ S - s^+_{\epsilon_p}, & S > s^+_{\epsilon_p}. \end{cases}$$

Upper clipping bounds $(\Phi^{-1})'$; either boundary can induce population bias. Empirical clipping can occur even for an interior population value.

Lemma 2 (Asymptotic efficiency). *Suppose Assumption 1 holds and $S \in (s^-_{\epsilon_p}, s^+_{\epsilon_p})$. Hold μ, σ, ϵ_p fixed as $N \to \infty$. Then*

$$\sqrt{N}(\widehat{S} - S) \Rightarrow \mathcal{N}\left(0,\, \frac{\Phi(S)(1-\Phi(S))}{\varphi(S)^2}\right).$$

The variance is the inverse single-sign Fisher information on either fixed-orientation submodel $y \in \{-1, +1\}$. Exponentially rare orientation errors preserve efficiency when y is unknown.

Proof. For $S > 0$,

$$\Pr(\widehat{y} \neq y) \leq \exp\left\{-2N\big(\Phi(S) - 1/2\big)^2\right\}.$$

With probability $1 - o(1)$, clipping is inactive and $\widehat{p}$ equals the positive-sign frequency of $y(2b^{(n)} - 1)$, with success probability $\Phi(S)$. The central limit theorem and delta method give the limit, with

$$I_1(S) = \frac{\varphi(S)^2}{\Phi(S)(1-\Phi(S))},$$

the single-sign Fisher information. □

Define

$$u^\star_{\epsilon_p} := y\, a(S_{\epsilon_p}), \qquad \widehat{u} := \widehat{y}\, a(\widehat{S}),$$

and set

$$\gamma(S) := \Phi(S) - \frac{1}{2}, \qquad \Delta_{\epsilon_p, N}(S) := \frac{K^2_{\epsilon_p}}{2N} + 8a(S_{\epsilon_p})^2 e^{-2N\gamma(S)^2}.$$

Theorem 2 (One-bit approximation of the RMS-normalized update). *Under Assumption 1,*

$$\mathbb{E}|\widehat{u} - u^\star_{\epsilon_p}| \leq \frac{K_{\epsilon_p}}{2\sqrt{N}} + 2a(S_{\epsilon_p}) e^{-2N\gamma(S)^2},$$

and

$$\mathbb{E}|\widehat{u} - u^\star_{\epsilon_p}|^2 \leq \Delta_{\epsilon_p, N}(S).$$

Moreover,

$$\mathbb{E}|\widehat{u} - u^\star| \leq |a(S_{\epsilon_p}) - a(S)| + \frac{K_{\epsilon_p}}{2\sqrt{N}} + 2a(S_{\epsilon_p}) e^{-2N\gamma(S)^2}.$$

Proof. For $s_1, s_2 \geq 0$, $y \in \{-1, 0, +1\}$, and $\widetilde{y} \in \{-1, 0, +1\}$,

$$|\widetilde{y}\, a(s_1) - y\, a(s_2)| \leq |a(s_1) - a(s_2)| + 2a(s_2)\mathbb{I}[\widetilde{y} \neq y].$$

Since $a'(s) = (1 + s^2)^{-3/2} \leq 1$,

$$|\widehat{u} - u^\star_{\epsilon_p}| \leq |\widehat{S} - S_{\epsilon_p}| + 2a(S_{\epsilon_p})\mathbb{I}[\widehat{y} \neq y].$$

For $S > 0$, Hoeffding's inequality gives

$$\Pr(\widehat{y} \neq y) \leq e^{-2N\gamma(S)^2};$$

at $S = 0$, this bound equals one. Lemma 1 and $(u+v)^2 \leq 2u^2 + 2v^2$ give the first two claims. The last follows from $|\widehat{u} - u^\star| \leq |\widehat{u} - u^\star_{\epsilon_p}| + |a(S_{\epsilon_p}) - a(S)|$. □

Corollary 1 (Consistency of one-bit Adam-style adaptivity)**.** *Under Assumption 1, for fixed μ, σ, ϵ_p with $S \in [s^-_{\epsilon_p}, s^+_{\epsilon_p}]$,*

$$\mathbb{E}|\widehat{u} - u^\star| = \mathcal{O}(N^{-1/2}) + \mathcal{O}\Big(e^{-2N\gamma(S)^2}\Big).$$

In particular, for every fixed $S > 0$ in the unclipped range,

$$\widehat{u} \xrightarrow[N\to\infty]{L^1} \frac{\mu}{\sqrt{\mu^2 + \sigma^2}}.$$

Remark 2 (Zero signal, ties, and the full-value reference)**.** *At $S = 0$, $u^\star = u^\star_{\epsilon_p} = 0$, whereas fixed lower clipping gives*

$$\mathbb{E}|\widehat{u}| \longrightarrow a(s^-_{\epsilon_p}) > 0, \qquad \mathbb{E}|\widehat{u}|^2 \longrightarrow a(s^-_{\epsilon_p})^2.$$

These limits follow from $\widehat{q} \to 1/2$ and boundedness. Symmetry gives

$$\mathbb{E}\widehat{u} = \begin{cases} a(s^-_{\epsilon_p}) \Pr(c = N/2), & N \text{ even}, \\ 0, & N \text{ odd}. \end{cases}$$

A tie produces $+a(s^-_{\epsilon_p})$. Mean cancellation therefore does not imply consistency at zero; Theorem 2 remains valid. Under Corollary 1, the law of large numbers and boundedness also give L^1 convergence of the full-value reference in Equation 67 to $u^\star$. This is asymptotic agreement, not a finite-worker reconstruction identity.

A.1 Statistical Consequences for Adaptive Scaling

Corollary 2 (Efficient recovery of the reliability scale)**.** *Under Lemma 2, let $a^\star := a(S)$. Then*

$$\sqrt{N}\big(a(\widehat{S}) - a^\star\big) \Rightarrow \mathcal{N}\left(0,\, \frac{\Phi(S)(1-\Phi(S))}{\varphi(S)^2(1+S^2)^3}\right).$$

On either fixed-orientation submodel $y \in \{-1, +1\}$, this is the inverse per-sign Fisher information for $a(S)$; the corresponding N-sample Cramér–Rao lower bound for unbiased estimators satisfying the regularity conditions is this quantity divided by N.

Proof. Apply the delta method with $a'(S) = (1 + S^2)^{-3/2}$. For $\alpha = a(S)$,

$$I^\alpha_1(\alpha) = I_1(S)\left(\frac{dS}{d\alpha}\right)^2 = \frac{\varphi(S)^2(1+S^2)^3}{\Phi(S)(1-\Phi(S))},$$

whose inverse is the displayed asymptotic variance. □

Proposition 1 (Majority direction is insufficient for adaptivity)**.** *Under Assumption 1, fix $s^-_{\epsilon_p} < S_- < S_+ < s^+_{\epsilon_p}$ for two coordinates with the same population direction. Any infinite-worker rule whose limiting magnitude depends only on that direction uses a common magnitude α; hence*

$$\max\big\{|\alpha - a(S_-)|,\, |\alpha - a(S_+)|\big\} \geq \frac{a(S_+) - a(S_-)}{2}.$$

By contrast, $\widehat{y}\, a(\widehat{S}) \to y\, a(S)$ in L^1 for each fixed unclipped $S > 0$.

Proof. The lower bound follows from $\max\{|\alpha - u|, |\alpha - v|\} \geq |u - v|/2$; the final claim follows from Corollary 1. □

A.2 Consensus with Non-Identical Workers

Let $\mathcal{F}$ be the sigma-field generated by the state before the current worker draws. Define

$$q_n := \Pr(m^{(n)} \geq 0 \mid \mathcal{F}), \quad \overline{q} := \frac{1}{N}\sum_n q_n, \quad \overline{p} := \max\{\overline{q}, 1 - \overline{q}\},$$

$$S_{\text{eff},\epsilon_p} := \Phi^{-1}(\text{clip}(\overline{p}, 1/2 + \epsilon_p, 1 - \epsilon_p)).$$

This clipped consensus parameter is finite at $\overline{q} \in \{0, 1\}$ and need not represent a shared momentum SNR. Retaining $c, \widehat{y}, \widehat{S}, \widehat{u}$, define

$$\gamma_{\text{eff}} := |\overline{q} - 1/2|, \qquad y_{\text{eff}} := \text{sign}(\overline{q} - 1/2), \qquad u_{\text{eff},\epsilon_p} := y_{\text{eff}}\, a(S_{\text{eff},\epsilon_p}).$$

Proposition 2 (Conditional Poisson-binomial consensus concentration)**.** *Suppose the bits $b^{(n)} = \mathbb{I}[m^{(n)} \geq 0]$ are mutually independent conditional on $\mathcal{F}$, with arbitrary probabilities $q_n \in [0,1]$. Then the count is conditionally Poisson-binomial, and, almost surely,*

$$V_N := \mathrm{Var}(\widehat{q} \mid \mathcal{F}) = \frac{1}{N^2}\sum_n q_n(1-q_n) = \frac{\overline{q}(1-\overline{q})}{N} - \frac{1}{N^2}\sum_n (q_n - \overline{q})^2 \leq \frac{1}{4N}, \tag{9}$$

$$\Pr(|\widehat{q} - \overline{q}| \geq \zeta \mid \mathcal{F}) \leq 2e^{-2N\zeta^2}, \qquad \zeta > 0. \tag{10}$$

With $\delta := |2c - N|$,

$$|\delta/N - |2\overline{q} - 1|| \leq 2|\widehat{q} - \overline{q}|, \tag{11}$$

$$\mathbb{E}[|\widehat{S} - S_{\mathrm{eff},\epsilon_p}|^2 \mid \mathcal{F}] \leq K_{\epsilon_p}^2 V_N, \tag{12}$$

$$\mathbb{E}[|a(\widehat{S}) - a(S_{\mathrm{eff},\epsilon_p})|^2 \mid \mathcal{F}] \leq K_{\epsilon_p}^2 V_N. \tag{13}$$

Both scale errors have first moment at most $K_{\epsilon_p}\sqrt{V_N}$, and

$$\Pr(|\widehat{S} - S_{\mathrm{eff},\epsilon_p}| \geq K_{\epsilon_p}\zeta \mid \mathcal{F}) \leq 2e^{-2N\zeta^2}.$$

The signed estimator satisfies

$$\mathbb{E}[|\widehat{u} - u_{\mathrm{eff},\epsilon_p}| \mid \mathcal{F}] \leq K_{\epsilon_p}\sqrt{V_N} + 2a(S_{\mathrm{eff},\epsilon_p})e^{-2N\gamma_{\mathrm{eff}}^2}, \tag{14}$$

$$\mathbb{E}[|\widehat{u} - u_{\mathrm{eff},\epsilon_p}|^2 \mid \mathcal{F}] \leq 2K_{\epsilon_p}^2 V_N + 8a(S_{\mathrm{eff},\epsilon_p})^2 e^{-2N\gamma_{\mathrm{eff}}^2}. \tag{15}$$

Consequently, for fixed ϵ_p,

$$\mathbb{E}[|a(\widehat{S}) - a(S_{\mathrm{eff},\epsilon_p})|^2 \mid \mathcal{F}] \leq \frac{K_{\epsilon_p}^2}{4N} \to 0.$$

The target may depend on N. Writing the N-dependent quantities with subscripts, if $\inf_{N\geq 1}\gamma_{\mathrm{eff},N} > 0$ almost surely, then

$$\mathbb{E}|\widehat{u}_N - u_{\mathrm{eff},\epsilon_p,N}|^2 \to 0.$$

No signed consistency is asserted when $\overline{q}_N = 1/2$.

Proof. Conditional variance additivity and expansion of $\sum_n (q_n - \overline{q})^2$ give the variance identity. Hoeffding's inequality and the Lipschitz argument of Lemma 1 apply without identical distributions. The direction-error events are

$$\{\widehat{y} \neq y_{\mathrm{eff}}\} = \begin{cases} \{\widehat{q} < 1/2\}, & \overline{q} > 1/2, \\ \{\widehat{q} \geq 1/2\}, & \overline{q} < 1/2, \\ \Omega, & \overline{q} = 1/2. \end{cases}$$

One-sided Hoeffding bounds and the deterministic decomposition in Theorem 2 give the signed bounds, including ties. Consistency follows from $V_N \leq 1/(4N)$ and dominated convergence. □

If Assumption 1 holds *conditional on* $\mathcal{F}$, with $\mathcal{F}$-measurable μ, σ, then

$$q_n = \Phi(\mu/\sigma), \qquad S_{\mathrm{eff},\epsilon_p} = S_{\epsilon_p}, \qquad y_{\mathrm{eff}} = y.$$

For trivial $\mathcal{F}$, Proposition 2 recovers Lemma 1 and Theorem 2 with unchanged constants. An unconditional Gaussian law does not imply this conditional identification; fixed clipping bias and heterogeneous RMS discrepancy may persist as $N \to \infty$.

Non-identifiability of pooled RMS. For $m^{(n)} \mid \mathcal{F} \sim \mathcal{N}(\mu_n, \sigma_n^2)$, $\sigma_n > 0$, the probabilities $q_n = \Phi(\mu_n/\sigma_n)$ do not generally identify

$$u_{\mathrm{pool}} := \frac{N^{-1}\sum_n \mu_n}{\sqrt{N^{-1}\sum_n(\mu_n^2 + \sigma_n^2)}}.$$

The independent pairs

$$\big(\mathcal{N}(1,1), \mathcal{N}(1,1)\big), \qquad \big(\mathcal{N}(1,1), \mathcal{N}(3,9)\big)$$

have identical joint sign laws but targets $1/\sqrt{2}$ and $2/\sqrt{10}$. Thus neither the complete sign law nor its pooled probit inverse identifies pooled RMS. For any $\mathcal{F}$-measurable reference v,

$$\mathbb{E}[|\widehat{u} - v|^2 \mid \mathcal{F}] \leq 2\mathbb{E}[|\widehat{u} - u_{\mathrm{eff},\epsilon_p}|^2 \mid \mathcal{F}] + 2|u_{\mathrm{eff},\epsilon_p} - v|^2.$$

The second term is modeling error rather than count-estimation error.

Dependent bits. With $V_N := \mathrm{Var}(c \mid \mathcal{F})/N^2$, the SNR and reliability-scale second-moment bounds through (13) hold without independence. Chebyshev replaces $2e^{-2N\zeta^2}$ by V_N/ζ^2. Magnitude consistency follows from $\mathbb{E}[V_N] \to 0$; sampled-state covariance measurements alone do not establish exponential concentration.

Empirical scope. Appendix D measures conditional shape, probit calibration, worker heterogeneity, and net covariance (Tables 21, 22, 23, and 24). Reset probes fix the realized model and persistent states and resample only the next mini-batches; the measurements establish neither exact Gaussian tails, mutual conditional independence, nor unconditional independence of worker histories. Table 26 and Figure 14 assess compression fidelity against the full-value reference, not exact LUT recovery, uniform population-error bounds, or gradient alignment.

A.3 Optimization-Level Perturbation

Let $\mathcal{F}_t$ be the pre-draw history, containing $\boldsymbol{\theta}_t$, $\{\mathbf{m}_{t-1}^{(n)}\}_n$, and the past sampling state, but excluding current draws and $\mathbf{m}_t^{(n)}$. Take deterministic $\boldsymbol{\theta}_0$, $\eta_t > 0$, and integer $d \geq 1$, with an integer worker count $N \geq 1$. For $\mathcal{F}_t$-measurable $\mu_{t,i} \in \mathbb{R}$ and $\sigma_{t,i} > 0$, define

$$\mathbb{E}_t[\cdot] := \mathbb{E}[\cdot \mid \mathcal{F}_t], \qquad S_{t,i} := \frac{|\mu_{t,i}|}{\sigma_{t,i}}, \qquad y_{t,i} := \mathrm{sign}(\mu_{t,i}),$$

and define

$$S_{\epsilon_p,t,i} := \Phi^{-1}\big(\mathrm{clip}\big(\Phi(S_{t,i}), \tfrac{1}{2} + \epsilon_p, 1 - \epsilon_p\big)\big), \qquad u^\star_{\epsilon_p,t,i} := y_{t,i} a(S_{\epsilon_p,t,i}).$$

For N workers, let

$$D_t(N) := \sum_{i=1}^{d} \Delta_{\epsilon_p,N}(S_{t,i}).$$

The explicit Gaussian specialization of Assumption 1 is the conditional product law

$$\mathcal{L}\Big((m_{t,i}^{(n)})_{n=1}^N \mid \mathcal{F}_t\Big) = \mathcal{N}(\mu_{t,i}, \sigma_{t,i}^2)^{\otimes N}, \qquad i = 1, \ldots, d.$$

Conditioning on EMA histories does not imply this product law, which requires neither coordinate independence nor unconditional independence of histories. The explicit Gaussian bounds use real-arithmetic $a(\widehat{S})$, $\varepsilon = 0$, and no update-norm clipping (Appendix E.1); the descent recursion takes $\lambda = 0$. The general formulation below does not require the product law.

Theorem 3 (Controlled perturbation of clipped RMS descent)**.** *Let $f : \mathbb{R}^d \to \mathbb{R}$ be L-smooth, $L \geq 0$. Under the conditional homogeneous law above,*

$$\mathbb{E}_t\|\widehat{\mathbf{u}}_t - \mathbf{u}^\star_{\epsilon_p,t}\|_2^2 \leq D_t(N), \qquad \mathbb{E}_t\|\widehat{\mathbf{u}}_t - \mathbf{u}^\star_{\epsilon_p,t}\|_2 \leq \sqrt{D_t(N)}.$$

If $\boldsymbol{\theta}_{t+1} = \boldsymbol{\theta}_t - \eta_t \widehat{\mathbf{u}}_t$, then

$$\mathbb{E}_t[f(\boldsymbol{\theta}_{t+1})] \leq f(\boldsymbol{\theta}_t) - \eta_t \langle \nabla f(\boldsymbol{\theta}_t), \mathbf{u}^\star_{\epsilon_p,t}\rangle + \eta_t \|\nabla f(\boldsymbol{\theta}_t)\|_2 \sqrt{D_t(N)} + \frac{L\eta_t^2 d}{2}.$$

For the general case, let $\mathbf{v}_t$ and integrable $B_t \geq 0$ be $\mathcal{F}_t$-measurable and satisfy

$$\mathbb{E}_t\|\widehat{\mathbf{u}}_t - \mathbf{v}_t\|_2^2 \leq B_t, \qquad \|\widehat{\mathbf{u}}_t\|_2^2 \leq d \quad \textit{almost surely}. \tag{16}$$

All three inequalities then hold with $(\mathbf{u}^\star_{\epsilon_p,t}, D_t(N))$ replaced by $(\mathbf{v}_t, B_t)$, without Gaussian or independence assumptions.

Proof. Sum Theorem 2 over coordinates and apply Jensen. Smoothness and $\|\widehat{\mathbf{u}}_t\|_2^2 \leq d$ give

$$f(\boldsymbol{\theta}_{t+1}) \leq f(\boldsymbol{\theta}_t) - \eta_t \langle \nabla f(\boldsymbol{\theta}_t), \mathbf{u}^\star_{\epsilon_p,t}\rangle - \eta_t \langle \nabla f(\boldsymbol{\theta}_t), \widehat{\mathbf{u}}_t - \mathbf{u}^\star_{\epsilon_p,t}\rangle + \frac{L\eta_t^2 d}{2}.$$

Take $\mathbb{E}_t$ and apply Cauchy–Schwarz. The general case uses $\mathbb{E}_t\|\widehat{\mathbf{u}}_t - \mathbf{v}_t\|_2 \leq \sqrt{B_t}$ in the same argument. □

Transfer to a general reference. The optimization theorems and corollaries below also hold under

$$\mathbf{u}^{\star}_{\epsilon_p,t} \mapsto \mathbf{v}_t, \qquad D_t(N) \mapsto B_t, \qquad \mathcal{G}_{\epsilon_p}(\boldsymbol{\theta}_t) \mapsto \mathcal{G}_t := \langle \nabla f(\boldsymbol{\theta}_t), \mathbf{v}_t \rangle,$$

provided $\mathbf{v}_t$ satisfies the stated nonnegativity or alignment conditions. For constant-step bounds, replace $\mathbb{E}[D_t(N)] \le \overline{D}(N)$ by $\mathbb{E}[B_t] \le \overline{B}$, as justified by Theorem 3.

Heterogeneous vector error. Apply Proposition 2 conditional on $\mathcal{F}_t$ for each coordinate, with

$$q^{(n)}_{t,i} := \Pr(m^{(n)}_{t,i} \ge 0 \mid \mathcal{F}_t), \qquad \overline{q}_{t,i} := \frac{1}{N}\sum_{n=1}^{N} q^{(n)}_{t,i}.$$

Use the corresponding scalar definitions of $S_{\mathrm{eff},\epsilon_p,t,i}$, $\gamma_{\mathrm{eff},t,i}$, and $u_{\mathrm{eff},\epsilon_p,t,i}$. Set

$$\mathbf{v}^{\mathrm{eff}}_t := (u_{\mathrm{eff},\epsilon_p,t,i})_{i=1}^{d},$$

$$V_{t,i} := \mathrm{Var}(\widehat{q}_{t,i} \mid \mathcal{F}_t) = \frac{1}{N^2}\sum_{n=1}^{N} q^{(n)}_{t,i}(1 - q^{(n)}_{t,i}) \le \frac{1}{4N},$$

and define

$$B^{\mathrm{eff}}_t := \sum_{i=1}^{d}\left[2K^2_{\epsilon_p} V_{t,i} + 8a(S_{\mathrm{eff},\epsilon_p,t,i})^2 e^{-2N\gamma^2_{\mathrm{eff},t,i}}\right]. \tag{17}$$

Summing the coordinate bounds, without coordinate independence, gives

$$\mathbb{E}_t\|\widehat{\mathbf{u}}_t - \mathbf{v}^{\mathrm{eff}}_t\|_2^2 \le B^{\mathrm{eff}}_t. \tag{18}$$

Moreover,

$$0 \le B^{\mathrm{eff}}_t \le \frac{dK^2_{\epsilon_p}}{2N} + 8d, \qquad B^{\mathrm{eff}}_t \text{ is } \mathcal{F}_t\text{-measurable}.$$

Thus B^{eff}_t is integrable and $(\mathbf{v}^{\mathrm{eff}}_t, B^{\mathrm{eff}}_t)$ satisfies (16). For any $\mathcal{F}_t$-measurable RMS reference $\mathbf{w}_t$,

$$\mathbb{E}_t\|\widehat{\mathbf{u}}_t - \mathbf{w}_t\|_2^2 \le 2B^{\mathrm{eff}}_t + 2\|\mathbf{v}^{\mathrm{eff}}_t - \mathbf{w}_t\|_2^2, \tag{19}$$

by $\|x+y\|_2^2 \le 2\|x\|_2^2 + 2\|y\|_2^2$. The second term is RMS modeling discrepancy. Stabilization, LUT rounding, and update-norm clipping require error accounting in B_t and the same update norm bound; empirical cosine and scale MAE imply neither conditional error control nor alignment.

A.4 Nonconvex Convergence

Define the clipped-RMS stationarity functional

$$\mathcal{G}_{\epsilon_p}(\boldsymbol{\theta}_t) := \langle \nabla f(\boldsymbol{\theta}_t), \mathbf{u}^{\star}_{\epsilon_p,t} \rangle. \tag{20}$$

This history-dependent functional need not characterize Euclidean stationarity. Coordinate-wise sign alignment gives

$$y_{t,i} = \mathrm{sign}(\nabla_i f(\boldsymbol{\theta}_t)), \qquad \mathcal{G}_{\epsilon_p}(\boldsymbol{\theta}_t) = \sum_{i=1}^{d} |\nabla_i f(\boldsymbol{\theta}_t)| a(S_{\epsilon_p,t,i}) \ge 0. \tag{21}$$

Theorem 4 (Convergence in clipped-RMS stationarity)**.** *Assume the conditions of Theorem 3 hold for $t = 0, \dots, T-1$, $f \ge f_{\inf}$, and $\mathcal{G}_{\epsilon_p}(\boldsymbol{\theta}_t) \ge 0$ almost surely. Then*

$$\begin{aligned}\sum_{t=0}^{T-1} \eta_t \mathbb{E}[\mathcal{G}_{\epsilon_p}(\boldsymbol{\theta}_t)] &\le f(\boldsymbol{\theta}_0) - f_{\inf} \\ &\quad + \sum_{t=0}^{T-1} \eta_t \mathbb{E}\Big[\|\nabla f(\boldsymbol{\theta}_t)\|_2 \sqrt{D_t(N)}\Big] + \frac{Ld}{2}\sum_{t=0}^{T-1}\eta_t^2.\end{aligned} \tag{22}$$

Let

$$H_T := \sum_{t=0}^{T-1}\eta_t, \qquad \Pr(\tau_T = t) := \frac{\eta_t}{H_T},$$

where τ_T is sampled independently of the algorithmic randomness. If $\|\nabla f(\boldsymbol{\theta}_t)\|_2 \le G$ almost surely, then

$$\mathbb{E}[\mathcal{G}_{\epsilon_p}(\boldsymbol{\theta}_{\tau_T})] \le \frac{f(\boldsymbol{\theta}_0) - f_{\inf}}{H_T} + \frac{G}{H_T}\sum_{t=0}^{T-1}\eta_t \mathbb{E}[\sqrt{D_t(N)}] + \frac{Ld}{2H_T}\sum_{t=0}^{T-1}\eta_t^2. \tag{23}$$

Consequently, if

$$H_T \to \infty, \qquad \frac{\sum_{t<T}\eta_t^2}{H_T} \to 0, \qquad \frac{\sum_{t<T}\eta_t \mathbb{E}[\sqrt{D_t(N)}]}{H_T} \to 0, \tag{24}$$

then

$$\mathbb{E}[\mathcal{G}_{\epsilon_p}(\boldsymbol{\theta}_{\tau_T})] \to 0. \tag{25}$$

Proof. Total expectation and telescoping in Theorem 3 give (22). Divide by H_T and use the law of τ_T. □

Assumption 2 (Population-gradient alignment)**.** *There exists $\kappa > 0$ such that, almost surely for every t,*

$$\langle \nabla f(\boldsymbol{\theta}_t), \mathbf{u}^{\star}_{\epsilon_p,t} \rangle \ge \kappa \|\nabla f(\boldsymbol{\theta}_t)\|_2^2. \tag{26}$$

A sufficient coordinate-wise condition for Assumption 2 is

$$y_{t,i} = \operatorname{sign}(\nabla_i f(\boldsymbol{\theta}_t)), \qquad a(S_{\epsilon_p,t,i}) \ge \kappa |\nabla_i f(\boldsymbol{\theta}_t)| \quad \text{for all } i.$$

Theorem 5 (Euclidean stationarity under population-gradient alignment)**.** *Suppose the conditions of Theorem 3 and Assumption 2 hold, and let $f \ge f_{\inf}$. Then*

$$\begin{aligned}\frac{\kappa}{2}\sum_{t=0}^{T-1}\eta_t \mathbb{E}\|\nabla f(\boldsymbol{\theta}_t)\|_2^2 \le f(\boldsymbol{\theta}_0) - f_{\inf} \\ + \frac{1}{2\kappa}\sum_{t=0}^{T-1}\eta_t \mathbb{E}[D_t(N)] + \frac{Ld}{2}\sum_{t=0}^{T-1}\eta_t^2.\end{aligned} \tag{27}$$

Consequently,

$$\frac{1}{H_T}\sum_{t=0}^{T-1}\eta_t \mathbb{E}\|\nabla f(\boldsymbol{\theta}_t)\|_2^2 \le \frac{2(f(\boldsymbol{\theta}_0) - f_{\inf})}{\kappa H_T} + \frac{\sum_{t<T}\eta_t \mathbb{E}[D_t(N)]}{\kappa^2 H_T} + \frac{Ld\sum_{t<T}\eta_t^2}{\kappa H_T}. \tag{28}$$

Equivalently, for the step-size-weighted iterate τ_T,

$$\mathbb{E}\|\nabla f(\boldsymbol{\theta}_{\tau_T})\|_2^2 \le \frac{2(f(\boldsymbol{\theta}_0) - f_{\inf})}{\kappa H_T} + \frac{\sum_{t<T}\eta_t \mathbb{E}[D_t(N)]}{\kappa^2 H_T} + \frac{Ld\sum_{t<T}\eta_t^2}{\kappa H_T}. \tag{29}$$

Proof. Theorem 3 and Assumption 2 imply

$$\begin{aligned}\mathbb{E}_t[f(\boldsymbol{\theta}_{t+1})] \le f(\boldsymbol{\theta}_t) - \kappa\eta_t \|\nabla f(\boldsymbol{\theta}_t)\|_2^2 \\ + \eta_t \|\nabla f(\boldsymbol{\theta}_t)\|_2 \sqrt{D_t(N)} + \frac{L\eta_t^2 d}{2}.\end{aligned} \tag{30}$$

Young's inequality gives

$$\|\nabla f(\boldsymbol{\theta}_t)\|_2 \sqrt{D_t(N)} \le \frac{\kappa}{2}\|\nabla f(\boldsymbol{\theta}_t)\|_2^2 + \frac{D_t(N)}{2\kappa}. \tag{31}$$

Hence

$$\mathbb{E}_t[f(\boldsymbol{\theta}_{t+1})] \le f(\boldsymbol{\theta}_t) - \frac{\kappa\eta_t}{2}\|\nabla f(\boldsymbol{\theta}_t)\|_2^2 + \frac{\eta_t D_t(N)}{2\kappa} + \frac{L\eta_t^2 d}{2}. \tag{32}$$

After taking total expectations and rearranging,

$$\begin{aligned}\frac{\kappa\eta_t}{2}\mathbb{E}\|\nabla f(\boldsymbol{\theta}_t)\|_2^2 \le \mathbb{E}[f(\boldsymbol{\theta}_t) - f(\boldsymbol{\theta}_{t+1})] \\ + \frac{\eta_t}{2\kappa}\mathbb{E}[D_t(N)] + \frac{L\eta_t^2 d}{2}.\end{aligned} \tag{33}$$

Summation and $f(\boldsymbol{\theta}_T) \ge f_{\inf}$ prove (27); division by $\kappa H_T/2$ proves the random-iterate bound. □

Corollary 3 (Constant-step complexity and perturbation floor)**.** *Suppose the conditions of Theorem 5 hold,*

$$\eta_t = \eta, \qquad \mathbb{E}[D_t(N)] \le \overline{D}(N).$$

Then

$$\frac{1}{T}\sum_{t=0}^{T-1} \mathbb{E}\|\nabla f(\boldsymbol{\theta}_t)\|_2^2 \le \frac{2(f(\boldsymbol{\theta}_0) - f_{\inf})}{\kappa\eta T} + \frac{\overline{D}(N)}{\kappa^2} + \frac{Ld\eta}{\kappa}. \tag{34}$$

If $L > 0$ and $f(\boldsymbol{\theta}_0) > f_{\inf}$, the choice

$$\eta = \sqrt{\frac{2(f(\boldsymbol{\theta}_0) - f_{\inf})}{LdT}} \tag{35}$$

yields

$$\frac{1}{T}\sum_{t=0}^{T-1} \mathbb{E}\|\nabla f(\boldsymbol{\theta}_t)\|_2^2 \le \frac{2\sqrt{2Ld(f(\boldsymbol{\theta}_0) - f_{\inf})}}{\kappa\sqrt{T}} + \frac{\overline{D}(N)}{\kappa^2}. \tag{36}$$

Proof. Substitute into (28) and minimize the $1/\eta$ and η terms. The constants $\kappa, \overline{D}(N)$ must hold for the selected run, uniformly over candidates when comparing a common bound. The unoptimized inequality also covers zero initial gap or $L = 0$. □

Corollary 4 (Asymptotic first-order stationarity)**.** *Under the conditions of Theorem 5, suppose*

$$\sum_{t=0}^{\infty} \eta_t = \infty, \qquad \sum_{t=0}^{\infty} \eta_t^2 < \infty, \qquad \sum_{t=0}^{\infty} \eta_t \mathbb{E}[D_t(N)] < \infty. \tag{37}$$

Then

$$\sum_{t=0}^{\infty} \eta_t \mathbb{E}\|\nabla f(\boldsymbol{\theta}_t)\|_2^2 < \infty, \tag{38}$$

and

$$\liminf_{t\to\infty} \mathbb{E}\|\nabla f(\boldsymbol{\theta}_t)\|_2^2 = 0. \tag{39}$$

Proof. Let $T \to \infty$ in (27). The first claim follows from (37); the second follows from $\sum_t \eta_t = \infty$. □

Rate interpretation. Equation (36) separates an $\mathcal{O}(T^{-1/2})$ optimization term from the perturbation floor $\overline{D}(N)/\kappa^2$. Under the conditional homogeneous model,

$$\gamma(S_{t,i}) \ge \gamma_0 > 0 \ \ \forall(t,i) \quad \Longrightarrow \quad D_t(N) \le \frac{dK_{\epsilon_p}^2}{2N} + 8de^{-2N\gamma_0^2}.$$

The diagnostics establish no uniform gap. A uniform nonzero gap gives $D_t(N) \to 0$, but fixed clipping can retain bias relative to the unclipped RMS target. Zero signal and modeling or numerical components of B_t may also prevent general-reference error from vanishing as $N \to \infty$. Corollary 4 establishes weighted gradient summability and a zero liminf under its assumptions; it does not establish convergence of every iterate or an unconditional guarantee for the implemented optimizer.

B Language-Model Training and Evaluation Protocol

The language-model results in Section 5.2 and sweeps in Appendix J share the protocol below. Within each model–budget setting, architecture, data order, batch geometry, training horizon, precision, and evaluation are fixed; only the explicitly swept optimizer hyperparameters vary.

Table 3: Language-model architectures and $1\times$ training budgets.

Model	Layers	d_{model}	Heads	d_{ff}	Parameters	$1\times$ updates	$1\times$ input tokens
60M	8	512	8	1376	58,073,600	10,000	1,310,720,000
130M	12	768	12	2048	134,105,856	20,000	2,621,440,000
350M	24	1024	16	2736	367,969,280	60,000	7,864,320,000
1B	24	2048	32	5461	1,339,082,752	100,000	13,107,200,000

B.1 Model configurations and training budgets

We train decoder-only LLaMA-style models from scratch. All configurations use a vocabulary of 32,000, SiLU gated MLPs, RMSNorm with $\epsilon_{\text{RMS}} = 10^{-6}$, rotary positional embeddings, bias-free linear projections, and an initialization standard deviation of 0.02. Table 3 gives the architectures, exact parameter counts, and reference horizons. The $1\times$ input-token budget is the update count times the global sequence batch (512) and sequence length (256).

All optimization-quality experiments use $N = 8$ synchronized workers, with one worker process mapped to each GPU. Each worker processes one microbatch of 64 sequences of length 256 per optimizer update, and framework gradient accumulation is disabled. The resulting global sequence batch is therefore $8 \times 64 = 512$, corresponding to $512 \times 256 = 131{,}072$ input tokens per optimizer update. At each optimizer update, all distributed methods operate on the same collection of eight local microbatches; only the optimizer-specific statistic constructed from those local observations and the corresponding synchronization rule differ. Where $2\times$, $4\times$, or $8\times$ budgets are reported, we multiply the corresponding $1\times$ optimizer-update horizon, and hence the input-token budget, by that factor while keeping the model architecture, sequence length, and global batch fixed.

B.2 Data, tokenization, and example order

We use the English configuration of C4 (`allenai/c4`), with its training and validation splits kept separate. The corpus is pretokenized using the `t5-base` tokenizer. Documents are tokenized without automatically adding per-document special tokens; the tokenizer EOS token is appended after each nonempty document, and the resulting token stream is partitioned into consecutive, non-overlapping blocks of length 256. Within each model–budget setting, every optimizer uses the same pretokenized training corpus and stored example order, without online reshuffling. Validation uses the same fixed, unshuffled C4 validation prefix for every hyperparameter candidate.

B.3 Precision, learning-rate schedule, and randomness

All methods maintain FP32 master weights and use BF16 for forward and backward computation. Accordingly, labels such as "full precision" in our optimizer comparisons refer to uncompressed or full-value optimizer information, not to the shared master-weight dtype. For all methods, the learning rate increases linearly from zero to the method-specific peak during warmup, then decays linearly to zero at the final optimizer update. Peak learning rate, warmup length, weight decay, clipping, and other method-specific optimizer hyperparameters are selected by the sweeps reported in Appendix J. Unless stated otherwise, runs use seed 0. Python, NumPy, and PyTorch random-number generators are seeded before model construction, and the initialized model parameters are synchronized across workers before training begins.

B.4 Validation and hyperparameter selection

Validation uses a fixed target of approximately one million next-token predictions. Workers reduce their local negative-log-likelihood sums and valid-target counts before division. The mean validation NLL is

$$\mathcal{L}_{\text{val}} = \frac{\sum_n L_n}{\sum_n C_n},$$

where L_n and C_n denote the loss sum and valid-target count on worker n, respectively. Validation perplexity is then $\text{PPL} = \exp(\mathcal{L}_{\text{val}})$.

Every hyperparameter candidate is trained to the prescribed model–budget endpoint and compared using validation loss at that endpoint. For the coordinate sweeps, hyperparameter coordinates are visited in the reported order, and a candidate replaces the current incumbent only if it improves endpoint validation loss by at least 0.003. The candidate grids and completed runs are reported in Appendix J. For each method–model–budget setting, the summary table reports the lowest endpoint validation loss among all completed candidates, converted to perplexity as above.

B.5 Selected configurations for reported results

Tables 4–13 give the optimizer hyperparameters for every result in Table 2; all other settings follow the common protocol above. Here, clip denotes the method-specific global-norm clipping threshold: update clipping for AdaSignSGD and gradient clipping under each baseline's implementation. Zero disables norm clipping; all selected AdaSignSGD configurations use clip$= 0$. WU is the warmup length in updates, and WD is weight decay. Complete candidate grids and run records appear in Appendix J.

Table 4: Selected AdamW configurations.

Setting	β_1	β_2	ϵ	η	clip	WU	WD
60M–1×	0.87	0.98	10^{-10}	0.002	1.0	2000	0.1
60M–2×	0.87	0.98	10^{-10}	0.002	1.0	2000	0.1
60M–4×	0.87	0.98	10^{-8}	0.008	0	2000	0.1
60M–8×	0.87	0.98	10^{-10}	0.008	0	2000	0.5
130M–1×	0.87	0.98	10^{-8}	0.002	1.0	2000	0.1
130M–2×	0.87	0.98	10^{-8}	0.001	1.0	2000	0.1
130M–4×	0.87	0.98	10^{-8}	0.001	1.0	2000	0.1
130M–8×	0.87	0.98	10^{-8}	0.001	1.0	2000	0.1
350M–1×	0.87	0.98	10^{-20}	0.001	0.5	2000	0.1
350M–2×	0.87	0.98	10^{-20}	0.001	0	2000	0.1
350M–4×	0.87	0.98	10^{-20}	0.001	0.5	2000	0.1
350M–8×	0.87	0.98	10^{-20}	0.001	0	2000	0.1
1B–1×	0.87	0.98	10^{-10}	5×10^{-4}	0.5	2000	0.1

Table 5: Selected AdaSignSGD (uncompressed) configurations.

Setting	β	ε	η	clip	WU	WD
60M–1×	0.90	10^{-20}	0.002	0	2000	0
60M–2×	0.90	10^{-20}	0.002	0	2000	0
60M–4×	0.95	10^{-10}	0.008	0	2000	0.5
60M–8×	0.95	10^{-10}	0.002	0	2000	0
130M–1×	0.95	10^{-8}	0.002	0	2000	0.5
130M–2×	0.95	10^{-8}	0.002	0	2000	0.5
130M–4×	0.95	10^{-8}	0.002	0	2000	0.5
130M–8×	0.95	10^{-8}	0.002	0	2000	0.5
350M–1×	0.95	10^{-20}	0.001	0.5	2000	0
350M–2×	0.95	10^{-20}	0.001	0.5	2000	0
350M–4×	0.95	10^{-20}	0.001	0.5	2000	0
350M–8×	0.95	10^{-20}	0.001	0.5	2000	0
1B–1×	0.95	10^{-20}	5×10^{-4}	0.5	2000	0.5

Table 6: Selected Birder configurations.

Setting	β	η	clip	WU	WD
60M–1×	0.95	0.002	0	2000	0.1
60M–2×	0.95	0.002	0	2000	0.1
60M–4×	0.95	0.002	0	2000	0.1
60M–8×	0.98	0.001	0	2000	0.1
130M–1×	0.95	0.001	0	1000	0.1
130M–2×	0.95	0.001	0	2000	0.1
130M–4×	0.98	0.001	0	2000	0.1
130M–8×	0.98	0.001	0	2000	0.1
350M–1×	0.95	0.001	0	2000	0.1
350M–2×	0.98	0.001	0	2000	0.1
350M–4×	0.98	0.001	0	2000	0.1
350M–8×	0.95	0.001	0	2000	0.1
1B–1×	0.98	0.001	0	2000	0.1

Table 7: Selected Birder + EC configurations.

Setting	β	η	clip	WU	WD
60M–1×	0.95	0.002	0	2000	0.1
60M–2×	0.95	0.002	0	2000	0.1
60M–4×	0.95	0.002	0	2000	0.1
60M–8×	0.98	0.001	0	2000	0.1
130M–1×	0.95	0.001	0	1000	0.1
130M–2×	0.95	0.001	0	1000	0.1
130M–4×	0.98	0.001	0	2000	0.1
130M–8×	0.98	0.001	0	2000	0.1
350M–1×	0.95	0.001	0	2000	0.1
350M–2×	0.98	0.001	0	2000	0.1
350M–4×	0.98	0.001	0	2000	0.1
350M–8×	0.95	0.001	0	2000	0.1
1B–1×	0.98	0.001	0	2000	0.1

Table 8: Selected 1-Bit Adam configurations.

Setting	β_1	β_2	ϵ	η	clip	WU	WD	T_c/T
60M–1×	0.95	0.98	10^{-8}	0.001	1.0	1000	0.1	0.05
60M–2×	0.87	0.98	10^{-8}	0.002	0.5	2000	0.1	0.10
60M–4×	0.87	0.95	10^{-8}	0.001	1.0	1000	0.1	0.10
60M–8×	0.87	0.90	10^{-8}	0.001	1.0	1000	0.1	0.10
130M–1×	0.87	0.95	10^{-8}	0.001	0.5	1000	0.1	0.10
130M–2×	0.87	0.90	10^{-8}	0.001	0.5	1000	0.1	0.10
130M–4×	0.87	0.95	10^{-8}	0.001	0.5	1000	0.1	0.10
130M–8×	0.87	0.95	10^{-8}	0.001	0.5	1000	0.1	0.10
350M–1×	0.98	0.98	10^{-8}	5×10^{-4}	0.5	1000	0.1	0.10
350M–2×	0.98	0.98	10^{-8}	5×10^{-4}	0.5	1000	0.1	0.10
350M–4×	0.98	0.98	10^{-8}	5×10^{-4}	0.5	1000	0.1	0.10
350M–8×	0.98	0.98	10^{-8}	5×10^{-4}	0.5	1000	0.1	0.10
1B–1×	0.98	0.98	10^{-8}	2.5×10^{-4}	0.5	1000	0.1	0.10

Table 9: Selected 0/1 Adam configurations.

Setting	β_1	β_2	ϵ	η	clip	WU	WD	L	T_v
60M–1×	0.98	0.999	10^{-8}	0.008	0.5	2000	0.1	4	8
60M–2×	0.98	0.999	10^{-8}	0.008	0.5	2000	0.1	4	8
60M–4×	0.98	0.999	10^{-8}	0.008	0.5	2000	0.1	4	8
60M–8×	0.98	0.999	10^{-8}	0.008	0.5	2000	0.1	4	8
130M–1×	0.98	0.999	10^{-8}	0.008	0.5	2000	0.1	4	8
130M–2×	0.98	0.999	10^{-8}	0.008	0.5	2000	0.1	4	8
130M–4×	0.98	0.999	10^{-8}	0.008	0.5	2000	0.1	4	8
130M–8×	0.98	0.999	10^{-8}	0.008	0.5	2000	0.1	4	8
350M–1×	0.98	0.98	10^{-8}	0.008	2.0	2000	0.1	4	16
350M–2×	0.98	0.98	10^{-8}	0.008	2.0	2000	0.1	4	16
350M–4×	0.98	0.98	10^{-8}	0.004	2.0	2000	0.1	4	8
350M–8×	0.98	0.98	10^{-8}	0.004	2.0	2000	0.1	4	8
1B–1×	0.98	0.98	10^{-8}	0.002	2.0	2000	0.1	4	16

Table 10: Selected SIGNUM-MV configurations.

Setting	β	η	clip	WU	WD
60M–1×	0.98	5×10^{-4}	0	2000	0.1
60M–2×	0.98	1.25×10^{-4}	0	2000	0.1
60M–4×	0.95	1.25×10^{-4}	0	1000	0
60M–8×	0.98	1.25×10^{-4}	0	1000	0.1
130M–1×	0.98	1.25×10^{-4}	0	2000	0.5
130M–2×	0.98	1.25×10^{-4}	0	1000	0
130M–4×	0.98	5×10^{-4}	0	1000	0.5
130M–8×	0.98	5×10^{-4}	0	2000	0.5
350M–1×	0.98	2.5×10^{-4}	0	2000	0.5
350M–2×	0.98	2.5×10^{-4}	0	1000	0.5
350M–4×	0.98	1.25×10^{-4}	0	1000	0.1
350M–8×	0.98	1.25×10^{-4}	0	1000	0.5
1B–1×	0.98	1.25×10^{-4}	0	2000	0.5

Table 11: Selected SIGNUM-MV+EF configurations.

Setting	β	η	clip	WU	WD
60M–1×	0.87	4.096	0	2000	0
60M–2×	0.87	4.096	0	2000	0
60M–4×	0.87	4.096	0	2000	0
60M–8×	0.87	4.096	1.0	2000	0
130M–1×	0.87	4.096	0	1000	0
130M–2×	0.87	4.096	0	2000	0
130M–4×	0.87	4.096	0	2000	0
130M–8×	0.87	4.096	0	2000	0
350M–1×	0.95	4.096	0	1000	0
350M–2×	0.95	4.096	0	2000	0
350M–4×	0.95	4.096	0	2000	0
350M–8×	0.95	4.096	0	2000	0
1B–1×	0.95	4.096	0	2000	0

Table 12: Selected D-Lion-MaVo configurations.

Setting	β_1	β_2	η	clip	WU	WD
60M–1×	0.98	0.98	2.5×10^{-4}	0	1000	0
60M–2×	0.98	0.98	1.25×10^{-4}	0	2000	0
60M–4×	0.98	0.98	1.25×10^{-4}	0	2000	0
60M–8×	0.98	0.98	2.5×10^{-4}	0	1000	0
130M–1×	0.95	0.98	2.5×10^{-4}	0	1000	0
130M–2×	0.95	0.98	2.5×10^{-4}	0	1000	0
130M–4×	0.95	0.98	2.5×10^{-4}	0	1000	0
130M–8×	0.95	0.98	2.5×10^{-4}	0	1000	0
350M–1×	0.98	0.98	1.25×10^{-4}	0	1000	1.0
350M–2×	0.98	0.98	1.25×10^{-4}	0	1000	1.0
350M–4×	0.98	0.98	6.25×10^{-5}	0	1000	1.0
350M–8×	0.98	0.98	1.25×10^{-4}	0	1000	0.5
1B–1×	0.98	0.98	6.25×10^{-5}	0	1000	1.0

Table 13: Selected ADASIGNSGD (ours) configurations.

Setting	β	ε	ϵ_p	η	clip	WU	WD
60M–1×	0.95	10^{-8}	5×10^{-3}	0.001	0	2000	0.5
60M–2×	0.95	10^{-10}	10^{-4}	0.001	0	2000	0.5
60M–4×	0.98	10^{-8}	10^{-4}	5×10^{-4}	0	2000	0.5
60M–8×	0.98	10^{-20}	10^{-4}	5×10^{-4}	0	2000	0.5
130M–1×	0.95	10^{-8}	10^{-4}	5×10^{-4}	0	2000	0.5
130M–2×	0.95	10^{-8}	10^{-4}	5×10^{-4}	0	2000	0.5
130M–4×	0.98	10^{-8}	10^{-4}	5×10^{-4}	0	2000	0.5
130M–8×	0.98	10^{-8}	10^{-4}	5×10^{-4}	0	2000	0.5
350M–1×	0.95	10^{-10}	10^{-4}	5×10^{-4}	0	2000	0.5
350M–2×	0.95	10^{-10}	10^{-4}	5×10^{-4}	0	2000	0.5
350M–4×	0.95	10^{-10}	10^{-4}	2.5×10^{-4}	0	2000	0.5
350M–8×	0.95	10^{-10}	10^{-4}	2.5×10^{-4}	0	2000	0.5
1B–1×	0.95	10^{-8}	10^{-4}	2.5×10^{-4}	0	2000	0.5

C FROZEN-CHECKPOINT DIAGNOSTICS OF SIGN AGREEMENT

This section provides the full experimental protocol and checkpoint-level results for Section 3. We test whether EMA momentum is sufficiently Gaussian for the probit model, whether sign agreement recovers its full-value SNR and bounded scale, and how accurately a finite number of worker signs recovers the resulting update. Throughout, we reuse the notation and definitions from the main text.

C.1 EXPERIMENTAL PROTOCOL

We train 60M, 130M, and 350M LLaMA-style causal language models on pre-tokenized C4 with sequence length 256. At each selected checkpoint, we fix $\boldsymbol{\theta}_t$ and sample fresh mini-batches, isolating stochastic gradient variation from parameter drift. We collect raw gradients and exponential moving average (EMA) momenta from the first, middle, and last transformer blocks and the seven attention and MLP projection tensors `q_proj`, `k_proj`, `v_proj`, `o_proj`, `gate_proj`, `up_proj`, and `down_proj`, for 21 tensors per checkpoint.

For each tensor, we uniformly sample $C = 2048$ coordinates. After 32 diagnostic EMA burn-in updates with $\beta = 0.9$, we collect $K = 128$ paired raw-gradient samples $g_{k,i}$ and momentum samples $m_{k,i}$. For momentum, we compute $\widehat{\mu}_i$, $\widehat{\sigma}_i$, S_i^{full}, $\widehat{p}_i$, and S_i^{sign} exactly as in Equations 1–2. We repeat the same computation for raw gradients by replacing $m_{k,i}$ with $g_{k,i}$; because every figure and table identifies the signal type, the same symbols denote the corresponding raw-gradient quantities. Finally, write $a_i^{\text{full}} := a(S_i^{\text{full}})$ and $a_i^{\text{sign}} := a(S_i^{\text{sign}})$ for the bounded map defined in Section 4. Table 14 lists the checkpoints and their training losses. In Figures 9–12, rows correspond to steps 50, 200, and 1000, from top to bottom. The three-panel figures order models as 60M, 130M, and 350M.

Table 14: Frozen-checkpoint diagnostic configuration and training losses.

Model	Params (M)	Checkpoints	Checkpoint losses	Seq. len.	K	Coords/ tensor C	Tensors
60M	58.1	$\{50, 200, 1000\}$	8.62/6.80/5.82	256	128	2048	21
130M	134.1	$\{50, 200, 1000\}$	7.93/6.72/5.74	256	128	2048	21
350M	368.0	$\{50, 200, 1000\}$	7.61/6.66/5.70	256	128	2048	21

Table 15: Mean standardized-residual shape and tail statistics.

Model	Signal type	Std(z)	Excess kurtosis	Skewness	$\Pr(\lvert z\rvert > 2)$	$\Pr(\lvert z\rvert > 3)$
60M	Raw gradient	0.994	5.550	0.009	0.047	0.010
60M	EMA momentum	0.996	0.182	-0.002	0.044	0.003
130M	Raw gradient	0.993	8.613	-0.040	0.045	0.009
130M	EMA momentum	0.996	0.368	-0.003	0.044	0.005
350M	Raw gradient	0.991	12.138	-0.056	0.044	0.009
350M	EMA momentum	0.995	0.617	-0.003	0.045	0.005

Table 16: Mean coordinate-wise calibration errors and full-value SNR.

Model	Signal type	$\lvert\Phi(S_i^{\text{full}}) - \widehat{p}_i\rvert$	$\lvert S_i^{\text{sign}} - S_i^{\text{full}}\rvert$	$\lvert a_i^{\text{sign}} - a_i^{\text{full}}\rvert$	Mean S_i^{full}
60M	Raw gradient	0.026	0.054	0.052	0.105
60M	EMA momentum	0.034	0.101	0.067	0.505
130M	Raw gradient	0.026	0.054	0.051	0.111
130M	EMA momentum	0.036	0.110	0.069	0.523
350M	Raw gradient	0.027	0.055	0.053	0.110
350M	EMA momentum	0.039	0.122	0.075	0.524

C.2 Gaussianity of gradients and momenta

We center and scale each coordinate using its empirical $\widehat{\mu}_i$ and $\widehat{\sigma}_i$, and denote the resulting standardized residuals by z. Figure 8 reports their excess kurtosis and tail probability at each checkpoint, and Table 15 averages the shape statistics over checkpoints and tensors. EMA substantially reduces heavy tails: model-level mean excess kurtosis falls from 5.55–12.14 for raw gradients to 0.18–0.62 for momenta. The standard-Gaussian references are zero excess kurtosis and skewness, $\Pr(|z| > 2) \approx 0.0455$, and $\Pr(|z| > 3) \approx 0.0027$.

C.3 Agreement, SNR, and scale calibration

Figure 9 compares the empirical agreement probability $\widehat{p}_i$ with the Gaussian prediction $\Phi(S_i^{\text{full}})$. Raw gradients occupy a relatively low-SNR regime, whereas EMA momenta span a wider range while remaining close to the prediction at every checkpoint.

Figure 10 compares the corresponding sign-only estimate S_i^{sign} with S_i^{full}. They track closely across models and checkpoints; the sparsely populated highest-SNR bins are the most affected by finite-sample clipping.

Applying the same bounded map to both estimates gives a_i^{sign} and a_i^{full}. Figure 11 shows that the former closely tracks the latter throughout training, directly evaluating the adaptive scale used by AdaSignSGD rather than absolute gradient magnitude.

Tables 16 and 17 report calibration errors averaged over the 21 tensors per checkpoint; Table 16 additionally averages over the three checkpoints. Raw gradients have smaller errors within their lower-SNR range, whereas EMA momenta remain well calibrated over the broader range used by AdaSignSGD. Lower errors indicate better calibration.

Table 17: Mean coordinate-wise calibration errors by checkpoint.

Model	Step	Signal type	$\lvert\Phi(S_i^{\mathrm{full}}) - \widehat{p}_i\rvert$	$\lvert S_i^{\mathrm{sign}} - S_i^{\mathrm{full}}\rvert$	$\lvert a_i^{\mathrm{sign}} - a_i^{\mathrm{full}}\rvert$
60M	50	Raw gradient	0.029	0.060	0.057
60M	50	EMA momentum	0.039	0.120	0.074
60M	200	Raw gradient	0.024	0.053	0.050
60M	200	EMA momentum	0.030	0.096	0.059
60M	1000	Raw gradient	0.024	0.049	0.048
60M	1000	EMA momentum	0.033	0.086	0.067
130M	50	Raw gradient	0.031	0.060	0.058
130M	50	EMA momentum	0.045	0.135	0.084
130M	200	Raw gradient	0.023	0.053	0.050
130M	200	EMA momentum	0.029	0.110	0.056
130M	1000	Raw gradient	0.023	0.048	0.047
130M	1000	EMA momentum	0.033	0.084	0.067
350M	50	Raw gradient	0.034	0.063	0.060
350M	50	EMA momentum	0.057	0.161	0.102
350M	200	Raw gradient	0.024	0.055	0.051
350M	200	EMA momentum	0.029	0.120	0.056
350M	1000	Raw gradient	0.023	0.048	0.047
350M	1000	EMA momentum	0.032	0.084	0.066

Table 18: Finite-worker scale error, direction accuracy, and vote statistics.

Model	N	Mean scale error	Median scale error	Direction accuracy	Mean LUT scale	Mean majority support
60M	4	0.311	0.281	0.729	0.521	0.753
60M	8	0.223	0.189	0.781	0.470	0.722
60M	16	0.167	0.136	0.827	0.436	0.702
60M	32	0.131	0.105	0.863	0.415	0.691
130M	4	0.309	0.279	0.733	0.526	0.756
130M	8	0.221	0.188	0.786	0.477	0.726
130M	16	0.168	0.136	0.829	0.444	0.706
130M	32	0.132	0.106	0.863	0.423	0.695
350M	4	0.310	0.282	0.731	0.528	0.758
350M	8	0.224	0.190	0.782	0.480	0.727
350M	16	0.171	0.138	0.824	0.446	0.708
350M	32	0.135	0.108	0.858	0.425	0.696

C.4 FINITE-WORKER RESAMPLING

At each checkpoint, we randomly draw N stored EMA momentum samples and treat them as worker momenta $m_i^{(n)}$, $n = 1, \ldots, N$. We then apply the exact count, majority direction $\widehat{y}_i$, and LUT scale $\widehat{a}_i$ from Algorithm 1, including its tie convention and numerical stabilizer. This single-GPU resampling isolates finite-worker sampling from the empirical marginal momentum distribution; it does not model cross-worker or temporal correlations in distributed training.

We measure scale error against the full-value target a_i^{full} and direction accuracy against $\mathrm{sgn}_+(\widehat{\mu}_i)$. Figure 12 shows that both improve with N at every checkpoint. At small N, choosing the observed majority biases its apparent support upward on low-SNR coordinates; consequently, the mean majority support and LUT scale in Table 18 decrease with N even as recovery improves.

Across all checkpoints, increasing N from 4 to 32 reduces mean scale error by 57.9%, 57.4%, and 56.5% for the 60M, 130M, and 350M models, respectively. The corresponding direction-accuracy gains are 0.134, 0.131, and 0.127. Table 18 averages over checkpoints $\{50, 200, 1000\}$; Table 19 reports each checkpoint, with error reduction measured relative to $N = 4$.

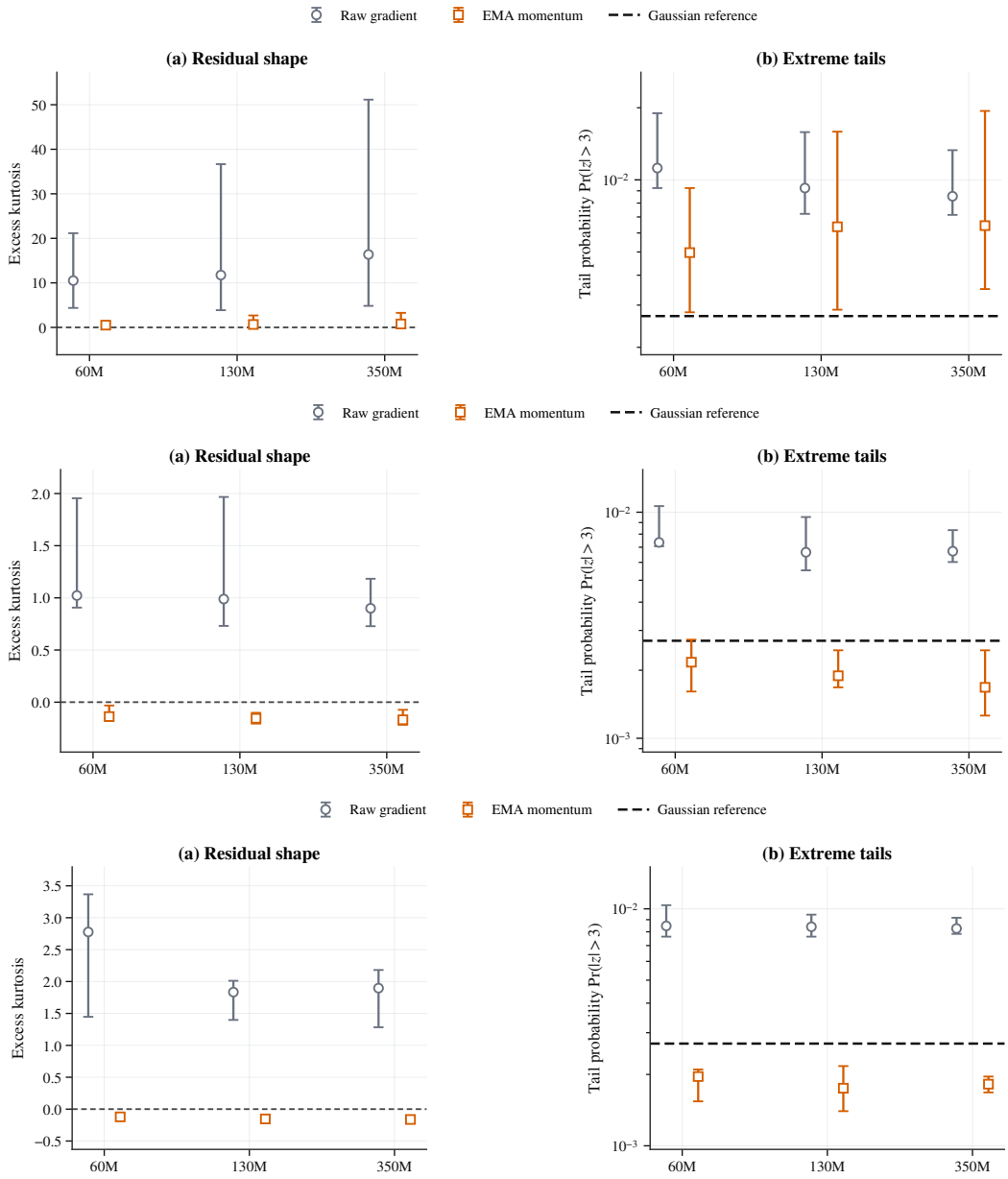

Figure 8: Gaussianity of gradients and EMA momenta. Left: excess kurtosis; right: $\Pr(|z| > 3)$. Points and error bars show tensor medians and interquartile ranges; dashed lines show standard-Gaussian references.

Table 19: Finite-worker recovery at $N = 4$ and $N = 32$ by checkpoint.

Model	Step	Scale error $N = 4$	Scale error $N = 32$	Error reduction (%)	Accuracy $N = 4$	Accuracy $N = 32$
60M	50	0.305	0.131	56.9	0.742	0.869
60M	200	0.301	0.125	58.5	0.753	0.883
60M	1000	0.328	0.137	58.2	0.692	0.837
130M	50	0.311	0.139	55.3	0.735	0.857
130M	200	0.286	0.118	58.7	0.777	0.898
130M	1000	0.329	0.137	58.3	0.685	0.835
350M	50	0.324	0.153	52.6	0.719	0.838
350M	200	0.277	0.115	58.5	0.787	0.903
350M	1000	0.329	0.137	58.5	0.688	0.834

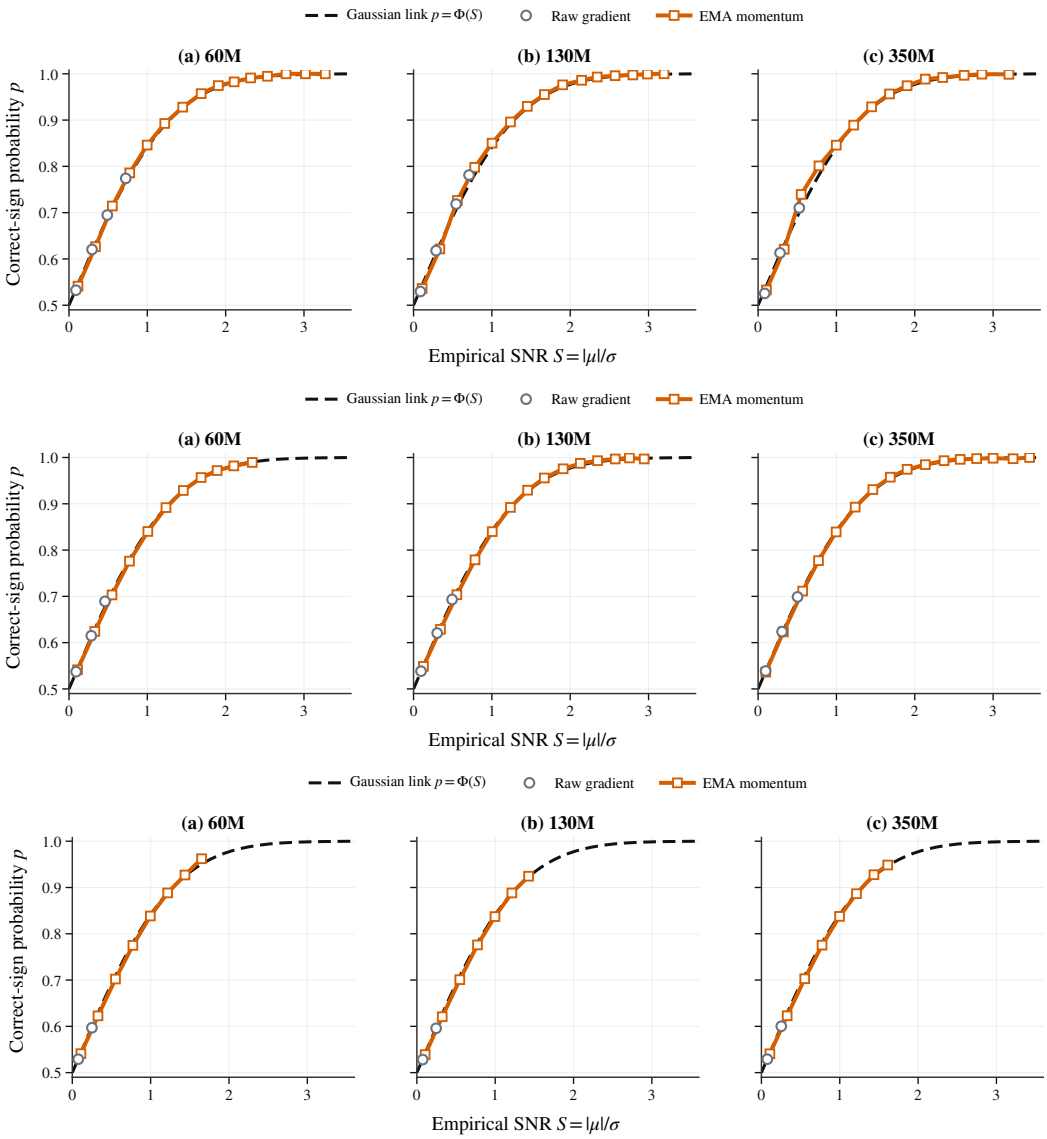

Figure 9: Agreement calibration across training. Binned $\widehat{p}_i$ versus S_i^{full}; dashed curves show $p = \Phi(S)$.

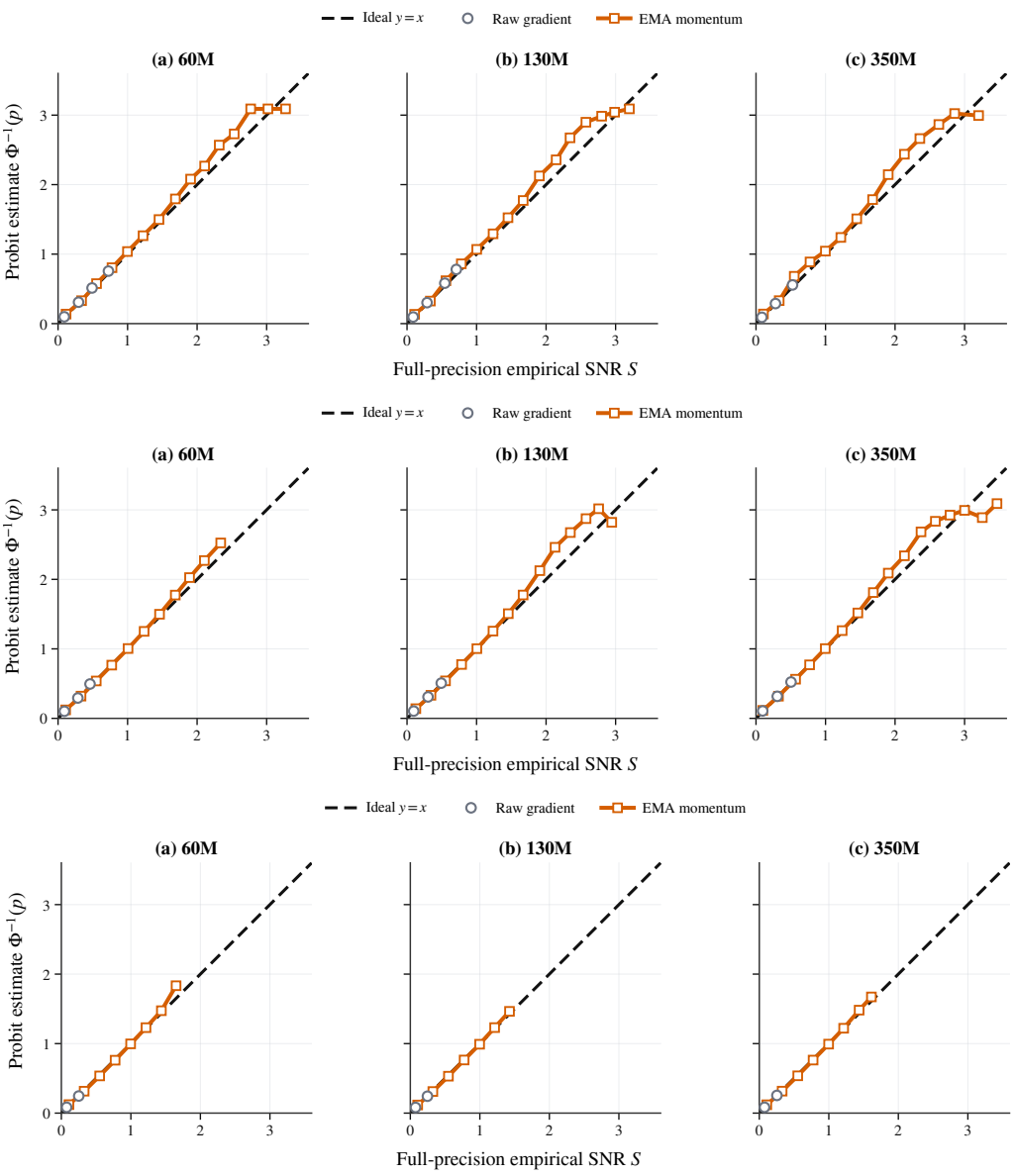

Figure 10: Sign-only SNR recovery across training. S_i^{sign} versus S_i^{full}; dashed lines show $y = x$.

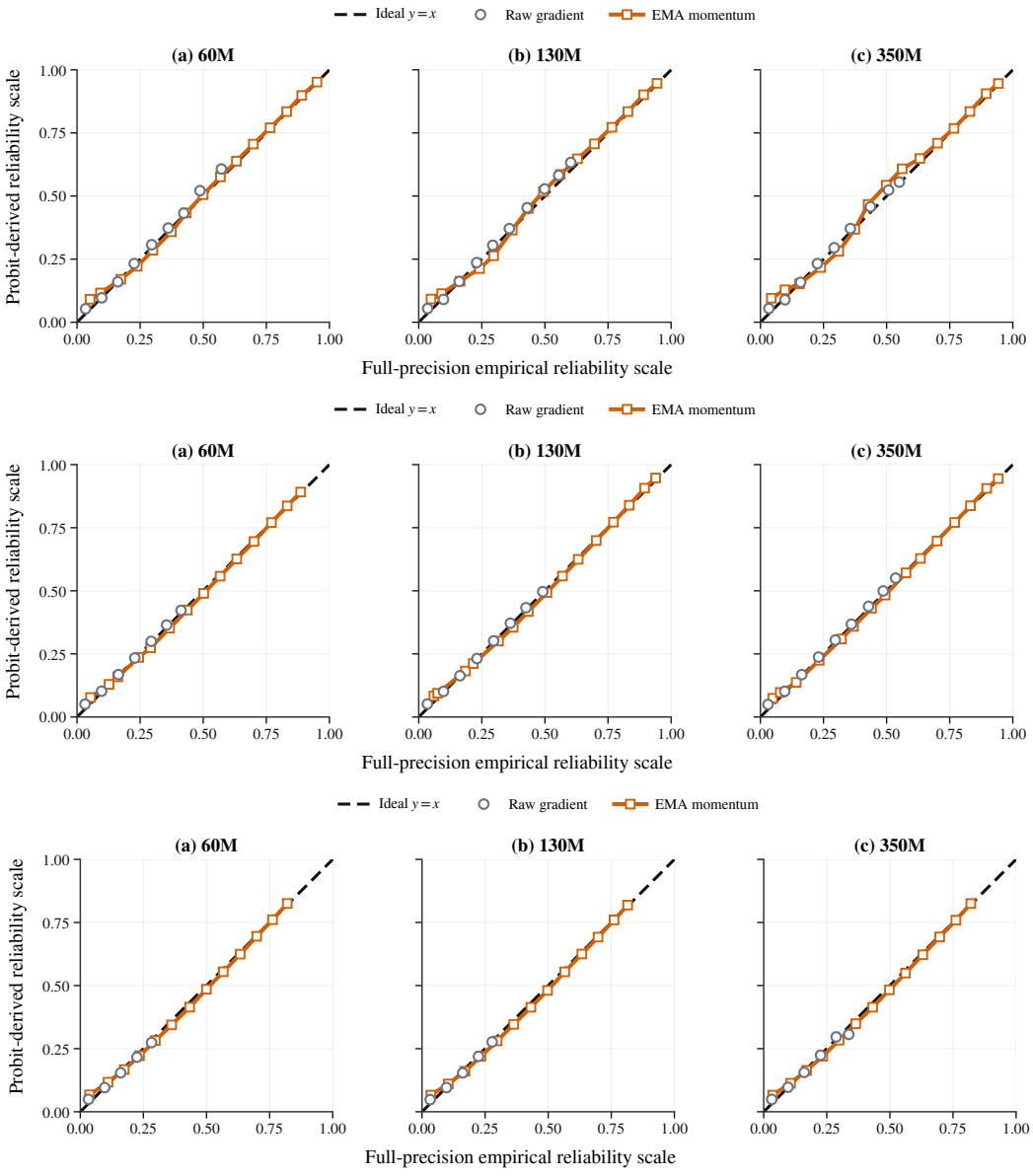

Figure 11: Adaptive-scale recovery across training. a_i^{sign} versus a_i^{full}; dashed lines show $y = x$.

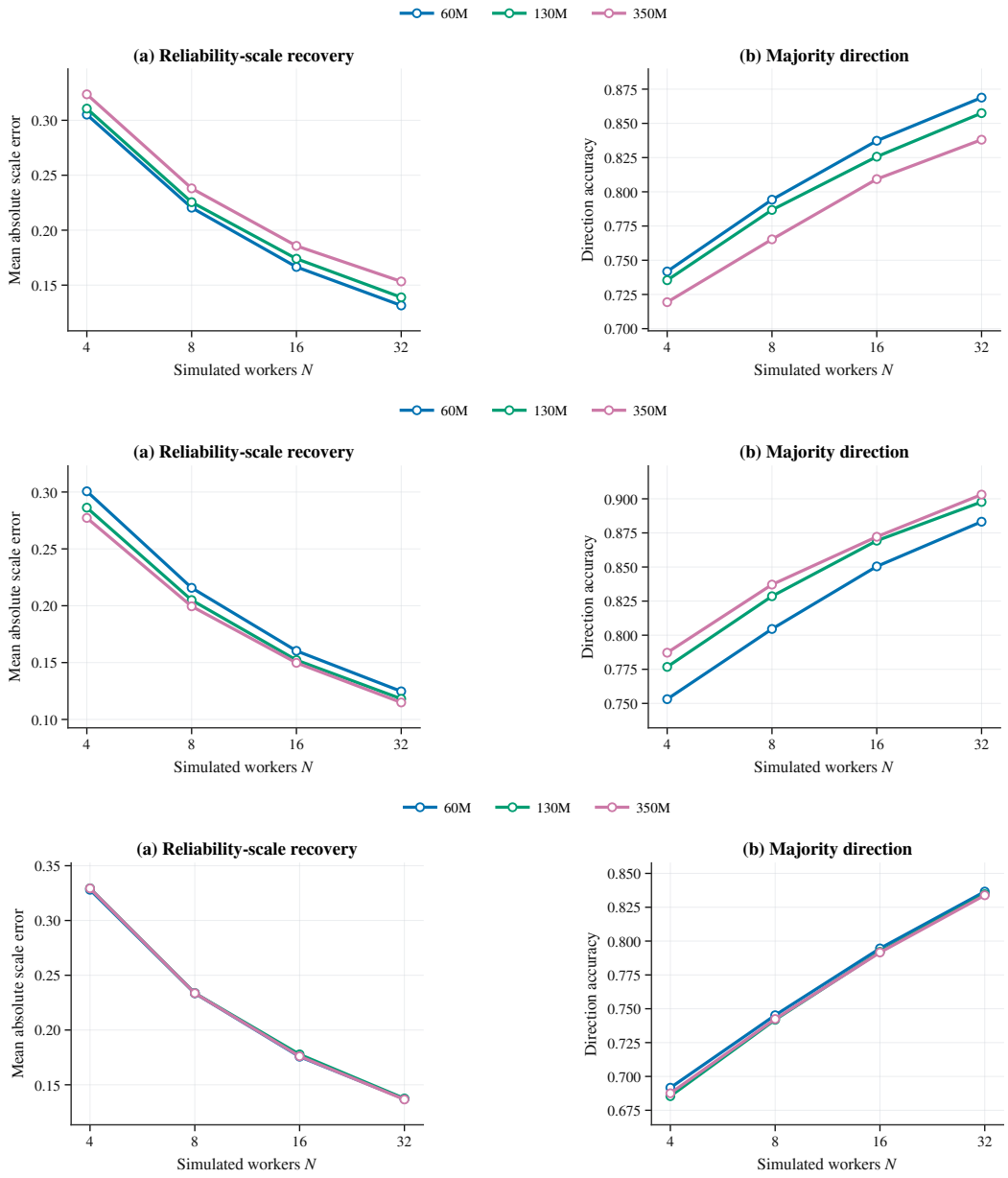

Figure 12: Finite-worker recovery. Left: mean $|\widehat{a}_i - a_i^{\text{full}}|$; right: the fraction of coordinates for which $\widehat{y}_i = \text{sgn}_+(\widehat{\mu}_i)$.

D TRAINING-TRAJECTORY DIAGNOSTICS OF THE GAUSSIAN WORKER MODEL

D.1 EXPERIMENTAL DESIGN

Models and training horizons. We train the 60M, 130M, and 350M LLaMA-style models with their $1\times$ ADASIGNSGD configurations, independently at each $N \in \{4, 8, 16, 32, 64, 128\}$. Each worker maintains a persistent BF16 momentum $\mathbf{m}_t^{(n)}$, data slice, and random-number stream. With global batch 512, the per-worker batch is

$$B = \frac{512}{N} \in \{128, 64, 32, 16, 8, 4\}. \tag{40}$$

All N use the model-specific hyperparameters selected at $N = 8$ in the main study, without diagnostic-specific tuning: sequence length 256, 2,000 warmup steps, clipping disabled, and weight decay 0.5. Table 20 gives the remaining configuration; tokens and data passes exclude reset probes.

Sampling schedule. We sample every 500 completed optimizer steps:

$$\begin{aligned} &\text{60M: } t \in \{500, 1000, \dots, 10000\}, \\ &\text{130M: } t \in \{500, 1000, \dots, 20000\}, \\ &\text{350M: } t \in \{500, 1000, \dots, 60000\}. \end{aligned} \tag{41}$$

All 18 model–worker-count runs are complete, contributing 20, 40, or 120 samples per 60M, 130M, or 350M trajectory, respectively. The first four samples, through $t = 2000$, cover warmup; the remainder span early, middle, and late post-warmup training. No aggregate includes a partial trajectory.

Fixed coordinate set. We sample 128 coordinates from each query, key, value, output, gate, up, and down projection matrix in the first, middle, and last transformer blocks, yielding

$$3 \text{ blocks} \times 7 \text{ tensors/block} \times 128 = 2688 \tag{42}$$

fixed coordinates per model. Using seed zero, we sample uniformly without replacement within each tensor before examining diagnostic results. Selection is independent of N and t, so every worker-count setting uses the same coordinates for a given model.

Persistent trajectory state. Following Section 4, we denote the saved post-update state at completed step t by $\boldsymbol{\theta}_t$ and $\{\mathbf{m}_t^{(n)}\}_{n=1}^N$. The sampler stores

$$\left\{ m_{t,i}^{(n)} : n = 1, \dots, N \right\}. \tag{43}$$

for each selected coordinate i. This *persistent trajectory state* comes directly from the EMA recursion in Algorithm 1 during training, without regeneration from a frozen checkpoint.

Reset one-step conditional probes. At each saved state, we draw $K = 128$ fresh probe worlds. For probe $k \in \{1, \dots, K\}$ and worker n, a fresh training mini-batch at fixed $\boldsymbol{\theta}_t$ produces gradient $\widetilde{\mathbf{g}}_{t+1}^{(n,k)}$ and candidate momentum

$$\widetilde{\mathbf{m}}_{t+1}^{(n,k)} = \beta \mathbf{m}_t^{(n)} + (1-\beta)\widetilde{\mathbf{g}}_{t+1}^{(n,k)}. \tag{44}$$

Table 20: Training horizons, sampled layers, and optimizer settings.

Model	Params	Steps	Checkpoints	Tokens	Data passes
60M	58.1M	10,000	20	1.311B	0.064
130M	134.1M	20,000	40	2.621B	0.128
350M	368.0M	60,000	120	7.864B	0.384

Model	Layers	β	ε	ϵ_p	Peak η
60M	$\{0, 4, 7\}$	0.95	10^{-8}	5×10^{-3}	10^{-3}
130M	$\{0, 6, 11\}$	0.95	10^{-8}	10^{-4}	5×10^{-4}
350M	$\{0, 12, 23\}$	0.95	10^{-10}	10^{-4}	5×10^{-4}

The two BF16 operations match training. Each probe restarts from the same $\mathbf{m}_t^{(n)}$; probe $k+1$ never inherits $\widetilde{\mathbf{m}}_{t+1}^{(n,k)}$. Probes leave parameters, persistent momenta, optimizer counters, the scheduler, model buffers, and the training data cursor unchanged. Parameters, momenta, buffers, random-number states, and scheduler state are checked before and after each collection, ensuring that probes do not alter the training trajectory.

Probe mini-batches are sampled with replacement from the same finite C4 training set through a separately seeded, reconstructible index stream. Each (t, k) world has an independently generated global batch of 512 examples, partitioned across N workers. Normal training uses sequential global batches with modulo cycling. The dataset contains 80 million packed examples; the 60M, 130M, and 350M horizons consume 0.064, 0.128, and 0.384 equivalent passes, respectively. No training trajectory completes a data cycle; replacement occurs only in the separate probe stream.

Only two BF16 arrays are retained at each sampled step: the persistent state in Equation 43, with shape $[N, 2688]$, and the reset one-step probes in Equation 44, with shape $[128, N, 2688]$. We do not retain gradients, input batches, logits, losses from probes, model checkpoints, full-model momenta, or continuation trajectories. The stored BF16 bit patterns are decoded to their FP32 values and all aggregate statistics are accumulated in FP64.

D.2 Analysis and results

D.2.1 Approximate conditional Gaussianity

We distinguish the realized cross-worker distribution at fixed (t, i), sampled over n in Equation 43, from the one-step conditional distribution at fixed (t, n, i), sampled over k in Equation 44. The former describes the current worker ensemble; the latter describes a worker's next momentum under fresh mini-batch randomness conditional on its actual history.

For a generic diagnostic group $x_1, \ldots, x_J$, let

$$\begin{aligned} \widetilde{\mu}_x &= \frac{1}{J}\sum_{j=1}^{J} x_j, \\ \widetilde{\sigma}_x^2 &= \frac{1}{J}\sum_{j=1}^{J}(x_j - \widetilde{\mu}_x)^2, \\ z_j &= \frac{x_j - \widetilde{\mu}_x}{\widetilde{\sigma}_x}. \end{aligned} \tag{45}$$

Zero-variance groups are recorded and omitted only from shape statistics that require division by $\widetilde{\sigma}_x$. We compute bias-corrected sample skewness.

We also compare the empirical CDF of z_j at the fixed threshold set

$$\mathcal{Z} = \{-3, -2.5, -2, \ldots, 2, 2.5, 3\} \tag{46}$$

with the corresponding matched-null CDF. The reported maximum gap is

$$D_{\text{state}} = \max_{z\in\mathcal{Z}} \left|\widetilde{F}_{\text{state}}(z) - F_{G,N}^{\text{fit}}(z)\right| \tag{47}$$

for persistent cross-worker states. Here $F_{G,N}^{\text{fit}}$ is obtained by repeatedly drawing N standard-normal observations and fitting their mean and variance internally in exactly the same way.

For the one-step conditional distribution, we use four-fold cross-fitting over whole probe worlds. In each fold, 96 probes fit a worker-coordinate mean and sample standard deviation, and the remaining 32 probes are standardized using that independent fit. The roles rotate so that every probe is evaluated once. The corresponding statistic is

$$D_{\text{cond}} = \max_{z\in\mathcal{Z}} \left|\widetilde{F}_{\text{cond}}(z) - F_{G,96}^{\text{heldout}}(z)\right|, \tag{48}$$

where the matched null draws a fresh standard-normal observation and standardizes it using the mean and sample standard deviation of 96 independent standard-normal fit observations. This null

Table 21: Persistent-state and conditional Gaussian-shape diagnostics. CDF gaps use 13 fixed thresholds and finite-sample matched nulls; ΔT_3 is the conditional excess tail probability.

		Persistent state		Reset one-step probes		
Model	N	Skew.	$10^3 D_{\mathrm{state}}$	Skew.	$10^3 D_{\mathrm{cond}}$	$10^3 \Delta T_3$
60M	4	-0.0076	1.220	0.0014	4.544	1.233
	8	0.0056	1.220	-0.0004	3.438	1.093
	16	0.0016	1.366	0.0002	3.641	1.227
	32	-0.0000	2.135	-0.0001	4.688	1.649
	64	-0.0006	2.534	0.0002	6.761	2.422
	128	-0.0007	2.497	0.0009	9.226	3.301
130M	4	-0.0043	0.862	-0.0002	3.031	0.930
	8	0.0054	1.043	0.0009	2.863	0.946
	16	-0.0002	0.982	0.0013	3.520	1.211
	32	0.0009	1.805	0.0008	4.797	1.710
	64	-0.0002	2.170	0.0008	6.579	2.375
	128	0.0007	1.968	0.0004	8.534	3.101
350M	4	0.0023	1.426	0.0021	9.130	2.256
	8	-0.0018	2.141	0.0001	7.056	2.209
	16	0.0017	1.949	-0.0004	7.531	2.527
	32	0.0009	2.738	-0.0003	8.333	2.879
	64	-0.0004	3.534	-0.0001	10.788	3.726
	128	-0.0026	4.453	-0.0009	14.161	4.725

includes finite-fit uncertainty; it is not the infinite-sample standard-normal CDF. We analogously report the excess held-out probability beyond three fitted standard deviations, ΔT_3.

Across all 18 settings, persistent-state signed skewness is -0.0076–0.0056 and the finite-N matched-null CDF gap is 0.00086–0.00445 (Table 21). Thus the realized worker ensemble is close to a fitted Gaussian at the tested thresholds. Shape agreement implies neither identical conditional means nor independence.

For reset one-step probes, signed skewness is -0.00093–0.00210, the held-out gap is $D_{\mathrm{cond}} = 0.00286$–$0.01416$, and excess probability beyond three fitted standard deviations is 0.00093–0.00473. Deviations tend to grow with smaller per-worker batch $B = 512/N$ and are larger for 350M. Persistent-state statistics aggregate over all checkpoints and coordinates; probe statistics use 128 observations per worker-coordinate.

Figure 13 places these shape results beside the sign-calibration, heterogeneity, and covariance measurements for the complete 18-setting matrix.

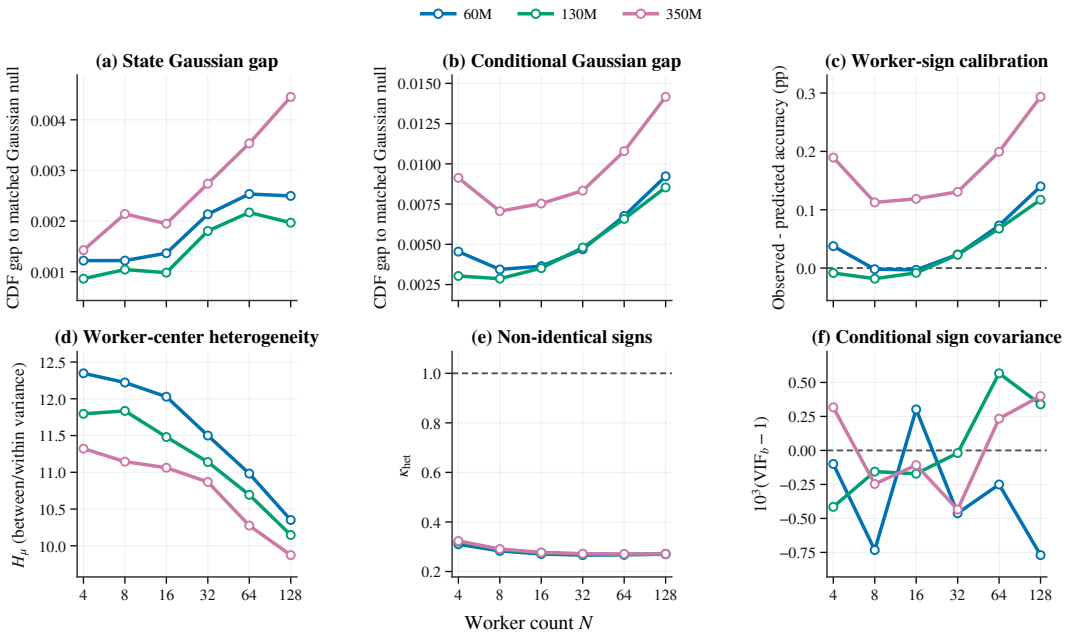

Figure 13: Shape, calibration, heterogeneity, and covariance. Top: persistent-state and conditional CDF gaps, and held-out sign-accuracy gap. Bottom: worker-center heterogeneity H_μ, independent non-identical variance ratio κ_{het}, and net sign covariance $10^3(\text{VIF}_b - 1)$.

D.2.2 Held-out worker-sign prediction

For fold f, let $\mathcal{A}_f$ denote its 96 fit probes and $\mathcal{B}_f$ its 32 evaluation probes. For each worker-coordinate pair, fit

$$\begin{aligned}\widetilde{\mu}_{t,i}^{(n,-f)} &= \frac{1}{|\mathcal{A}_f|}\sum_{k\in\mathcal{A}_f}\widetilde{m}_{t+1,i}^{(n,k)},\\ \left(\widetilde{\sigma}_{t,i}^{(n,-f)}\right)^2 &= \frac{1}{|\mathcal{A}_f|-1}\sum_{k\in\mathcal{A}_f}\left(\widetilde{m}_{t+1,i}^{(n,k)}-\widetilde{\mu}_{t,i}^{(n,-f)}\right)^2.\end{aligned} \tag{49}$$

The Gaussian prediction for the nonnegative-sign probability is then the same probit link as Equation 4:

$$\widetilde{q}_{G,t,i}^{(n,-f)} = \Phi\left(\frac{\widetilde{\mu}_{t,i}^{(n,-f)}}{\widetilde{\sigma}_{t,i}^{(n,-f)}}\right). \tag{50}$$

The held-out bit uses the main-text zero convention:

$$\widetilde{b}_{t+1,i}^{(n,k)} = \mathbb{I}\left[\widetilde{m}_{t+1,i}^{(n,k)} \geq 0\right], \qquad k \in \mathcal{B}_f. \tag{51}$$

The predicted hard-sign accuracy is $\max\{\widetilde{q}_{G,t,i}^{(n,-f)}, 1-\widetilde{q}_{G,t,i}^{(n,-f)}\}$, while the observed accuracy is the fraction of held-out bits whose encoded signs match $\text{sgn}_+(\widetilde{\mu}_{t,i}^{(n,-f)})$. Disjoint fit and evaluation observations prevent bias from selecting the fitted direction on the evaluation set.

For probability calibration, we use the Brier score

$$\text{BS}(\widetilde{q}) = \frac{1}{J}\sum_{j=1}^{J}\left(\widetilde{b}_j - \widetilde{q}_j\right)^2, \tag{52}$$

where J is the number of relevant held-out sign observations. We compare the Gaussian probability with a direct sign-frequency estimate formed on the same 96 fit probes, using the finite-sample adjustment

$$\widetilde{q}_{E,t,i}^{(n,-f)} = \frac{\frac{1}{2}+\sum_{k\in\mathcal{A}_f}\widetilde{b}_{t+1,i}^{(n,k)}}{1+|\mathcal{A}_f|}. \tag{53}$$

Table 22: Four-fold held-out worker-sign prediction and calibration. Gap is observed minus predicted accuracy; G and E denote Gaussian and empirical sign-frequency predictors.

		Hard-sign accuracy			Probability calibration			
Model	N	Pred. (%)	Obs. (%)	Gap (pp)	BS_G	BS_E	$10^3\Delta\mathrm{BS}$	$10^3\mathrm{ECE}$
60M	4	91.04	91.07	0.038	0.0632	0.0633	-0.118	1.535
	8	90.85	90.85	-0.002	0.0644	0.0647	-0.211	1.072
	16	90.82	90.81	-0.003	0.0648	0.0650	-0.221	1.175
	32	90.72	90.74	0.023	0.0653	0.0655	-0.213	1.487
	64	90.54	90.61	0.073	0.0664	0.0666	-0.199	2.085
	128	90.34	90.48	0.140	0.0675	0.0676	-0.169	2.873
130M	4	90.71	90.70	-0.008	0.0655	0.0658	-0.205	1.001
	8	90.73	90.71	-0.018	0.0655	0.0657	-0.227	0.985
	16	90.65	90.64	-0.008	0.0660	0.0662	-0.232	1.115
	32	90.59	90.61	0.023	0.0663	0.0665	-0.223	1.503
	64	90.45	90.52	0.068	0.0670	0.0672	-0.207	2.040
	128	90.29	90.41	0.117	0.0679	0.0681	-0.192	2.659
350M	4	90.43	90.62	0.189	0.0665	0.0665	0.082	2.998
	8	90.49	90.60	0.113	0.0665	0.0666	-0.097	2.322
	16	90.48	90.60	0.119	0.0665	0.0667	-0.122	2.436
	32	90.42	90.55	0.131	0.0669	0.0671	-0.146	2.609
	64	90.30	90.50	0.199	0.0674	0.0675	-0.109	3.365
	128	90.16	90.45	0.293	0.0679	0.0680	-0.042	4.418

We report $\Delta\mathrm{BS} = \mathrm{BS}(\widetilde{q}_G) - \mathrm{BS}(\widetilde{q}_E)$ and expected calibration error (ECE) over ten fixed probability bins. Near-zero $\Delta\mathrm{BS}$ indicates similar predictive accuracy for Gaussian moments and direct sign frequencies. Neither metric uses the ADASIGNSGD lookup table or fits a count-to-scale map.

Observed held-out sign accuracy is 90.41%–91.07%, with an observed-minus-Gaussian-predicted gap of -0.018 to $+0.293$ percentage points (Table 22). Worker-specific ECE is 0.00099–0.00442, and $\Delta\mathrm{BS}$ ranges from -2.32×10^{-4} to 8.19×10^{-5}. Negative differences can arise from finite-sample smoothing by the Gaussian two-moment fit. Despite measurable tail deviations, the conditional Gaussian model therefore accurately predicts the sign event communicated by ADASIGNSGD.

The same calibration is already visible during warmup for every tested worker count. Across the 18 model–N cells in Tables 27–29, the held-out worker-specific observed accuracy through step 2000 is 90.70%–91.49%, and the observed-minus-predicted gap is -0.066–0.159 percentage points.

D.2.3 WORKERS ARE CONDITIONALLY NON-IDENTICAL

The $K = 128$ reset probes separate between-worker history differences from within-worker current-batch variation. For fixed (t, i), define

$$\begin{aligned} \widetilde{\mu}_{t,i}^{(n)} &= \frac{1}{K}\sum_{k=1}^{K} \widetilde{m}_{t+1,i}^{(n,k)}, \\ \left(\widetilde{\sigma}_{t,i}^{(n)}\right)^2 &= \frac{1}{K-1}\sum_{k=1}^{K}\left(\widetilde{m}_{t+1,i}^{(n,k)} - \widetilde{\mu}_{t,i}^{(n)}\right)^2. \end{aligned} \tag{54}$$

The within-worker variance is

$$V_{\mathrm{within},t,i} = \frac{1}{N}\sum_{n=1}^{N}\left(\widetilde{\sigma}_{t,i}^{(n)}\right)^2. \tag{55}$$

The variance of $\{\widetilde{\mu}_{t,i}^{(n)}\}_{n=1}^{N}$ includes center differences and finite-K estimation noise. We remove the leading noise contribution:

$$V_{\mathrm{between},t,i} = \left[\mathrm{Var}_n\left(\widetilde{\mu}_{t,i}^{(n)}\right) - \frac{V_{\mathrm{within},t,i}}{K}\right]_+, \tag{56}$$

where $[x]_+ = \max\{x, 0\}$. The global center-heterogeneity ratio is a ratio of aggregated variances,

$$H_\mu = \frac{\sum_{t,i} V_{\text{between},t,i}}{\sum_{t,i} V_{\text{within},t,i}}. \tag{57}$$

Thus $H_\mu = 1$ equates between-center and mean within-worker variance; $\sqrt{H_\mu}$ compares standard deviations. We report CV_σ, the coefficient of variation of worker conditional standard deviations averaged across coordinates.

For sign heterogeneity, define

$$\widetilde{q}_{t,i}^{(n)} = \frac{1}{K}\sum_{k=1}^{K} \widetilde{b}_{t+1,i}^{(n,k)}, \qquad \widetilde{\overline{q}}_{t,i} = \frac{1}{N}\sum_{n=1}^{N} \widetilde{q}_{t,i}^{(n)}. \tag{58}$$

We analogously subtract the finite-K estimation variance from $\text{Var}_n(\widetilde{q}_{t,i}^{(n)})$ and normalize its aggregate by $\widetilde{\overline{q}}_{t,i}(1 - \widetilde{\overline{q}}_{t,i})$; the resulting ratio is denoted H_q. Positive H_q indicates different conditional nonnegative-sign probabilities, which can coexist with mutually independent fresh sign draws.

The consequence for the vote count can be stated directly. If workers are conditionally independent but not identically distributed, define the population probabilities

$$\begin{aligned} q_{t,i}^{(n)} &= \Pr\left[\widetilde{m}_{t+1,i}^{(n,k)} \geq 0 \,\middle|\, \boldsymbol{\theta}_t, \{\mathbf{m}_t^{(j)}\}_{j=1}^N\right], \\ \overline{q}_{t,i} &= \frac{1}{N}\sum_{n=1}^{N} q_{t,i}^{(n)}. \end{aligned} \tag{59}$$

Then

$$\text{Var}(c_{t+1,i}) = \sum_{n=1}^{N} q_{t,i}^{(n)}\left(1 - q_{t,i}^{(n)}\right), \tag{60}$$

rather than the i.i.d.-binomial variance $N\overline{q}_{t,i}(1 - \overline{q}_{t,i})$. We summarize their aggregate ratio as

$$\kappa_{\text{het}} = \frac{\sum_{t,i,n} q_{t,i}^{(n)}(1 - q_{t,i}^{(n)})}{\sum_{t,i} N\overline{q}_{t,i}(1 - \overline{q}_{t,i})}, \tag{61}$$

using unbiased finite-probe variance estimates in the numerical computation. A ratio below one indicates lower conditional count variance than the binomial model with common probability $\overline{q}_{t,i}$.

Across complete runs, $H_\mu = 9.87$–12.35 and $\sqrt{H_\mu} = 3.14$–3.51 (Table 23): the conditional-center standard deviation is roughly 3.1–3.5 times the one-step mini-batch-noise standard deviation. The coefficient of variation of conditional scales is only 6.33%–10.87%. Heterogeneity is therefore dominated by history-dependent center shifts.

The sign-space calculation reaches the same conclusion: $H_q = 0.734$–0.921. If the workers were independent but were allowed to retain these different positive-sign probabilities, their count variance would be only 0.266–0.323 of the single-probability i.i.d.-binomial variance. The literal identically distributed part of Equation 3 is therefore not an accurate description of the conditioned worker ensemble.

Table 23: Conditional worker heterogeneity and count-variance ratios. κ_{emp} denotes the empirical count variance relative to the i.i.d.-binomial reference.

Model	N	H_μ	$\sqrt{H_\mu}$	$100\,\mathrm{CV}_\sigma$	H_q	κ_{het}	κ_{emp}
60M	4	12.35	3.51	6.51	0.921	0.310	0.310
	8	12.22	3.50	6.57	0.819	0.283	0.283
	16	12.03	3.47	6.71	0.778	0.271	0.271
	32	11.50	3.39	7.08	0.758	0.266	0.266
	64	10.98	3.31	7.34	0.745	0.267	0.267
	128	10.35	3.22	7.86	0.736	0.270	0.270
130M	4	11.80	3.43	6.33	0.907	0.320	0.320
	8	11.83	3.44	6.49	0.815	0.288	0.287
	16	11.48	3.39	6.70	0.773	0.275	0.275
	32	11.14	3.34	6.96	0.754	0.270	0.270
	64	10.69	3.27	7.32	0.742	0.269	0.270
	128	10.15	3.19	7.65	0.734	0.271	0.272
350M	4	11.32	3.36	7.72	0.903	0.323	0.323
	8	11.14	3.34	7.61	0.811	0.291	0.291
	16	11.06	3.33	7.93	0.771	0.277	0.277
	32	10.87	3.30	8.20	0.752	0.272	0.272
	64	10.28	3.21	8.98	0.741	0.271	0.271
	128	9.87	3.14	10.87	0.735	0.271	0.271

D.2.4 NET ONE-STEP CONDITIONAL COVARIANCE IS NEGLIGIBLE

For the reset probes, the variance of the sign count satisfies the exact identity

$$\mathrm{Var}_k\left(\sum_{n=1}^{N}\widetilde{b}_{t+1,i}^{(n,k)}\right)=\sum_{n=1}^{N}\mathrm{Var}_k\left(\widetilde{b}_{t+1,i}^{(n,k)}\right)+2\sum_{n<j}\mathrm{Cov}_k\left(\widetilde{b}_{t+1,i}^{(n,k)},\widetilde{b}_{t+1,i}^{(j,k)}\right). \tag{62}$$

We summarize the net contribution of all $N(N-1)/2$ worker pairs through the centered aggregate sign variance-inflation factor

$$\mathrm{VIF}_b-1=\frac{\sum_{t,i}\mathrm{Var}_k(\sum_n\widetilde{b}_{t+1,i}^{(n,k)})}{\sum_{t,i,n}\mathrm{Var}_k(\widetilde{b}_{t+1,i}^{(n,k)})}-1. \tag{63}$$

This is the net covariance contribution normalized by independent-worker variance; its sign indicates net positive, zero, or negative conditional covariance. The continuous analogue is

$$\mathrm{VIF}_m-1=\frac{\sum_{t,i}\mathrm{Var}_k(\sum_n\widetilde{m}_{t+1,i}^{(n,k)})}{\sum_{t,i,n}\mathrm{Var}_k(\widetilde{m}_{t+1,i}^{(n,k)})}-1. \tag{64}$$

For compact comparison across N, we also report

$$\rho_{\mathrm{eff}}=\frac{\mathrm{VIF}-1}{N-1}. \tag{65}$$

This is the equal-correlation value that would produce the observed aggregate variance inflation. It is not an unweighted mean of the pairwise Pearson correlations.

Let κ_{emp} denote the empirical count variance divided by the i.i.d.-binomial variance. Equations 61 and 63 give the exact aggregate decomposition

$$\kappa_{\mathrm{emp}}=\kappa_{\mathrm{het}}\left[1+(\mathrm{VIF}_b-1)\right]. \tag{66}$$

Different worker probabilities affect κ_{het}; cross-worker covariance affects VIF_b-1.

Table 24 reports $10^3(\mathrm{VIF}_m-1)=-1.314$–$4.267$ and $10^3(\mathrm{VIF}_b-1)=-0.770$–$0.567$. The corresponding variance-equivalent dependences are between -4.38×10^{-4} and 2.59×10^{-4} for momenta and between -1.38×10^{-4} and 1.06×10^{-4} for signs. Thus the sum of all pairwise

Table 24: Net conditional covariance of one-step momenta and signs. Zero VIF offset indicates no net covariance contribution; ρ_{eff} is the variance-equivalent correlation.

		Candidate momentum		Sign bit	
Model	N	$10^3(\text{VIF}-1)$	$10^4\rho_{\text{eff}}$	$10^3(\text{VIF}-1)$	$10^4\rho_{\text{eff}}$
60M	4	-1.314	-4.379	-0.100	-0.332
	8	-0.183	-0.261	-0.733	-1.047
	16	-0.818	-0.545	0.302	0.201
	32	1.038	0.335	-0.462	-0.149
	64	4.267	0.677	-0.250	-0.040
	128	0.267	0.021	-0.770	-0.061
130M	4	-1.284	-4.280	-0.415	-1.383
	8	1.815	2.593	-0.156	-0.222
	16	-0.325	-0.216	-0.171	-0.114
	32	2.115	0.682	-0.019	-0.006
	64	-0.097	-0.015	0.567	0.090
	128	-0.972	-0.076	0.339	0.027
350M	4	-0.281	-0.936	0.317	1.057
	8	-0.354	-0.506	-0.247	-0.352
	16	2.295	1.530	-0.108	-0.072
	32	-0.253	-0.082	-0.435	-0.140
	64	0.321	0.051	0.234	0.037
	128	0.686	0.054	0.400	0.032

conditional covariances changes the relevant aggregate variance by much less than one percent in every complete setting.

Consistent with Equation 66, the empirical count-variance ratio $\kappa_{\text{emp}} = 0.266$–$0.323$ is numerically almost identical to κ_{het}. The large departure from an i.i.d. binomial count is therefore explained by different worker probabilities, not by net covariance among the independently resampled current mini-batches.

The test conditions on realized $\boldsymbol{\theta}_t$ and all persistent $\mathbf{m}_t^{(n)}$, varying only the next independently sampled mini-batches. Thus $\text{VIF} - 1 \approx 0$ indicates negligible net one-step conditional covariance, not unconditional independence of persistent histories. Independent complete histories and continuation trajectories were not collected; dependence induced by shared historical updates is absorbed into the conditioned state and lies outside this estimand.

D.2.5 Vote support and full-value update fidelity

We evaluate the vote margin along training using the unmodified lookup. The finite-worker full-value analogue of $u^\star$ in Equation 5 is

$$\widetilde{u}_{t,i}^{\text{FV}} = \frac{N^{-1}\sum_{n=1}^{N} m_{t,i}^{(n)}}{\sqrt{N^{-1}\sum_{n=1}^{N}(m_{t,i}^{(n)})^2}}. \tag{67}$$

This quantity uses the full momentum values of the same workers that produce $b_{t,i}^{(n)}$, the exact count $c_{t,i}$, the margin $\delta_{t,i}$, and the update $u_{t,i} = \widehat{y}_{t,i}\widehat{a}_{t,i}$, where $\widehat{a}_{t,i} = \text{LUT}_N[\delta_{t,i}]$, in Section 4. For a reset probe we replace $m_{t,i}^{(n)}$ by $\widetilde{m}_{t+1,i}^{(n,k)}$ everywhere in Equation 67. In both cases $u_{t,i}$ is computed with the BF16 lookup table saved from the actual run; no scalar is fitted and no alternative mapping is introduced.

We condition the descriptive results on the number of workers supporting the majority, $\max\{c_{t,i}, N - c_{t,i}\}$. For each support level, we report the probability that the majority direction $\text{sgn}_+(2c_{t,i} - N)$ agrees with the sign of the numerator in Equation 67. Aggregate correspondence is summarized by the cosine between the selected-coordinate vectors of $\mathbf{u}_t$ and their full-value counterparts, computed per tensor and world and then averaged, and by

$$\text{MAE}_{\text{scale}} = \mathbb{E}\left[\left||u_{t,i}| - |\widetilde{u}_{t,i}^{\text{FV}}|\right|\right]. \tag{68}$$

Table 25: Persistent-state direction agreement at exact ties (support $N/2$) and positive reference margins.

		Exact tie	Positive reference support		
Model	N	Direction (%)	Support	$\delta_{t,i}/N$	Direction (%)
60M	4	50.33	3/4	0.50	91.73
	8	50.72	5/8	0.25	82.75
	16	49.95	10/16	0.25	90.77
	32	50.17	20/32	0.25	97.25
	64	50.19	40/64	0.25	99.26
	128	51.34	80/128	0.25	100.00
130M	4	49.69	3/4	0.50	91.70
	8	49.96	5/8	0.25	82.70
	16	49.98	10/16	0.25	90.93
	32	50.92	20/32	0.25	97.22
	64	50.01	40/64	0.25	99.75
	128	49.46	80/128	0.25	100.00
350M	4	50.16	3/4	0.50	91.42
	8	49.77	5/8	0.25	82.62
	16	49.95	10/16	0.25	90.67
	32	49.64	20/32	0.25	96.83
	64	49.83	40/64	0.25	99.67
	128	50.15	80/128	0.25	99.92

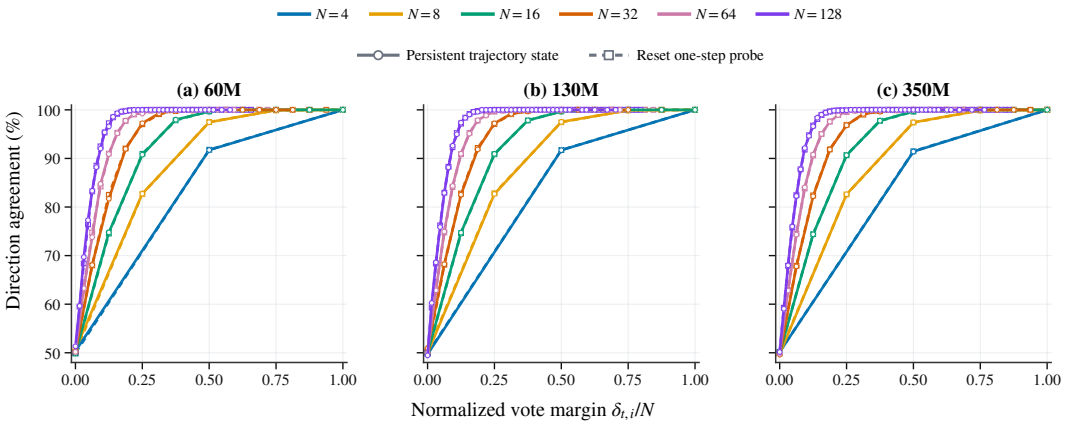

Figure 14: Direction agreement versus normalized vote margin. Columns: 60M, 130M, and 350M; colors: $N \in \{4, 8, 16, 32, 64, 128\}$.

This compression-fidelity diagnostic measures information retained about discarded full values. Because the count and reference share the same workers, it is not an independent calibration test of $p = \Phi(S)$. Scale MAE is unnormalized.

Figure 14 and Table 25 report support-conditioned direction agreement for all six worker counts and all three model sizes. Across all 18 model–N settings, agreement increases over adjacent support bins containing at least 100 persistent-state observations, up to two fluctuations of at most 0.16 percentage points when agreement is already at least 99.85%. At an exact tie, direction agreement is 49.46%–51.34% across all settings.

At the common normalized margin $\delta_{t,i}/N = 0.25$, direction agreement rises from 82.62%–82.75% at $N = 8$ to 90.67%–90.93% at $N = 16$, 96.83%–97.25% at $N = 32$, 99.26%–99.75% at $N = 64$, and 99.92%–100.00% at $N = 128$. The $N = 4$ grid does not contain 0.25; at its nearest positive margin, $\delta_{t,i}/N = 0.50$, agreement is 91.42%–91.73%. Over all displayed support bins, the reset-probe estimates differ from the persistent-state estimates by at most 1.30 percentage points.

Table 26 reports compression fidelity for every worker count. The mean tensor cosine remains between 0.804 and 0.847 over all settings. Across model sizes, the persistent-state scale-MAE ranges are 0.201–0.211 at $N = 4$, 0.157–0.161 at $N = 8$, 0.117–0.119 at $N = 16$, 0.086–0.087

Table 26: Fidelity of the unmodified ADASIGNSGD lookup: scale MAE and mean tensor cosine.

		Persistent state		Reset one-step probes	
Model	N	Scale MAE	Cosine	Scale MAE	Cosine
60M	4	0.201	0.847	0.201	0.847
	8	0.157	0.821	0.157	0.823
	16	0.117	0.815	0.117	0.815
	32	0.086	0.814	0.086	0.813
	64	0.062	0.818	0.062	0.818
	128	0.045	0.829	0.045	0.829
130M	4	0.210	0.847	0.210	0.847
	8	0.160	0.824	0.160	0.824
	16	0.119	0.811	0.119	0.811
	32	0.087	0.809	0.087	0.809
	64	0.063	0.812	0.063	0.812
	128	0.045	0.823	0.045	0.822
350M	4	0.211	0.845	0.211	0.845
	8	0.161	0.820	0.160	0.820
	16	0.119	0.808	0.119	0.808
	32	0.087	0.804	0.087	0.805
	64	0.063	0.808	0.063	0.808
	128	0.046	0.814	0.046	0.814

at $N = 32$, 0.062–0.063 at $N = 64$, and 0.045–0.046 at $N = 128$. Reset-probe results differ from their persistent-state counterparts by at most 0.0005 in scale MAE and 0.0016 in mean tensor cosine. These cross-N values are descriptive, not causal comparisons, because the per-worker batch is $512/N$.

D.2.6 STABILITY ACROSS TRAINING STAGES

Figure 15 shows every 500-step checkpoint for all 18 model–N trajectories. Tables 27, 28, and 29 give all 72 prespecified model–N–stage aggregates. Warmup contains the four checkpoints through step 2,000; remaining checkpoints are split into three equal-progress intervals, denoted Early, Middle, and Late. Accuracy is the observed worker-specific held-out sign accuracy, and Gap is observed minus Gaussian-predicted accuracy. Worker-specific sign accuracy remains between 90.35% and 91.49% over these cells. The $|\mathrm{VIF}_b - 1|$ remains at most 2.11×10^{-3}, H_μ remains between 9.49 and 14.06, and persistent-state cosine remains between 0.802 and 0.862. The held-out conditional CDF gap remains at most 0.0156. Thus the Gaussian sign prediction, conditional covariance result, worker-center heterogeneity, and full-value update correspondence persist from warmup through late training across every tested worker count.

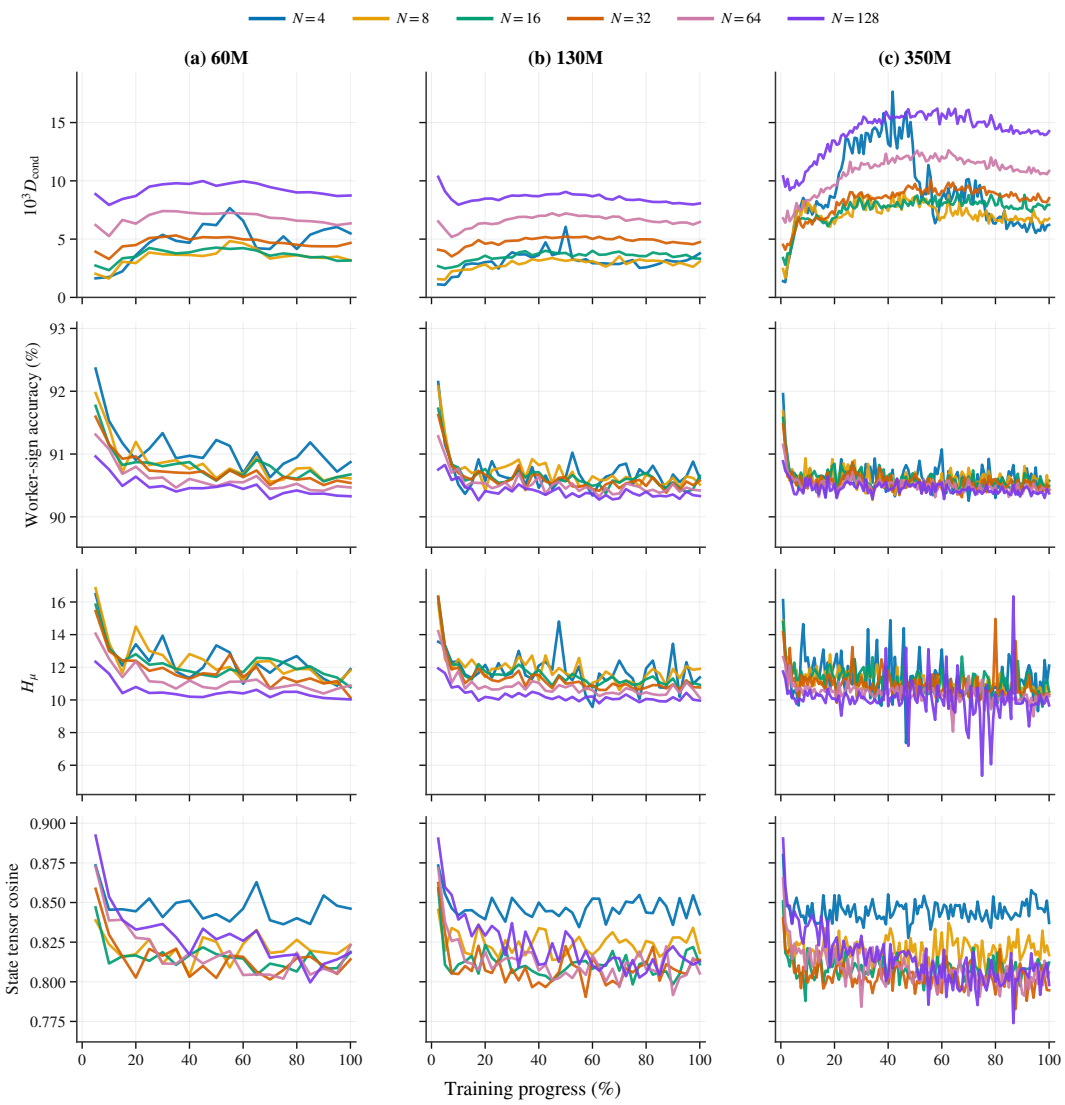

Figure 15: Diagnostics throughout training. Columns: 60M, 130M, and 350M; each panel includes all six worker counts and every 500-step sample. Rows: conditional CDF gap, held-out sign accuracy, H_μ, and mean tensor cosine. Normalized progress aligns the three horizons.

Table 27: Training-stage diagnostics for 60M across all worker counts.

N	Stage	Ckpts.	$10^3 D_{\text{cond}}$	Acc. (%)	Gap (pp)	H_μ	10^3 $(\text{VIF}_b - 1)$	Cosine
4	Warmup	4	2.309	91.49	-0.029	13.76	-0.501	0.852
	Early	5	5.044	91.05	0.045	12.26	1.518	0.847
	Middle	5	5.598	90.94	0.065	12.14	-0.537	0.846
	Late	6	5.300	90.92	0.053	11.79	-0.813	0.844
8	Warmup	4	2.411	91.33	-0.022	13.94	0.806	0.824
	Early	5	3.681	90.84	0.004	12.54	-1.413	0.818
	Middle	5	4.118	90.71	0.020	12.00	-1.204	0.822
	Late	6	3.411	90.66	-0.012	11.53	-0.738	0.821
16	Warmup	4	2.974	91.15	-0.020	13.42	0.413	0.823
	Early	5	4.005	90.82	0.011	11.86	-0.190	0.816
	Middle	5	4.050	90.74	0.005	12.06	1.156	0.813
	Late	6	3.439	90.64	-0.010	11.62	-0.071	0.812
32	Warmup	4	4.019	91.16	0.004	13.19	-1.031	0.827
	Early	5	5.145	90.72	0.035	11.60	0.032	0.814
	Middle	5	4.985	90.64	0.035	11.62	0.489	0.809
	Late	6	4.504	90.58	0.017	10.98	-1.299	0.811
64	Warmup	4	6.111	90.97	0.057	12.46	0.504	0.845
	Early	5	7.260	90.57	0.086	10.96	-0.046	0.816
	Middle	5	7.118	90.54	0.078	10.97	0.177	0.810
	Late	6	6.481	90.46	0.068	10.72	-1.249	0.810
128	Warmup	4	8.483	90.71	0.120	11.17	-0.500	0.854
	Early	5	9.735	90.45	0.156	10.30	1.174	0.830
	Middle	5	9.696	90.45	0.154	10.40	-2.107	0.826
	Late	6	8.909	90.37	0.129	10.23	-1.444	0.813

Table 28: Training-stage diagnostics for 130M across all worker counts.

N	Stage	Ckpts.	$10^3 D_{\text{cond}}$	Acc. (%)	Gap (pp)	H_μ	10^3 $(\text{VIF}_b - 1)$	Cosine
4	Warmup	4	1.426	91.17	-0.066	13.12	-0.739	0.854
	Early	12	3.201	90.64	-0.006	11.79	-0.899	0.845
	Middle	12	3.409	90.63	0.006	11.63	-0.089	0.844
	Late	12	3.018	90.67	-0.006	11.44	-0.156	0.847
8	Warmup	4	1.910	91.24	-0.039	13.61	0.149	0.833
	Early	12	2.785	90.76	-0.019	12.06	-1.039	0.823
	Middle	12	3.132	90.63	-0.007	11.49	0.548	0.823
	Late	12	2.990	90.58	-0.021	11.52	-0.083	0.824
16	Warmup	4	2.611	91.14	-0.025	13.95	-2.051	0.821
	Early	12	3.492	90.65	-0.011	11.74	0.293	0.812
	Middle	12	3.780	90.55	-0.002	11.20	-0.214	0.809
	Late	12	3.589	90.55	-0.006	11.06	-0.001	0.809
32	Warmup	4	3.854	91.09	-0.003	13.72	0.963	0.826
	Early	12	4.775	90.62	0.021	11.38	-0.471	0.807
	Middle	12	5.133	90.52	0.034	10.92	0.558	0.805
	Late	12	4.797	90.53	0.022	10.83	-0.459	0.810
64	Warmup	4	5.739	90.94	0.044	12.67	0.634	0.841
	Early	12	6.516	90.53	0.065	10.86	1.279	0.813
	Middle	12	7.002	90.44	0.080	10.55	0.189	0.806
	Late	12	6.498	90.45	0.066	10.53	0.219	0.809
128	Warmup	4	8.948	90.70	0.131	11.36	1.797	0.861
	Early	12	8.522	90.42	0.117	10.29	-0.093	0.828
	Middle	12	8.757	90.35	0.123	10.08	-0.072	0.816
	Late	12	8.185	90.36	0.106	10.06	0.710	0.812

Table 29: Training-stage diagnostics for 350M across all worker counts.

N	Stage	Ckpts.	$10^3 D_{\mathrm{cond}}$	Acc. (%)	Gap (pp)	H_μ	10^3 $(\mathrm{VIF}_b - 1)$	Cosine
4	Warmup	4	2.200	91.20	-0.026	14.06	1.830	0.855
	Early	38	10.059	90.63	0.200	11.67	0.162	0.844
	Middle	39	11.206	90.62	0.272	11.24	-0.065	0.844
	Late	39	6.898	90.57	0.119	11.00	0.702	0.845
8	Warmup	4	2.667	91.07	-0.021	13.68	0.296	0.831
	Early	38	7.245	90.63	0.126	11.51	0.108	0.819
	Middle	39	7.711	90.55	0.132	10.96	-0.665	0.820
	Late	39	6.727	90.57	0.093	10.80	-0.223	0.820
16	Warmup	4	3.417	91.07	-0.005	13.39	-0.315	0.824
	Early	38	7.019	90.65	0.110	11.57	-0.019	0.808
	Middle	39	8.060	90.57	0.132	10.99	-0.502	0.807
	Late	39	7.923	90.53	0.127	10.73	0.219	0.808
32	Warmup	4	4.421	90.96	0.019	13.09	0.838	0.825
	Early	38	7.474	90.57	0.106	11.08	-0.300	0.804
	Middle	39	9.121	90.53	0.154	10.77	-0.898	0.803
	Late	39	8.783	90.51	0.143	10.72	-0.227	0.802
64	Warmup	4	6.788	90.87	0.080	12.05	-0.437	0.842
	Early	38	9.657	90.52	0.162	10.58	0.158	0.814
	Middle	39	11.951	90.50	0.238	10.23	0.304	0.806
	Late	39	11.138	90.45	0.210	10.12	0.303	0.802
128	Warmup	4	9.807	90.73	0.159	11.14	1.520	0.862
	Early	38	12.781	90.46	0.251	10.07	0.470	0.826
	Middle	39	15.581	90.47	0.339	10.32	0.640	0.809
	Late	39	14.561	90.40	0.303	9.49	-0.017	0.804

E VoteReduce Implementation and Exactness

Algorithm 1 is the complete optimizer specification and the sole definition of the margin-to-scale table LUT_N. This appendix specifies only the exact distributed collective behind its VoteReduce line. Separating optimizer semantics from transport details avoids introducing a second update rule.

E.1 Conventions inherited from Algorithm 1

All workers begin a step with identical $\boldsymbol{\theta}_t$, while worker $n \in \{1, \dots, N\}$ retains its local momentum $\mathbf{m}_{t-1}^{(n)}$. Momentum is initialized as $\mathbf{m}_0^{(n)} = \mathbf{0}$. Standard first-moment bias correction multiplies momentum by a positive scalar and therefore cannot change the communicated bits

$$b_{t,i}^{(n)} := \mathbb{I}[m_{t,i}^{(n)} \geq 0]. \tag{69}$$

The collective must return, on every worker, exactly

$$c_{t,i} := \sum_{n=1}^{N} b_{t,i}^{(n)} \in \{0, \dots, N\}. \tag{70}$$

Once this vector is available, every worker obtains the same $\widehat{\mathbf{y}}_t$ and $\boldsymbol{\delta}_t$, applies the same precomputed LUT_N to $\boldsymbol{\delta}_t$, and hence obtains the same $\mathbf{u}_t$. The method communicates no gradient magnitude, floating-point momentum, second moment, or error-feedback residual.

For reference, if

$$\mathbf{B}_t := \left[\mathbf{b}_t^{(1)}, \dots, \mathbf{b}_t^{(N)}\right] \in \{0,1\}^{d \times N}, \tag{71}$$

then the required semantics are simply

$$\mathbf{c}_t = \mathbf{B}_t \mathbf{1}_N. \tag{72}$$

The conceptual matrix $\mathbf{B}_t$ is never materialized.

E.2 EXACT OWNER-SHARDED VOTE AGGREGATION

VOTEREDUCE uses a fixed balanced partition of the d coordinates. Worker s temporarily owns shard $\mathcal{I}_s$, where

$$\mathcal{I}_s \cap \mathcal{I}_{s'} = \varnothing \ (s \neq s'), \qquad \bigcup_{s=1}^{N} \mathcal{I}_s = \{1, \dots, d\}, \qquad \big|\, |\mathcal{I}_s| - |\mathcal{I}_{s'}| \,\big| \leq 1. \tag{73}$$

In Phase I, worker n bit-packs $\mathbf{b}_t^{(n)}[\mathcal{I}_s]$ for every owner s and routes the packed shards with ALL-TOALL. Owner s decodes and sums the N received bit vectors into

$$\mathbf{c}_t^{[s]} := \sum_{n=1}^{N} \mathbf{b}_t^{(n)}[\mathcal{I}_s] \in \{0, \dots, N\}^{|\mathcal{I}_s|}. \tag{74}$$

The brackets in $\mathbf{c}_t^{[s]}$ mark an owner shard, whereas parentheses in $\mathbf{b}_t^{(n)}$ mark a worker-local vector.

In Phase II, the owners encode and ALLGATHER only these reduced count shards. Every worker decodes and assembles

$$\mathbf{c}_t = \mathrm{Assemble}_{\{\mathcal{I}_s\}_{s=1}^{N}} \left(\mathbf{c}_t^{[1]}, \dots, \mathbf{c}_t^{[N]} \right). \tag{75}$$

The count encoding is either the minimum exact width

$$w_N := \lceil \log_2(N+1) \rceil \tag{76}$$

or byte-aligned `uint8`. The w_N-bit path is exact for every N; `uint8` is exact when $N \leq 255$, which covers all experiments. The two paths differ only in Phase II representation, not in counts or optimizer updates.

Fixed-width count layout. For a logical sequence $x = (x_1, \dots, x_L)$ with $x_j \in \{0, \dots, 2^w - 1\}$, Pack_w concatenates the unsigned w-bit fields in coordinate order. If $P \in \{0,1\}^{Lw}$ is the resulting logical bitstream, then

$$P_{(j-1)w+\ell+1} := \left\lfloor \frac{x_j}{2^\ell} \right\rfloor \bmod 2, \qquad j = 1, \dots, L, \quad \ell = 0, \dots, w-1, \tag{77}$$

and the inverse decoder recovers

$$x_j = \sum_{\ell=0}^{w-1} 2^\ell P_{(j-1)w+\ell+1}. \tag{78}$$

Fields are contiguous and may cross byte or machine-word boundaries. Any final transport padding follows the Lw logical bits and is discarded using the known shard length. Hence $\mathrm{EncodeCounts}_{w_N}$ and $\mathrm{DecodeCounts}_{w_N}$ are exact inverses because $c_{t,i} \leq N < 2^{w_N}$.

Algorithm 2 VOTEREDUCE: exact owner-sharded aggregation on worker n

Require: Local bits $\mathbf{b}_t^{(n)} \in \{0,1\}^d$; worker count N; balanced partition $\{\mathcal{I}_s\}_{s=1}^N$; count encoding $\chi \in \{w_N\text{-bit}, \texttt{uint8}\}$, with `uint8` requiring $N \le 255$
Ensure: Exact $\mathbf{c}_t \in \{0,\dots,N\}^d$, identical on every worker

Phase I: route packed signs and reduce at owners

for $s = 1,\dots,N$ **do**
  $\mathbf{z}_{n\to s} \leftarrow \mathrm{PackBits}(\mathbf{b}_t^{(n)}[\mathcal{I}_s])$
end for
$\{\mathbf{z}_{j\to n}\}_{j=1}^N \leftarrow \mathrm{AllToAll}(\{\mathbf{z}_{n\to s}\}_{s=1}^N)$ ▷ worker n owns $\mathcal{I}_n$
$\mathbf{c}_t^{[n]} \leftarrow \mathbf{0}_{|\mathcal{I}_n|}$
for $j = 1,\dots,N$ **do**
  $\mathbf{c}_t^{[n]} \leftarrow \mathbf{c}_t^{[n]} + \mathrm{UnpackBits}(\mathbf{z}_{j\to n}, |\mathcal{I}_n|)$
end for

Phase II: redistribute only reduced count shards

$\mathbf{h}_n \leftarrow \mathrm{EncodeCounts}_\chi(\mathbf{c}_t^{[n]})$
$\{\mathbf{h}_s\}_{s=1}^N \leftarrow \mathrm{AllGather}(\mathbf{h}_n)$
for $s = 1,\dots,N$ **do**
  $\mathbf{c}_t^{[s]} \leftarrow \mathrm{DecodeCounts}_\chi(\mathbf{h}_s, |\mathcal{I}_s|)$
end for
$\mathbf{c}_t \leftarrow \mathrm{Assemble}_{\{\mathcal{I}_s\}_{s=1}^N}(\mathbf{c}_t^{[1]},\dots,\mathbf{c}_t^{[N]})$
return $\mathbf{c}_t$

Exactness and replica synchronization. Equation 73 assigns each coordinate to one owner. That owner receives one losslessly decoded bit from every worker, so Equation 74 equals the count in Equation 70. Phase II then losslessly returns each owner shard to every worker in the original coordinate order. Consequently all workers obtain the same $\mathbf{c}_t$, compute the same $\widehat{\mathbf{y}}_t$ and $\boldsymbol{\delta}_t$, look up the same $\widehat{\mathbf{a}}_t$, and remain synchronized. Replacing VOTEREDUCE with any exact implementation of Equation 72 leaves the optimization trajectory unchanged.

Runtime state and temporary storage. Worker-local momentum $\mathbf{m}_t^{(n)}$ is the sole persistent model-sized optimizer state. Packed votes, owner-side count accumulators, encoded counts, and lookup outputs are step-local temporaries. The implementation streams owner-side accumulation in chunks rather than retaining all received sign shards. Appendices G and H report temporary memory, traffic, and latency for both exact count encodings.

F PHYSICAL SYSTEMS, NETWORK TOPOLOGY, AND SOFTWARE STACK

This section records the physical and software environments used by the memory, communication-path, and complete-step studies. Each distributed process is bound to one GPU, with four ranks per node. The P100 experiments scale to 32 nodes ($N = 128$), whereas the H200 experiments scale to eight nodes ($N = 32$). We separately inventoried two allocated nodes from each Slurm GPU class; the paired inventories agreed on every field reported below.

F.1 Physical Node and Fabric Configuration

Table 30: Physical systems used for the systems experiments. Link rates are nominal per-port line rates reported by `ibstat`, not measured collective bandwidth. The NCCL data-path row is obtained from the per-rank NCCL transport traces of the independent calibration below.

Property	P100 suite	H200 suite
Accelerators per node	4× Tesla P100 PCIe 16 GB; Pascal, compute capability 6.0, 56 SMs/GPU	4× NVIDIA H200 SXM 141 GB; Hopper, compute capability 9.0, 132 SMs/GPU
GPU–host interface	PCIe Gen3 ×16	PCIe Gen5 ×16
Intra-node GPU topology	PCIe only; no active NVLink. GPU pairs $(0, 1)$ and $(2, 3)$ are PIX-connected; cross-pair paths are NODE-connected	Fully connected NV6: every GPU pair is joined by six bonded NVLinks; 18 active links/GPU at 26.562 GB/s per link as reported by `nvidia-smi`
Host processors	2× Intel Xeon E5-2680 v4, 14 cores/socket (28 physical cores/node; SMT disabled)	2× AMD EPYC 9554, 64 cores/socket (128 physical cores/node; SMT disabled)
Host memory	Approximately 256 GB across two NUMA domains	Approximately 1.5 TB across two NUMA domains
Inter-node data fabric	1× Mellanox ConnectX-4, 100-Gb/s EDR InfiniBand, active MTU 4096	4× Mellanox ConnectX-7, each 400-Gb/s NDR InfiniBand, active MTU 4096; 1.6-Tb/s aggregate nominal line rate/node
GPU–HCA locality	One HCA shared by four GPUs; PHB distance from GPUs 0–1 and NODE distance from GPUs 2–3	One PXB-local ConnectX-7 HCA per GPU
Observed NCCL inter-node data path	Single EDR rail; the selected channels are host-staged (no `/GDRDMA` channel in the trace)	All four NDR rails with balanced channel assignment and GPUDirect RDMA
Largest evaluated allocation	32 nodes, 128 GPUs	8 nodes, 32 GPUs
Rank placement	Four ranks/node; one process and one GPU per rank	

F.2 Software Environment and Runtime Conventions

Table 31: Software environment recorded inside allocated compute jobs. CUDA denotes the build reported by `torch.version.cuda`; the "CUDA Version" displayed by `nvidia-smi` is only the driver's compatibility ceiling and is not used here.

Property	P100 suite	H200 suite
Operating system	Red Hat Enterprise Linux 9.5; Linux 5.14.0-427.13.1.el9_4.x86_64	Red Hat Enterprise Linux 9.5; Linux 5.14.0-427.13.1.el9_4.x86_64
Python	3.10.19	3.10.19
PyTorch	2.5.1+cu121	2.5.1+cu121
CUDA build	12.1	12.1
cuDNN	9.1.0 (version integer 90100)	9.1.0 (version integer 90100)
NCCL	2.21.5	2.21.5
NVIDIA driver	575.57.08	580.126.16
Arithmetic for latency studies	FP32 (P100 does not support TF32)	FP32 with TF32 disabled
DDP bucketization	25-MiB main buckets and a 1-MiB first bucket	25-MiB main buckets and a 1-MiB first bucket
Reported distributed time	Synchronized CUDA timing; maximum elapsed time across ranks	Synchronized CUDA timing; maximum elapsed time across ranks

The H200 nodes also expose a 25-Gb/s Ethernet/RoCE interface. NCCL enumerated that interface, but no collective data channel selected it: all inter-node payload channels used the four 400-Gb/s InfiniBand devices, with equal channel counts and `/GDRDMA` in the transport trace. Ethernet was used only for out-of-band bootstrap/control traffic. Intra-node H200 channels used `P2P/CUMEM`. NCCL selected zero NVLS channels, so we describe the measured four-GPU topology as fully connected NVLink (NV6), not NVSwitch/NVLS. On P100, all inter-node payload channels used the single 100-Gb/s InfiniBand device; the selected channels lacked `/GDRDMA`, consistent with the logged PCIe-distance check and host-staged path.

F.3 Independent NCCL Fabric Calibration

We independently calibrate the hardware path with FP32 sum all-reduce, separate from every optimizer-specific timing result. The intra-node test uses one node and four GPUs; the inter-node test uses two nodes and eight GPUs. Each iteration is timed with CUDA events, and the maximum time across ranks is used. For payload size M, we report algorithmic bandwidth $\mathrm{algbw} = M/t$ and the conventional all-reduce bus bandwidth

$$\mathrm{busbw} = \mathrm{algbw}\ \frac{2(N-1)}{N}. \tag{79}$$

The probe uses 10 warm-up and 100 measured calls through 1 MiB, 8 and 50 calls at 4–16 MiB, 5 and 25 calls at 64 MiB, and 3 and 10 calls at 256 MiB.

Table 32: Independent NCCL FP32 all-reduce calibration. Entries report mean time in milliseconds and conventional all-reduce bus bandwidth in GB/s. Intra-node uses one node/four GPUs and inter-node uses two nodes/eight GPUs. Bus bandwidth is a collective-normalized metric, not a direct per-HCA throughput counter.

	P100				H200			
	Intra-node		Inter-node		Intra-node		Inter-node	
Message	ms	GB/s	ms	GB/s	ms	GB/s	ms	GB/s
4 KiB	0.1467	0.042	0.1394	0.051	0.0453	0.135	0.0512	0.140
16 KiB	0.1443	0.170	0.1438	0.199	0.0424	0.579	0.0512	0.560
64 KiB	0.1614	0.609	0.1368	0.838	0.0430	2.287	0.0546	2.101
256 KiB	0.1537	2.558	0.1943	2.361	0.0450	8.740	0.0534	8.587
1 MiB	0.2492	6.312	0.4660	3.938	0.0459	34.298	0.0594	30.876
4 MiB	0.8766	7.177	1.3773	5.329	0.0463	135.995	0.1194	61.482
16 MiB	3.4659	7.261	5.3780	5.459	0.0983	255.932	0.2444	120.126
64 MiB	13.0387	7.720	20.7141	5.670	0.3328	302.454	0.8583	136.836
256 MiB	48.9052	8.233	79.4644	5.912	1.2149	331.435	2.7346	171.787

At 256 MiB, H200 reaches 331.435 GB/s intra-node and 171.787 GB/s inter-node, whereas P100 reaches 8.233 and 5.912 GB/s, respectively. These values characterize the machines but are not used to reconstruct any optimizer latency: all communication-path and complete-step results are timed directly under their own protocols.

G Persistent-State and Peak-Memory Efficiency

We distinguish the algorithmic number of persistent model-sized optimizer states, the persistent CUDA allocation after method-specific lazy initialization, and complete-step peak allocated memory. The peak includes PyTorch-visible activations, gradients, packed communication buffers, and temporary update tensors. These measures separate state storage from the remaining training footprint.

G.1 State-Counting Convention

A *model-sized optimizer state* contains $\Theta(d)$ elements and persists across optimizer steps, where d is the number of trainable parameters. Parameters, gradients, activations, scalar counters, $\mathcal{O}(N)$ lookup tables, and step-local buffers are excluded. Compression residuals and auxiliary accumulators are included whenever they persist across steps and scale as $\Theta(d)$.

Let L_o and G_o denote the numbers of worker-local and system-level model-sized states. With global states evenly sharded across N workers, the amortized per-worker count is

$$\nu_o^{\mathrm{state}}(N) = L_o + \frac{G_o}{N}. \tag{80}$$

Replicating the global state instead gives $L_o + G_o$ states per worker. Measured persistent allocation distinguishes physical replication from the amortized sharded count.

Table 33: Persistent model-sized optimizer states. L and G are worker-local and system-level counts; bold marks the minimum among the evaluated 1-bit methods. Both ADASIGNSGD count encodings use the same persistent states.

Optimizer	L	G	Amortized per worker	Worker-local states	Global state
Full-precision references					
AdamW	2	0	2	$\boldsymbol{m}$, $\boldsymbol{v}$	–
ADASIGNSGD (FP)	2	0	2	$\boldsymbol{\mu}$, $\boldsymbol{c}$	–
Communication-efficient 1-bit optimizers					
SIGNUM-MV	1	0	**1**	$\boldsymbol{m}$	–
SIGNUM-MV+EF	2	1	$2 + 1/N$	$\boldsymbol{m}$, $\boldsymbol{e}_{\mathrm{w}}$	$\boldsymbol{e}_{\mathrm{g}}$
Birder	2	0	2	$\boldsymbol{m}$, $\boldsymbol{s}$	–
Birder+EC	3	1	$3 + 1/N$	$\boldsymbol{m}$, $\boldsymbol{s}$, $\boldsymbol{e}_{\mathrm{w}}$	$\boldsymbol{e}_{\mathrm{g}}$
1-Bit Adam	3	1	$3 + 1/N$	$\boldsymbol{m}$, $\widehat{\boldsymbol{v}}$, $\boldsymbol{e}_{\mathrm{w}}$	$\boldsymbol{e}_{\mathrm{g}}$
0/1 Adam	4	1	$4 + 1/N$	$\boldsymbol{m}$, $\boldsymbol{v}$, $\boldsymbol{a}$, $\boldsymbol{e}_{\mathrm{w}}$	$\boldsymbol{e}_{\mathrm{g}}$
D-Lion-MaVo	1	0	**1**	$\boldsymbol{m}$	–
ADASIGNSGD (ours)	**1**	**0**	**1**	$\boldsymbol{m}$	–

State notation. $\boldsymbol{m}$: momentum or first moment; $\boldsymbol{v}$: variance or second moment; $\widehat{\boldsymbol{v}}$: frozen variance; $\boldsymbol{\mu}$: mean; $\boldsymbol{c}$: second central moment; $\boldsymbol{s}$: adaptive scale; $\boldsymbol{a}$: accumulator; $\boldsymbol{e}_{\mathrm{w}}$ and $\boldsymbol{e}_{\mathrm{g}}$: worker and global residuals. Birder is reported without its optional AMSGrad-style pre-max buffer, which would add one local state. FP denotes full precision.

For a model with d trainable parameters and state element width b_o bytes, the sharded persistent-state storage predicted by the count is

$$M_o^{\mathrm{state}}(N) = d\, b_o \left(L_o + \frac{G_o}{N} \right). \tag{81}$$

Equation (81) is evaluated using the actual state dtype: tensor counts are dtype-independent, whereas storage in GiB is not.

G.2 PHYSICAL MEMORY MEASUREMENT PROTOCOL

Hardware and workloads. We use a controlled FP32 storage probe at sequence lengths 256, 512, and 1024, with world size $N = 4$, per-GPU microbatch size $B_{\mathrm{mb}} = 1$, and one microstep per optimizer step ($A = 1$). The global sequence batch is therefore four, corresponding to global token batches of $1{,}024$, $2{,}048$, and $4{,}096$, respectively. This probe fixes the per-GPU microbatch rather than the optimizer-step token batch, allowing peak memory to be compared under a common local activation workload.

The 60M, 130M, and 350M models are measured on four P100 GPUs within one node, while the 1B model is measured on four H200 GPUs within one node. Comparisons are made only within each model–hardware block. Parameters, gradients, and persistent model-sized optimizer states use FP32 storage, while each method retains its native communication representation. Activation checkpointing is enabled for all reported configurations.

The complete optimizer suite is evaluated at every sequence length. Within each model and sequence-length block, all methods use the same sequence packing, DDP bucket configuration, gradient-bucket view setting, kernels, and CUDA allocator configuration. Each model–optimizer configuration is executed in a fresh distributed process invocation so that allocator and process-group state from a preceding measurement cannot carry into the next configuration. We additionally run P100 consistency probes at $N \in \{8, 16, 32\}$ under the same local workload to verify the predicted world-size dependence of sharded global state.

Quiescent persistent memory. To include all lazily allocated states, we measure memory independently on each rank after the following synchronized initialization:

1. construct and wrap the model in DDP, stabilize the reducer buckets with two backward passes, synchronize, and record the allocator-tracked baseline $M_{\mathrm{model},n}$;
2. construct the optimizer, register its communication path, and execute the method-specific initialization needed to instantiate every persistent state;

3. release gradients and temporary references, synchronize, and record the quiescent allocator-tracked allocation $M_{\mathrm{quiet},o,n}$.

For methods with phase-dependent state, initialization covers every relevant phase before the quiescent measurement. In particular, both the dense and compressed phases of 1-Bit Adam are exercised, while the variance-refresh, local, and synchronization paths of 0/1 Adam are initialized. The numerical model/optimizer state and RNG/input stream are then restored before peak measurement so that initialization itself does not alter the measured trajectory.

The implementation-level persistent delta on rank n is

$$\Delta M_{o,n}^{\mathrm{persistent}} := M_{\mathrm{quiet},o,n} - M_{\mathrm{model},n}. \tag{82}$$

The table reports the mean per-GPU persistent delta,

$$\overline{\Delta M}_{o}^{\mathrm{persistent}} := \frac{1}{N}\sum_{n=1}^{N} \Delta M_{o,n}^{\mathrm{persistent}}. \tag{83}$$

Because every model-sized persistent state uses FP32 in this controlled measurement, the measured equivalent-state count is

$$\widehat{\nu}_{o}^{\mathrm{state}} := \frac{\overline{\Delta M}_{o}^{\mathrm{persistent}}}{4d}, \tag{84}$$

where d is the exact number of trainable parameters and the denominator is expressed in bytes.

PyTorch's allocator-tracked memory is attributable to each process and rank. Tensor names, shapes, dtypes, and storage pointers verify ownership and sharding against this allocation. The device-global used-memory delta from `cudaMemGetInfo` diagnoses allocator-external CUDA/NCCL storage but includes device-global caching effects; it does not enter $\widehat{\nu}_{o}^{\mathrm{state}}$.

Steady-state peak memory. After persistent-state initialization, each relevant step type is warmed up for 20 steps. We then synchronize, reset the CUDA peak-memory statistics, and measure the following 50 steps. The reported peak is

$$M_{o}^{\mathrm{peak}} := \max_{n,\pi,t} M_{o,n,\pi}^{\mathrm{allocated}}(t), \tag{85}$$

where n, π, and t index ranks, phase-dependent step types, and measured steps. The corresponding `max_memory_reserved` value is reported as an allocator diagnostic.

Validation, training-state checkpoint writes, and experiment logging are disabled. PyTorch-visible compression buffers, packed collective payloads, vote aggregation, and update temporaries are excluded from the persistent-state count but included in the complete-step peak.

Temporary vote aggregation. The conceptual matrix $\mathbf{B}_t \in \{0,1\}^{d\times N}$ in Equation 72 is never materialized. Local signs are bit-packed before communication and routed to temporary owner ranks. Each owner accumulates only its approximately d/N coordinate shard in an unsigned 8-bit vote-count buffer. The reduced counts are then redistributed by VoteReduce using either $w_N = \lceil \log_2(N+1) \rceil$-bit packing or byte-aligned uint8 encoding. The q-bit path packs the owner counts into bit planes before redistribution; the uint8 path redistributes the byte-aligned counts directly. Both decode into the existing live DDP bucket, so neither path materializes an additional full-size floating-point count tensor. All packing, routing, decoding, and chunk-level temporaries are included in the peak-memory measurement.

For 1B/H200, the complete 33-configuration sweep is repeated in a second independent job. Persistent entries below average the two repetition-level rank means, while peak allocated and reserved entries take the larger maximum across the two repetitions. Peak allocated memory differs by at most 0.7 MiB between the two H200 sweeps for every configuration.

G.3 Measured Persistent and Peak Memory

Tables 34–36 report the controlled $N = 4$ measurements. The allocator-tracked persistent state closely follows the analytical count. Across all reported $N = 4$ configurations, the largest relative deviation between $\widehat{\nu}_{o}^{\mathrm{state}}$ and $L_o + G_o/N$ is below 0.26%. For AdaSignSGD, the measured

Table 34: Persistent and peak memory at sequence length 256. All memory values are GiB/GPU.

Optimizer	Predicted equiv.	Measured equiv.	Persistent delta (GiB)	Peak allocated (GiB)	Peak reserved (GiB)
60M, 4×*P100*					
AdamW	2	2.004	0.434	1.020	1.107
ADASIGNSGD (FP)	2	2.004	0.434	1.020	1.107
SIGNUM-MV	1	1.002	0.217	0.803	0.891
SIGNUM-MV+EF	2.25	2.254	0.488	1.175	1.447
Birder	2	2.004	0.434	1.106	1.264
Birder+EC	3.25	3.256	0.704	1.291	1.387
1-Bit Adam	3.25	3.256	0.704	1.291	1.465
0/1 Adam	4.25	4.258	0.921	1.652	2.037
D-Lion-MaVo	1	1.002	0.217	0.803	0.891
ADASIGNSGD (q-bit)	**1**	1.002	0.217	0.803	0.891
ADASIGNSGD (uint8)	**1**	1.002	0.217	0.803	0.891
130M, 4×*P100*					
AdamW	2	2.002	1.000	2.195	2.320
ADASIGNSGD (FP)	2	2.002	1.000	2.195	2.320
SIGNUM-MV	1	1.001	0.500	1.695	1.820
SIGNUM-MV+EF	2.25	2.253	1.125	2.430	2.629
Birder	2	2.002	1.000	2.287	2.459
Birder+EC	3.25	3.254	1.625	2.821	2.945
1-Bit Adam	3.25	3.254	1.625	2.853	2.969
0/1 Adam	4.25	4.254	2.125	3.533	3.859
D-Lion-MaVo	1	1.001	0.500	1.695	1.820
ADASIGNSGD (q-bit)	**1**	1.001	0.500	1.695	1.820
ADASIGNSGD (uint8)	**1**	1.001	0.500	1.695	1.820
350M, 4×*P100*					
AdamW	2	2.000	2.742	5.740	5.986
ADASIGNSGD (FP)	2	2.000	2.742	5.740	5.986
SIGNUM-MV	1	1.000	1.371	4.369	4.615
SIGNUM-MV+EF	2.25	2.252	3.087	6.312	6.793
Birder	2	2.000	2.742	5.874	6.234
Birder+EC	3.25	3.253	4.459	7.456	7.674
1-Bit Adam	3.25	3.253	4.459	7.684	7.928
0/1 Adam	4.25	4.253	5.830	9.286	10.035
D-Lion-MaVo	1	1.000	1.371	4.369	4.615
ADASIGNSGD (q-bit)	**1**	1.000	1.371	4.369	4.615
ADASIGNSGD (uint8)	**1**	1.000	1.371	4.369	4.615
1B, 4×*H200*					
AdamW	2	2.000	9.977	20.404	20.789
ADASIGNSGD (FP)	2	2.000	9.977	20.404	20.789
SIGNUM-MV	1	1.000	4.988	15.416	15.811
SIGNUM-MV+EF	2.25	2.250	11.226	22.732	23.170
Birder	2	2.000	9.977	20.728	21.967
Birder+EC	3.25	3.250	16.215	26.642	27.064
1-Bit Adam	3.25	3.250	16.215	27.720	28.225
0/1 Adam	4.25	4.250	21.203	33.126	35.039
D-Lion-MaVo	1	1.000	4.988	15.416	15.811
ADASIGNSGD (q-bit)	**1**	1.000	4.988	15.416	15.811
ADASIGNSGD (uint8)	**1**	1.000	4.988	15.416	15.801

persistent allocation is essentially one FP32 model-sized state at every model scale: 0.217 GiB at 60M, 0.500 GiB at 130M, 1.371 GiB at 350M, and 4.988 GiB at 1B. These values are, to allocator precision, one half of the corresponding AdamW persistent allocation.

At sequence length 256, the reduction in persistent state remains visible in the complete training step. Relative to AdamW, ADASIGNSGD reduces peak allocated memory by 21.3% at 60M, 22.8% at 130M, 23.9% at 350M, and 24.4% at 1B. At 1B, the peak is 15.416 GiB for ADASIGNSGD (q-bit), compared with 20.404 GiB for AdamW, 22.732 GiB for SIGNUM-MV+EF, 26.642 GiB

Table 35: Persistent and peak memory at sequence length 512. All memory values are GiB/GPU.

Optimizer	Predicted equiv.	Measured equiv.	Persistent delta (GiB)	Peak allocated (GiB)	Peak reserved (GiB)
60M, 4×*P100*					
AdamW	2	2.004	0.434	1.145	1.229
ADASIGNSGD (FP)	2	2.004	0.434	1.145	1.229
SIGNUM-MV	1	1.002	0.217	0.929	1.012
SIGNUM-MV+EF	2.25	2.256	0.488	1.200	1.525
Birder	2	2.004	0.434	1.145	1.385
Birder+EC	3.25	3.258	0.705	1.417	1.465
1-Bit Adam	3.25	3.258	0.705	1.417	1.543
0/1 Adam	4.25	4.260	0.922	1.652	2.115
D-Lion-MaVo	1	1.002	0.217	0.929	1.014
ADASIGNSGD (q-bit)	**1**	1.002	0.217	0.929	1.012
ADASIGNSGD (uint8)	**1**	1.002	0.217	0.929	1.012
130M, 4×*P100*					
AdamW	2	2.002	1.000	2.298	2.486
ADASIGNSGD (FP)	2	2.002	1.000	2.298	2.486
SIGNUM-MV	1	1.001	0.500	1.798	1.994
SIGNUM-MV+EF	2.25	2.253	1.125	2.430	2.752
Birder	2	2.002	1.000	2.298	2.578
Birder+EC	3.25	3.254	1.625	2.924	3.068
1-Bit Adam	3.25	3.254	1.625	2.924	3.068
0/1 Adam	4.25	4.254	2.125	3.533	3.844
D-Lion-MaVo	1	1.001	0.500	1.798	1.992
ADASIGNSGD (q-bit)	**1**	1.001	0.500	1.798	1.990
ADASIGNSGD (uint8)	**1**	1.001	0.500	1.798	1.994
350M, 4×*P100*					
AdamW	2	2.000	2.742	5.829	6.109
ADASIGNSGD (FP)	2	2.000	2.742	5.829	6.109
SIGNUM-MV	1	1.000	1.371	4.458	4.740
SIGNUM-MV+EF	2.25	2.252	3.088	6.313	6.924
Birder	2	2.000	2.742	5.874	6.232
Birder+EC	3.25	3.253	4.459	7.546	7.807
1-Bit Adam	3.25	3.253	4.459	7.684	7.867
0/1 Adam	4.25	4.253	5.830	9.285	9.975
D-Lion-MaVo	1	1.000	1.371	4.458	4.740
ADASIGNSGD (q-bit)	**1**	1.000	1.371	4.458	4.740
ADASIGNSGD (uint8)	**1**	1.000	1.371	4.458	4.742
1B, 4×*H200*					
AdamW	2	2.000	9.977	20.518	20.926
ADASIGNSGD (FP)	2	2.000	9.977	20.519	20.918
SIGNUM-MV	1	1.000	4.988	15.530	15.936
SIGNUM-MV+EF	2.25	2.251	11.227	22.731	23.236
Birder	2	2.000	9.977	20.728	21.902
Birder+EC	3.25	3.251	16.215	26.757	27.252
1-Bit Adam	3.25	3.250	16.215	27.720	28.221
0/1 Adam	4.25	4.251	21.204	33.127	35.016
D-Lion-MaVo	1	1.000	4.988	15.530	15.928
ADASIGNSGD (q-bit)	**1**	1.000	4.988	15.531	15.928
ADASIGNSGD (uint8)	**1**	1.000	4.988	15.530	15.928

for Birder+EC, 27.720 GiB for 1-Bit Adam, and 33.126 GiB for 0/1 Adam. Thus, relative to 1-Bit Adam and 0/1 Adam, the corresponding complete-step peak reductions are 44.4% and 53.5%.

SIGNUM-MV and D-Lion-MaVo also retain one model-sized state and have nearly identical memory footprints in this probe. ADASIGNSGD matches this minimum without a model-sized communication residual; multi-state compressed optimizers additionally retain residuals, frozen statistics, or accumulators.

Table 36: Persistent and peak memory at sequence length 1024. All memory values are GiB/GPU.

Optimizer	Predicted equiv.	Measured equiv.	Persistent delta (GiB)	Peak allocated (GiB)	Peak reserved (GiB)
60M, 4×*P100*					
AdamW	2	2.004	0.434	1.405	1.512
ADASIGNSGD (FP)	2	2.004	0.434	1.405	1.512
SIGNUM-MV	1	1.002	0.217	1.188	1.299
SIGNUM-MV+EF	2.25	2.254	0.488	1.460	1.637
Birder	2	2.004	0.434	1.405	1.510
Birder+EC	3.25	3.256	0.704	1.676	1.857
1-Bit Adam	3.25	3.256	0.704	1.676	1.854
0/1 Adam	4.25	4.258	0.921	1.893	2.070
D-Lion-MaVo	1	1.002	0.217	1.188	1.301
ADASIGNSGD (q-bit)	**1**	1.002	0.217	1.188	1.299
ADASIGNSGD (uint8)	**1**	1.002	0.217	1.188	1.299
130M, 4×*P100*					
AdamW	2	2.002	1.000	2.567	2.664
ADASIGNSGD (FP)	2	2.002	1.000	2.567	2.664
SIGNUM-MV	1	1.001	0.500	2.067	2.176
SIGNUM-MV+EF	2.25	2.254	1.126	2.695	2.951
Birder	2	2.002	1.000	2.567	2.664
Birder+EC	3.25	3.255	1.626	3.195	3.268
1-Bit Adam	3.25	3.255	1.626	3.195	3.268
0/1 Adam	4.25	4.256	2.126	3.695	3.951
D-Lion-MaVo	1	1.001	0.500	2.067	2.174
ADASIGNSGD (q-bit)	**1**	1.001	0.500	2.067	2.172
ADASIGNSGD (uint8)	**1**	1.001	0.500	2.067	2.178
350M, 4×*P100*					
AdamW	2	2.000	2.742	6.125	6.283
ADASIGNSGD (FP)	2	2.000	2.742	6.125	6.285
SIGNUM-MV	1	1.000	1.371	4.754	4.920
SIGNUM-MV+EF	2.25	2.255	3.091	6.473	7.141
Birder	2	2.000	2.742	6.125	6.283
Birder+EC	3.25	3.255	4.462	7.844	8.023
1-Bit Adam	3.25	3.255	4.462	7.844	8.021
0/1 Adam	4.25	4.255	5.833	9.288	10.006
D-Lion-MaVo	1	1.000	1.371	4.754	4.920
ADASIGNSGD (q-bit)	**1**	1.000	1.371	4.754	4.918
ADASIGNSGD (uint8)	**1**	1.000	1.371	4.754	4.920
1B, 4×*H200*					
AdamW	2	2.000	9.977	20.755	21.305
ADASIGNSGD (FP)	2	2.000	9.977	20.755	21.291
SIGNUM-MV	1	1.000	4.988	15.766	16.305
SIGNUM-MV+EF	2.25	2.250	11.225	22.736	23.719
Birder	2	2.000	9.977	20.755	22.029
Birder+EC	3.25	3.250	16.214	26.992	27.730
1-Bit Adam	3.25	3.250	16.214	27.726	28.162
0/1 Adam	4.25	4.250	21.202	33.132	34.980
D-Lion-MaVo	1	1.000	4.988	15.766	16.311
ADASIGNSGD (q-bit)	**1**	1.000	4.988	15.766	16.334
ADASIGNSGD (uint8)	**1**	1.000	4.988	15.766	16.301

Table 37: World-size consistency of persistent-state accounting. Entries are predicted/measured equivalent-state counts for 60M at sequence length 512 on P100 GPUs. The q-bit and uint8 ADASIGNSGD variants have identical persistent-state counts; q-bit is shown.

Optimizer	$N = 4$	$N = 8$	$N = 16$	$N = 32$
SIGNUM-MV	1.000/1.002	1.000/1.002	1.000/1.002	1.000/1.002
SIGNUM-MV+EF	2.250/2.256	2.125/2.134	2.062/2.069	2.031/2.035
Birder+EC	3.250/3.258	3.125/3.136	3.062/3.071	3.031/3.038
1-Bit Adam	3.250/3.258	3.125/3.136	3.062/3.071	3.031/3.038
0/1 Adam	4.250/4.260	4.125/4.138	4.062/4.073	4.031/4.040
D-Lion-MaVo	1.000/1.002	1.000/1.002	1.000/1.002	1.000/1.002
ADASIGNSGD (q-bit)	1.000/1.002	1.000/1.002	1.000/1.002	1.000/1.002

Longer sequences increase the common activation component of the step, but the state-derived saving remains visible. For 60M, ADASIGNSGD's peak reduction relative to AdamW changes from 21.3% at sequence length 256 to 18.9% at 512 and 15.4% at 1024 as activations occupy a larger fraction of the footprint. At 1B, the same reduction remains essentially constant: 24.4%, 24.3%, and 24.0% at sequence lengths 256, 512, and 1024, respectively. At sequence length 1024, ADASIGNSGD peaks at 15.766 GiB, versus 20.755 GiB for AdamW, 27.726 GiB for 1-Bit Adam, and 33.132 GiB for 0/1 Adam.

The two VoteReduce count encodings are indistinguishable at the level of peak allocated memory. Across all 36 matched P100 settings spanning $N \in \{4, 8, 16, 32\}$, three model sizes, and three sequence lengths, q-bit and uint8 produce identical reported peak allocated values. On H200, their largest difference is below 0.7 MiB. This is consistent with the implementation: both variants retain the same one-state optimizer footprint and the same owner-sharded uint8 vote accumulator, while only the redistributed count representation changes. Reserved memory is somewhat noisier because it reflects allocator caching and is therefore treated only as a diagnostic.

World-size consistency. The additional P100 world-size sweep verifies the $L + G/N$ accounting without changing the primary $N = 4$ comparison. Table 37 shows a complete 60M, sequence-length-512 slice. One-state methods remain flat as N grows, whereas methods with one sharded global residual approach their worker-local count at the predicted $1/N$ rate. Across all available P100 world-size measurements, the largest relative deviation between the allocator-tracked equivalent-state count and the analytical prediction is below 0.43%.

The allocation measurements confirm that ADASIGNSGD retains one model-sized state without a replicated residual or statistic. Its state savings remain visible in complete-step peaks, including communication temporaries: ADASIGNSGD matches the lowest persistent footprint in the evaluated 1-bit suite and uses less peak memory than AdamW and the multi-state compressed Adam-style baselines.

H COMMUNICATION-PATH VOLUME AND TIME

We account for average per-rank model-sized traffic relative to FP32 dense DDP and measure the implemented communication path when synchronization is exposed. P100 provides a broad matched-stack comparison through $N = 128$ in a communication-constrained regime. H200 compares hardware-optimized VoteReduce with optimized dense DDP; restricting this comparison avoids confounding algorithmic differences with uneven method-specific kernel engineering.

H.1 COMMUNICATION-VOLUME ACCOUNTING

Unless stated otherwise, V denotes average per-rank send volume for model-sized payloads. Sub-model-size metadata is excluded from analytical payloads but included in physical timing. Let d denote the number of communicated coordinates and N the world size. For an FP32 tensor, the standard reduce-scatter/all-gather ring model gives the per-rank send volume

$$V_{\mathrm{DDP}}(N) = 8\frac{N-1}{N}d \quad \text{bytes.} \tag{86}$$

We use this quantity only as a common accounting reference.

For the two-way 1-bit paths used by SIGNUM-MV, D-Lion-MaVo, and the compressed phases of the error-feedback baselines, the model-sized component is

$$V_{\mathrm{1bit}}(N) = \frac{1}{32} V_{\mathrm{DDP}}(N). \tag{87}$$

Scaled-sign paths additionally communicate one FP32 scale per bucket for each compressed exchange. This $O(K)$ metadata, where K is the number of communication buckets, is excluded from the $\Theta(d)$ analytical payload but included in all physical timing measurements.

Birder uses its hierarchical 1-bit communication path. With G GPUs per node, H nodes, total worker count $N = GH$, and FP32 payload $P = 4d$ bytes, its steady-state per-rank volume is

$$V_{\mathrm{Birder}}(N) = P\left[\frac{G-1}{G} + \frac{G-1}{32G} + \frac{2(H-1)}{32GH}\right]. \tag{88}$$

The three terms correspond respectively to the intra-node FP32 reduce-scatter, the intra-node 1-bit all-gather, and the two inter-node 1-bit exchanges. We evaluate (88) with $G = 4$, matching the four-GPU-per-node layout used in our physical benchmarks. Birder+EC has the same communication volume because its compensation modifies the local residual but not the transmitted representation.

For 1-Bit Adam, let T_c be the number of full-precision steps and T the training horizon. Its exact model-sized average is

$$\overline{V}_{\mathrm{1bitAdam}} = \frac{T_c}{T} V_{\mathrm{DDP}} + \left(1 - \frac{T_c}{T}\right) V_{\mathrm{1bit}}, \tag{89}$$

with $T_c/T = 0.10$ in our experiments. For 0/1 Adam, we use the fixed $L = 4$, $T_v = 8$ configuration selected by our baseline hyperparameter sweep, where L is the compressed-synchronization interval and T_v is the full-precision variance-refresh interval. Let $K_u(T)$ and $K_v(T)$ denote the numbers of compressed synchronization and full-precision variance refreshes, respectively. Under our configured schedule,

$$K_u(T) = \left\lceil \frac{T}{4} \right\rceil, \qquad K_v(T) = \left\lceil \frac{T}{8} \right\rceil, \tag{90}$$

and

$$\overline{V}_{0/1} = \frac{K_u(T)}{T} V_{\mathrm{1bit}} + \frac{K_v(T)}{T} V_{\mathrm{DDP}}. \tag{91}$$

When compressed synchronization and variance refresh coincide, both corresponding payloads are included in the accounting above. Under these configurations, the model-sized traffic ratio is 12.81% for 1-Bit Adam, while the fixed-schedule 0/1 Adam ratio approaches 13.28% as T grows. Finite-horizon values use (89) and (91) exactly.

VoteReduce. Let $b_{t,i}^{(n)} \in \{0,1\}$ denote worker n's packed sign bit for coordinate i. AdaSignSGD requires only the exact vote count $c_{t,i} = \sum_{n=1}^{N} b_{t,i}^{(n)}$, rather than the identity of the contributing workers. VoteReduce partitions coordinates across temporary owners, routes packed 1-bit sign shards to those owners, and redistributes only the exact reduced counts. The minimum exact width is $w_N = \lceil \log_2(N+1) \rceil$ bits, while the byte-aligned implementation uses one `uint8` per count, which is exact for all evaluated $N \le 128$. Their per-rank volumes are

$$V_{\text{VR-q}}(N) = \frac{w_N + 1}{8} \frac{N-1}{N} d, \tag{92}$$

$$V_{\text{VR-u8}}(N) = \frac{9}{8} \frac{N-1}{N} d. \tag{93}$$

The 1-bit term accounts for routing packed sign shards to temporary owners, while the w_N-bit (or 8-bit) term accounts for redistributing the reduced count shards. Both compute exactly the same vote counts and therefore exactly the same AdaSignSGD update; they differ only in count packing and transport.

We report the relative reduction $100\big(1 - V_o(N)/V_{\mathrm{DDP}}(N)\big)\%$. Both the analytical accounting and physical timing extend to $N = 128$.

Table 38: Analytical per-rank communication reduction. Reduction in model-sized send volume relative to FP32 dense DDP (%); higher is better. The accounting and schedules are defined in Section H.1.

Optimizer	Implemented path / schedule	Reduction vs. FP32 dense DDP (%)					
		$N = 4$	$N = 8$	$N = 16$	$N = 32$	$N = 64$	$N = 128$
Full-precision reference							
AdamW	Dense DDP	0.00	0.00	0.00	0.00	0.00	0.00
Communication-efficient baselines							
SIGNUM-MV	1-bit majority	96.88	96.88	96.88	96.88	96.88	96.88
D-Lion-MaVo	1-bit majority	96.88	96.88	96.88	96.88	96.88	96.88
SIGNUM-MV+EF	Scaled-sign + EF	96.88	96.88	96.88	96.88	96.88	96.88
Birder	Hierarchical 1-bit	48.44	55.36	58.13	59.38	59.97	60.26
Birder+EC	Hierarchical 1-bit + EC	48.44	55.36	58.13	59.38	59.97	60.26
1-Bit Adam	10% dense + 1-bit EF	≈ 87.19	≈ 87.19	≈ 87.19	≈ 87.19	≈ 87.19	≈ 87.19
0/1 Adam	$L = 4,\ T_v = 8$	≈ 86.72	≈ 86.72	≈ 86.72	≈ 86.72	≈ 86.72	≈ 86.72
ADASIGNSGD *implementations*							
ADASIGNSGD (q-bit)	VoteReduce	93.75	92.19	90.63	89.06	87.50	85.94
ADASIGNSGD (uint8)	VoteReduce	85.94	85.94	85.94	85.94	85.94	85.94

Sign-only methods retain only the aggregate direction, whereas VoteReduce preserves the exact agreement count for adaptive scaling (Table 38). At $N = 32$, the w_N-bit and `uint8` variants reduce model-sized traffic by 89.06% and 85.94%, respectively.

H.2 COMMUNICATION-PATH TIME

The timed communication path begins when the local gradient payload is available and ends when the synchronized optimizer representation is ready. It includes applicable sign extraction, scaling, quantization, error-feedback processing, packing, communication, reduction or reconstruction, and decoding. For ADASIGNSGD, this includes VoteReduce routing, vote-count reduction, count redistribution, unpacking, and count-to-scale lookup. The parameter update is excluded.

For the P100 suite, each table cell is measured in an independent job. Both hardware suites use 100 warm-up calls followed by 500 measured calls per stationary or phase-specific path. Within each model–hardware block, all methods operate on the same logical model payload and use the same model-level bucketization: 25 MiB main buckets and a 1 MiB first bucket. Encoding, physical message size, and collective structure follow each method's implemented communication path.

For stationary paths, the reported center is the mean over all measured communication-path calls. For methods with nonstationary communication, we form the schedule-weighted mean from the phase-specific measurements using the exact training schedule. In particular, for 1-Bit Adam,

$$\overline{\tau}_{\mathrm{1bitAdam}} = \frac{T_c}{T}\tau_{\mathrm{dense}} + \left(1 - \frac{T_c}{T}\right)\tau_{\mathrm{comp}}, \tag{94}$$

where τ_{dense} and τ_{comp} are the mean communication-path times of the corresponding phases, and T is the finite training horizon used to instantiate the corresponding schedule. For 0/1 Adam, we analogously amortize over the exact fixed $L = 4, T_v = 8$ schedule, including local-only, compressed-synchronization, variance-refresh, and coincident synchronization–refresh steps.

For the P100 tables, the reported percentile interval is computed from the corresponding schedule-weighted empirical per-step time distribution. Consequently, the mean need not lie inside the reported 10th–90th percentile interval: the 10% dense phase of 1-Bit Adam can occupy the upper tail of the mixture, while local-only steps in 0/1 Adam contribute zero communication-path time. These percentiles summarize within-job call-to-call variability rather than confidence intervals across independent allocations.

Across the completed P100 settings in Tables 39 and 40, ADASIGNSGD `uint8` reduces communication-path time relative to dense DDP by 34.1–80.3%. At $N = 32$, the reductions are 57.4%, 56.0%, and 57.9% for 60M, 130M, and 350M, respectively; at $N = 128$, the 60M, 130M, and 350M measurements yield 79.8%, 80.3%, and 80.0% reductions, respectively.

Table 39: P100 matched-stack communication-path time at $N \in \{4, 8, 16\}$. Mean milliseconds per scheduled optimizer step; parentheses give the within-job 10th–90th percentile under the schedule-weighted protocol.

Optimizer	Implemented path / schedule	$N = 4$	$N = 8$	$N = 16$
60M				
AdamW	Dense DDP	43.0 (42.9–43.2)	53.3 (53.1–53.4)	53.4 (53.3–53.5)
SIGNUM-MV	1-bit majority	21.5 (21.4–21.5)	21.8 (21.8–21.9)	23.3 (23.2–23.4)
D-Lion-MaVo	1-bit majority	24.5 (24.5–24.6)	24.8 (24.8–24.9)	26.3 (26.3–26.4)
SIGNUM-MV+EF	Scaled-sign + EF	38.3 (38.3–38.3)	37.3 (37.3–37.4)	38.1 (38.0–38.2)
Birder	Hierarchical 1-bit	43.9 (43.8–43.9)	53.1 (53.0–53.2)	50.1 (50.0–50.2)
Birder+EC	Hierarchical 1-bit + EC	45.5 (45.4–45.6)	55.7 (55.6–55.8)	52.2 (52.1–52.3)
1-Bit Adam	10% dense + 1-bit EF	41.8 (41.6–41.9)	41.9 (40.6–41.2)	42.6 (41.3–42.7)
0/1 Adam	$L = 4,\ T_v = 8$	15.6 (0.0–85.5)	16.6 (0.0–94.3)	16.9 (0.0–95.7)
AdaSignSGD (q-bit)	VoteReduce	43.8 (43.8–43.9)	48.7 (48.7–48.8)	56.0 (56.0–56.1)
AdaSignSGD (uint8)	VoteReduce	28.4 (28.3–28.4)	30.2 (30.1–30.3)	31.3 (31.2–31.3)
130M				
AdamW	Dense DDP	105.2 (105.0–105.4)	123.8 (123.2–124.6)	151.8 (151.5–152.1)
SIGNUM-MV	1-bit majority	51.3 (51.3–51.4)	54.0 (54.0–54.1)	58.0 (57.9–58.1)
D-Lion-MaVo	1-bit majority	58.6 (58.5–58.7)	61.4 (61.4–61.5)	65.1 (65.0–65.3)
SIGNUM-MV+EF	Scaled-sign + EF	90.6 (90.6–90.7)	89.8 (89.7–89.8)	91.9 (91.7–92.0)
Birder	Hierarchical 1-bit	105.5 (105.4–105.6)	122.0 (121.8–122.3)	126.8 (126.4–127.2)
Birder+EC	Hierarchical 1-bit + EC	109.4 (109.2–109.5)	128.4 (127.8–128.7)	132.1 (131.2–132.7)
1-Bit Adam	10% dense + 1-bit EF	98.8 (98.0–98.3)	99.8 (97.1–97.6)	104.4 (99.0–99.9)
0/1 Adam	$L = 4,\ T_v = 8$	37.4 (0.0–206.5)	39.5 (0.0–223.7)	43.0 (0.0–249.5)
AdaSignSGD (q-bit)	VoteReduce	103.9 (103.8–103.9)	115.7 (115.6–115.7)	138.7 (138.6–138.8)
AdaSignSGD (uint8)	VoteReduce	67.8 (67.7–67.8)	72.6 (72.6–72.7)	79.1 (78.9–79.2)
350M				
AdamW	Dense DDP	309.2 (308.4–309.9)	375.4 (374.5–376.3)	368.9 (368.0–369.8)
SIGNUM-MV	1-bit majority	142.0 (141.9–142.1)	142.8 (142.6–143.1)	149.2 (149.1–149.4)
D-Lion-MaVo	1-bit majority	162.2 (162.1–162.3)	161.8 (161.7–161.8)	168.9 (168.8–169.0)
SIGNUM-MV+EF	Scaled-sign + EF	249.9 (249.8–250.0)	241.6 (241.5–241.8)	242.7 (242.6–242.8)
Birder	Hierarchical 1-bit	295.7 (295.4–295.9)	354.5 (353.9–355.1)	339.6 (339.1–340.2)
Birder+EC	Hierarchical 1-bit + EC	306.0 (305.8–306.3)	373.2 (370.9–374.2)	356.6 (356.2–357.1)
1-Bit Adam	10% dense + 1-bit EF	273.9 (269.8–270.8)	273.0 (261.4–262.3)	272.7 (261.9–262.7)
0/1 Adam	$L = 4,\ T_v = 8$	104.8 (0.0–582.6)	111.1 (0.0–640.1)	110.5 (0.0–634.9)
AdaSignSGD (q-bit)	VoteReduce	286.1 (286.0–286.2)	316.0 (315.8–316.2)	348.2 (348.1–348.3)
AdaSignSGD (uint8)	VoteReduce	189.5 (189.4–189.6)	198.7 (198.4–199.0)	202.3 (202.2–202.4)

Table 40: P100 matched-stack communication-path time at $N \in \{32, 64, 128\}$. Mean milliseconds per scheduled optimizer step; parentheses give the within-job 10th–90th percentile under the schedule-weighted protocol.

Optimizer	Implemented path / schedule	$N = 32$	$N = 64$	$N = 128$
60M				
AdamW	Dense DDP	88.4 (88.2–88.7)	90.1 (89.8–90.4)	718.2 (717.4–719.0)
SIGNUM-MV	1-bit majority	25.9 (25.8–26.0)	27.8 (27.3–28.2)	61.7 (61.4–62.2)
D-Lion-MaVo	1-bit majority	29.0 (28.9–29.0)	30.9 (30.4–31.3)	64.1 (63.8–64.5)
SIGNUM-MV+EF	Scaled-sign + EF	40.7 (40.6–40.7)	42.6 (42.2–43.0)	78.5 (78.0–78.9)
Birder	Hierarchical 1-bit	68.3 (68.2–68.5)	68.5 (68.3–68.7)	388.0 (387.4–388.6)
Birder+EC	Hierarchical 1-bit + EC	70.4 (70.2–70.5)	70.4 (70.2–70.6)	390.2 (389.6–390.8)
1-Bit Adam	10% dense + 1-bit EF	48.3 (43.8–45.7)	50.2 (45.4–51.6)	145.4 (81.4–83.3)
0/1 Adam	$L = 4,\ T_v = 8$	21.9 (0.0–133.5)	22.6 (0.0–136.4)	110.2 (0.0–801.2)
ADASIGNSGD (q-bit)	VoteReduce	66.0 (66.0–66.1)	77.0 (76.9–77.1)	181.9 (181.5–182.4)
ADASIGNSGD (uint8)	VoteReduce	37.7 (37.6–37.8)	40.5 (40.4–40.7)	144.8 (144.3–145.2)
130M				
AdamW	Dense DDP	210.1 (209.7–210.5)	213.4 (212.9–214.0)	1663.6 (1661.7–1665.5)
SIGNUM-MV	1-bit majority	61.9 (61.7–62.0)	66.2 (65.9–66.6)	150.8 (150.2–151.4)
D-Lion-MaVo	1-bit majority	69.0 (68.9–69.1)	75.6 (75.0–76.3)	157.4 (156.8–157.9)
SIGNUM-MV+EF	Scaled-sign + EF	96.0 (95.8–96.1)	100.3 (100.0–100.6)	190.0 (189.4–190.7)
Birder	Hierarchical 1-bit	162.7 (162.4–163.0)	164.0 (163.6–164.7)	908.8 (907.7–909.8)
Birder+EC	Hierarchical 1-bit + EC	168.1 (167.2–168.6)	169.2 (168.6–170.4)	913.9 (912.8–914.8)
1-Bit Adam	10% dense + 1-bit EF	114.2 (103.4–104.6)	118.6 (107.5–112.8)	343.8 (196.4–199.8)
0/1 Adam	$L = 4,\ T_v = 8$	52.0 (0.0–317.2)	53.4 (0.0–323.7)	257.3 (0.0–1865.0)
ADASIGNSGD (q-bit)	VoteReduce	162.1 (161.9–162.2)	186.0 (185.7–186.3)	415.7 (415.0–416.6)
ADASIGNSGD (uint8)	VoteReduce	92.4 (92.2–92.5)	107.6 (107.4–107.9)	328.2 (327.5–328.9)
350M				
AdamW	Dense DDP	601.6 (600.7–602.5)	611.0 (610.0–612.1)	4564.0 (4561.4–4566.8)
SIGNUM-MV	1-bit majority	171.8 (171.4–172.2)	180.0 (179.4–180.8)	413.9 (413.0–414.9)
D-Lion-MaVo	1-bit majority	191.3 (191.2–191.5)	201.1 (200.5–201.8)	436.2 (435.3–437.2)
SIGNUM-MV+EF	Scaled-sign + EF	264.0 (263.7–264.4)	272.2 (270.9–273.4)	513.9 (508.5–519.1)
Birder	Hierarchical 1-bit	454.6 (453.5–455.6)	461.1 (458.2–464.2)	2519.6 (2516.1–2523.1)
Birder+EC	Hierarchical 1-bit + EC	470.1 (466.8–471.7)	475.8 (472.0–479.6)	2536.2 (2534.6–2537.9)
1-Bit Adam	10% dense + 1-bit EF	315.7 (283.6–286.2)	324.8 (292.3–296.4)	937.0 (533.0–536.5)
0/1 Adam	$L = 4,\ T_v = 8$	145.4 (0.0–892.4)	148.0 (0.0–905.5)	702.5 (0.0–5099.0)
ADASIGNSGD (q-bit)	VoteReduce	420.1 (419.6–420.5)	502.2 (501.6–502.8)	1151.5 (1149.9–1153.0)
ADASIGNSGD (uint8)	VoteReduce	253.5 (253.1–253.9)	274.3 (273.6–275.1)	910.8 (909.1–912.6)

Table 41: H200 isolated communication-path time. Mean milliseconds; parentheses give ADASIGNSGD's speedup over dense DDP in the same model–world-size block.

Optimizer	Implemented path	$N = 4$	$N = 8$	$N = 16$	$N = 32$
60M					
AdamW	Dense DDP	1.338	3.178	4.581	4.763
ADASIGNSGD (q-bit)	VoteReduce	0.983 (1.36×)	1.461 (2.17×)	1.886 (2.43×)	2.444 (1.95×)
ADASIGNSGD (uint8)	VoteReduce	0.877 (1.52×)	1.548 (2.05×)	1.858 (2.47×)	2.243 (2.12×)
130M					
AdamW	Dense DDP	3.085	6.993	10.601	11.027
ADASIGNSGD (q-bit)	VoteReduce	2.339 (1.32×)	3.598 (1.94×)	4.680 (2.27×)	5.855 (1.88×)
ADASIGNSGD (uint8)	VoteReduce	2.193 (1.41×)	3.683 (1.90×)	4.396 (2.41×)	5.579 (1.98×)
350M					
AdamW	Dense DDP	8.436	19.749	29.078	31.728
ADASIGNSGD (q-bit)	VoteReduce	6.563 (1.29×)	9.599 (2.06×)	12.496 (2.33×)	16.644 (1.91×)
ADASIGNSGD (uint8)	VoteReduce	5.943 (1.42×)	10.253 (1.93×)	11.672 (2.49×)	15.889 (2.00×)
1B					
AdamW	Dense DDP	29.523	71.212	104.866	114.815
ADASIGNSGD (q-bit)	VoteReduce	21.741 (1.36×)	31.594 (2.25×)	38.849 (2.70×)	57.420 (2.00×)
ADASIGNSGD (uint8)	VoteReduce	19.452 (1.52×)	34.668 (2.05×)	37.125 (2.82×)	52.776 (2.18×)

The `uint8` path is faster than Birder, Birder+EC, and 1-Bit Adam in every completed model–world-size block. It also outperforms SIGNUM-MV+EF in 13 of the 15 settings through $N = 64$, including all 12 through $N = 32$, while retaining the agreement signal for adaptive scaling. The pure 1-bit majority paths, SIGNUM-MV and D-Lion-MaVo, remain faster because they retain only the aggregate direction. The lower schedule-amortized time of 0/1 Adam reflects its many local-only steps, not a cheaper synchronization event.

Despite its smaller analytical payload, the w_N-bit representation is 25.7–90.0% slower than the byte-aligned `uint8` path across the completed P100 settings. On this hardware, the additional minimum-width packing and unpacking cost therefore outweighs the corresponding reduction in count traffic.

H200: optimized modern-hardware communication path. We next test CUDA-optimized implementations of both VoteReduce encodings against dense DDP on H200 GPUs for the 60M, 130M, 350M, and 1B models at $N \in \{4, 8, 16, 32\}$. These configurations occupy one, two, four, and eight four-GPU nodes, respectively. All paths use FP32 with TF32 disabled and follow the common payload, bucketization, and timing protocol defined above. We collect the cases sequentially within one 32-GPU allocation using synchronized max-rank CUDA timing (PyTorch 2.5.1+cu121, NCCL 2.21.5).

Table 41 shows that both VoteReduce encodings outperform dense DDP in every one of the 16 model–world-size blocks, with 1.29–2.82× communication-path speedups. At $N = 32$, optimized VoteReduce approximately halves isolated communication-path time across all four model sizes, yielding 1.88–2.18× speedups. For the 1B model, dense DDP requires 114.815 ms, compared with 57.420 ms for the q-bit path and 52.776 ms for the `uint8` path. At every reported world size, both optimized VoteReduce implementations were separately validated bitwise against the reference under identical gradient inputs. An order-reversed repeat on the same allocation changed each of the three 1B/$N = 32$ means by less than 0.1%.

Neither count encoding uniformly dominates on H200: the q-bit path is faster in all four $N = 8$ blocks, whereas `uint8` is faster at $N \in \{4, 16, 32\}$. Their largest difference is approximately 12%. Together with the P100 results, this indicates that the balance between transmitted bytes and local packing overhead can depend on both implementation and hardware.

H.3 Boundary to Complete-Step Latency

Native bucket overlap determines how much communication remains on the training-step critical path. Appendix I therefore measures the complete forward-to-update interval separately.

I Complete-Step Latency

Complete-step latency is measured from forward-pass start through the final parameter write, including model computation, overlapped and exposed communication, and optimizer work. It is measured directly rather than reconstructed from the isolated communication path in Appendix H.

The main-paper optimization study fixes the global sequence batch at 512. This systems probe instead fixes the per-worker workload, so increasing the world size changes synchronization scale without changing local forward/backward computation.

I.1 Complete-Step Protocol

Hardware roles and workload. The hardware roles follow Appendix H. P100 evaluates 60M, 130M, and 350M at $N \in \{4, 8, 16, 32, 64, 128\}$ using a matched implementation stack. H200 evaluates 60M, 130M, 350M, and 1B at $N \in \{4, 8, 16, 32\}$ on one, two, four, and eight four-GPU nodes, respectively. H200 compares optimized AdaSignSGD with the fastest measured dense AdamW path; prior methods without comparable H200 kernel engineering are excluded.

Both suites use FP32 and sequence length 256. Each worker processes one sequence per optimizer step, with per-GPU microbatch size $B_{\mathrm{mb}} = 1$ and one microstep per optimizer step ($A = 1$). Thus the global sequence batch is N and the global token batch is

$$B_{\mathrm{tok}}(N) = 256N. \tag{95}$$

This weak-scaling probe holds local computation fixed as N increases. Absolute latencies are compared only within each model–world-size block and are not used to compare optimization quality across world sizes.

Within each model–world-size–hardware block, every optimizer uses the same model implementation, token shape, per-worker batch, precision, and parameter-aware bucketization (25 MiB main buckets and a 1 MiB first bucket). Pre-tokenized input batches are prepared before timing. Data loading, host-to-device input staging, validation, checkpointing, experiment logging, and filesystem I/O are excluded. The H200 runs disable TF32 and use PyTorch 2.5.1+cu121 with NCCL 2.21.5.

Timing boundary. A measured step begins immediately before the model forward pass and ends only after the optimizer has completed the corresponding parameter update. The timed region therefore includes forward and backward computation, exposed communication, native communication–computation overlap, method-specific compression and reconstruction, optimizer-state arithmetic, and the final parameter write. Dense AdamW retains standard bucketed DDP with native backward/communication overlap. The H200 reference further uses `torch.optim.AdamW(fused=True)`, which is uniformly faster than the standard AdamW update in the measured matrix.

Before every timed sample, ranks synchronize with an untimed barrier, after which timing begins immediately before the model forward pass. After the parameter update, CUDA work is synchronized and the sample latency is taken as the maximum elapsed time across ranks, so a reported complete step cannot finish before its slowest worker. Each model–optimizer–world-size cell uses a fresh distributed process invocation. The H200 cases are collected sequentially within one 32-GPU allocation to avoid cross-allocation comparisons.

Native optimizer paths and phase schedules. Each method uses the same implemented communication path characterized in Appendix H. The P100 study uses the byte-aligned `uint8` VoteReduce path; the w_N-bit path is omitted there because it is algorithmically identical and consistently slower in the P100 communication benchmark. The H200 study retains both optimized encodings because their realized ordering depends on world size. Both H200 AdaSignSGD paths launch

bucket-level VoteReduce during backward computation and use the fused final parameter update unless an ablation explicitly disables either mechanism.

Stationary methods are timed directly. For optimizers with phase-dependent execution, every distinct step type is timed separately and combined using the same exact schedule as in Appendix H. Let Π_o denote the set of execution phases for optimizer o, $w_{o,\pi}$ the fraction of optimizer steps in phase π, and $\tau_{o,\pi}^{\mathrm{step}}$ the measured mean complete-step latency of that phase. We report

$$\overline{\tau}_o^{\mathrm{step}} = \sum_{\pi \in \Pi_o} w_{o,\pi}\, \tau_{o,\pi}^{\mathrm{step}}. \tag{96}$$

Thus 1-Bit Adam retains its dense-to-compressed schedule, while 0/1 Adam retains its local, synchronization, variance-refresh, and coincident synchronization–refresh steps. Unlike the isolated communication-path measurement, a local-only 0/1 Adam step still incurs the full model forward/backward and local optimizer computation and therefore has nonzero complete-step latency.

Repetitions and statistics. Each stationary or phase-specific path is warmed up for 100 complete steps and measured for the following 500 steps. P100 tables report the mean and within-job 10th–90th percentile interval. For nonstationary methods, we form an empirical mixture over phase-specific samples using the exact schedule frequencies. H200 tables report the mean and a paired speedup within the same model–world-size block; the focused controls additionally report their raw means. These summaries measure call-to-call variation rather than uncertainty across cluster allocations.

I.2 P100 Matched-Stack Results

Tables 42 and 43 compare P100 complete-step latency against dense-DDP AdamW. AdasignSGD uses `uint8` VoteReduce.

At $N = 32$, AdasignSGD requires 75.86, 172.36, and 412.69 ms for the 60M, 130M, and 350M models, compared with 119.32, 231.53, and 610.05 ms for dense AdamW. At $N = 128$, the corresponding AdasignSGD latencies are 209.43, 414.01, and 1075.74 ms, compared with 750.84, 1739.76, and 4719.05 ms, yielding reductions of 72.11%, 76.20%, and 77.20%. Across completed P100 blocks with both measurements, the reduction ranges from 9.41% to 77.20%. The matched-stack communication gains therefore persist after including model and optimizer computation in this communication-constrained regime; P100 is not presented as a modern deployment system.

Table 42: P100 complete-step latency at $N \in \{4, 8, 16\}$. Mean milliseconds; parentheses give the within-job 10th–90th percentile. Phase-dependent methods use exact schedule weights.

Optimizer	Implemented path / schedule	$N = 4$	$N = 8$	$N = 16$
60M				
AdamW	Dense DDP	72.02 (71.84–72.17)	82.25 (82.05–82.43)	93.11 (92.03–98.33)
SIGNUM-MV	1-bit majority	53.99 (53.84–54.16)	58.81 (58.64–58.93)	83.17 (80.07–85.02)
D-Lion-MaVo	1-bit majority	62.83 (62.73–62.90)	61.31 (61.18–61.43)	83.67 (82.34–86.30)
SIGNUM-MV+EF	Scaled-sign + EF	74.77 (74.61–74.91)	73.44 (73.26–73.59)	89.09 (88.32–89.83)
Birder	Hierarchical 1-bit	78.24 (78.12–78.33)	86.58 (86.40–86.77)	111.28 (104.00–117.18)
Birder+EC	Hierarchical 1-bit + EC	81.48 (81.31–81.63)	88.55 (85.05–90.59)	106.71 (106.12–107.14)
1-Bit Adam	10% dense + 1-bit EF	80.80 (75.21–81.71)	79.50 (74.85–81.32)	96.68 (96.22–97.18)
0/1 Adam	$L = 4,\ T_v = 8$	56.26 (39.40–130.84)	60.21 (39.41–144.36)	98.24 (86.13–163.75)
ADASIGNSGD (ours)	VoteReduce (uint8)	65.24 (65.05–65.42)	66.14 (66.03–66.23)	82.45 (82.15–82.86)
130M				
AdamW	Dense DDP	226.72 (226.28–227.15)	248.34 (247.89–248.80)	180.66 (180.05–182.99)
SIGNUM-MV	1-bit majority	113.56 (113.35–113.73)	117.48 (117.15–117.79)	118.91 (118.66–119.13)
D-Lion-MaVo	1-bit majority	119.61 (119.44–119.78)	126.12 (124.02–127.01)	126.58 (126.38–126.76)
SIGNUM-MV+EF	Scaled-sign + EF	155.36 (155.14–155.57)	149.68 (149.50–149.84)	151.67 (151.48–151.86)
Birder	Hierarchical 1-bit	204.26 (203.93–204.55)	220.72 (220.35–221.06)	183.89 (183.62–184.11)
Birder+EC	Hierarchical 1-bit + EC	209.39 (209.13–209.64)	227.48 (227.16–227.79)	187.02 (186.77–187.26)
1-Bit Adam	10% dense + 1-bit EF	172.04 (165.19–166.54)	176.15 (166.99–168.40)	165.44 (163.28–165.30)
0/1 Adam	$L = 4,\ T_v = 8$	119.70 (69.87–362.88)	126.68 (74.11–387.26)	117.07 (74.05–316.37)
ADASIGNSGD (ours)	VoteReduce (uint8)	136.73 (136.55–136.91)	142.86 (142.28–143.12)	137.94 (137.63–138.30)
350M				
AdamW	Dense DDP	422.41 (421.79–423.03)	474.28 (473.39–475.11)	484.62 (484.16–485.06)
SIGNUM-MV	1-bit majority	278.71 (278.46–278.95)	278.62 (278.44–278.81)	285.43 (285.27–285.60)
D-Lion-MaVo	1-bit majority	307.53 (307.21–307.85)	297.97 (297.75–298.20)	312.28 (311.87–312.71)
SIGNUM-MV+EF	Scaled-sign + EF	380.44 (380.18–380.70)	371.41 (371.17–371.66)	373.43 (373.20–373.68)
Birder	Hierarchical 1-bit	434.03 (433.69–434.34)	477.69 (477.36–478.01)	484.36 (483.69–485.85)
Birder+EC	Hierarchical 1-bit + EC	452.05 (451.69–452.43)	496.79 (496.49–497.10)	495.59 (493.98–501.79)
1-Bit Adam	10% dense + 1-bit EF	418.99 (416.93–417.92)	415.90 (407.65–408.61)	418.26 (409.30–411.92)
0/1 Adam	$L = 4,\ T_v = 8$	282.70 (170.66–789.47)	293.75 (178.11–832.51)	297.24 (179.12–852.79)
ADASIGNSGD (ours)	VoteReduce (uint8)	326.37 (326.12–326.65)	333.07 (332.67–333.45)	345.87 (345.97–346.64)

Table 43: P100 complete-step latency at $N \in \{32, 64, 128\}$. Mean milliseconds; parentheses give the within-job 10th–90th percentile. Phase-dependent methods use exact schedule weights.

Optimizer	Implemented path / schedule	$N = 32$	$N = 64$	$N = 128$
60M				
AdamW	Dense DDP	119.32 (118.96–119.69)	127.23 (126.87–127.62)	750.84 (749.87–751.80)
SIGNUM-MV	1-bit majority	64.32 (64.17–64.49)	104.83 (104.41–105.37)	123.05 (122.65–123.53)
D-Lion-MaVo	1-bit majority	67.57 (67.35–67.71)	86.07 (84.92–86.90)	129.64 (129.23–130.04)
SIGNUM-MV+EF	Scaled-sign + EF	78.19 (77.98–78.39)	93.26 (92.73–93.88)	143.39 (142.55–144.02)
Birder	Hierarchical 1-bit	104.21 (103.95–104.46)	128.72 (128.28–129.21)	420.18 (419.55–420.76)
Birder+EC	Hierarchical 1-bit + EC	102.79 (102.57–103.03)	117.81 (117.10–118.55)	422.76 (422.16–423.40)
1-Bit Adam	10% dense + 1-bit EF	91.44 (87.47–89.40)	106.69 (102.54–121.65)	210.05 (148.99–153.15)
0/1 Adam	$L = 4,\ T_v = 8$	68.94 (45.72–185.72)	104.66 (87.32–203.19)	194.29 (85.89–881.08)
AdaSignSGD (ours)	VoteReduce (uint8)	75.86 (75.63–76.14)	94.31 (92.05–97.51)	209.43 (208.75–210.11)
130M				
AdamW	Dense DDP	231.53 (231.11–231.99)	270.12 (269.47–270.58)	1739.76 (1738.45–1741.13)
SIGNUM-MV	1-bit majority	141.18 (140.65–141.75)	153.07 (146.89–165.30)	252.75 (251.77–253.54)
D-Lion-MaVo	1-bit majority	158.50 (147.67–171.40)	153.74 (153.03–154.47)	244.40 (243.71–245.05)
SIGNUM-MV+EF	Scaled-sign + EF	162.86 (161.23–170.90)	179.88 (179.16–180.32)	292.53 (291.74–293.19)
Birder	Hierarchical 1-bit	219.76 (211.05–223.62)	236.08 (235.20–237.34)	969.75 (968.78–970.64)
Birder+EC	Hierarchical 1-bit + EC	226.58 (226.11–227.07)	228.65 (225.37–230.54)	976.97 (976.00–977.90)
1-Bit Adam	10% dense + 1-bit EF	191.86 (186.54–206.75)	183.88 (173.21–177.25)	434.58 (284.25–302.73)
0/1 Adam	$L = 4,\ T_v = 8$	170.17 (128.74–387.71)	131.19 (75.27–412.41)	388.12 (135.57–1983.71)
AdaSignSGD (ours)	VoteReduce (uint8)	172.36 (171.77–173.04)	171.11 (170.75–171.46)	414.01 (410.21–415.48)
350M				
AdamW	Dense DDP	610.05 (609.36–610.49)	741.14 (740.26–742.03)	4719.05 (4716.24–4721.75)
SIGNUM-MV	1-bit majority	324.54 (323.68–325.34)	347.34 (345.86–348.46)	632.64 (629.66–637.35)
D-Lion-MaVo	1-bit majority	350.43 (349.71–350.90)	368.78 (367.56–370.09)	633.42 (631.40–635.05)
SIGNUM-MV+EF	Scaled-sign + EF	404.01 (402.90–405.13)	427.68 (423.77–431.20)	696.37 (692.96–708.17)
Birder	Hierarchical 1-bit	585.62 (563.40–596.78)	615.07 (611.84–618.58)	2676.48 (2674.91–2678.07)
Birder+EC	Hierarchical 1-bit + EC	600.01 (589.05–629.71)	626.62 (622.26–631.70)	2704.59 (2702.69–2706.39)
1-Bit Adam	10% dense + 1-bit EF	463.26 (443.51–447.16)	509.58 (481.78–486.58)	1135.71 (722.32–771.23)
0/1 Adam	$L = 4,\ T_v = 8$	383.44 (262.12–1003.02)	399.39 (250.42–1146.33)	966.26 (259.88–5396.69)
AdaSignSGD (ours)	VoteReduce (uint8)	412.69 (411.99–413.45)	447.58 (434.39–452.81)	1075.74 (1074.03–1077.80)

I.3 H200 Optimized Results

The H200 comparison includes native backward overlap and fused optimizer arithmetic for both optimized VoteReduce encodings and fused AdamW, the fastest measured dense path.

Table 44: H200 optimized complete-step latency. Mean milliseconds at sequence length 256 and per-GPU microbatch one; parentheses give AdaSignSGD's speedup over fused AdamW in the same block. Both VoteReduce encodings use backward overlap and fused final updates.

Optimizer	Implemented path	$N = 4$	$N = 8$	$N = 16$	$N = 32$
60M					
AdamW	Dense DDP + fused update	14.566	15.267	15.780	16.119
AdaSignSGD (q-bit)	Overlapped VoteReduce	17.056 (0.85×)	17.441 (0.88×)	17.616 (0.90×)	17.700 (0.91×)
AdaSignSGD (uint8)	Overlapped VoteReduce	16.469 (0.88×)	17.169 (0.89×)	17.286 (0.91×)	17.745 (0.91×)
130M					
AdamW	Dense DDP + fused update	22.414	22.984	23.949	24.343
AdaSignSGD (q-bit)	Overlapped VoteReduce	26.665 (0.84×)	27.020 (0.85×)	27.170 (0.88×)	27.656 (0.88×)
AdaSignSGD (uint8)	Overlapped VoteReduce	26.587 (0.84×)	26.759 (0.86×)	27.070 (0.88×)	27.163 (0.90×)
350M					
AdamW	Dense DDP + fused update	44.355	45.447	49.081	50.705
AdaSignSGD (q-bit)	Overlapped VoteReduce	52.047 (0.85×)	52.496 (0.87×)	54.022 (0.91×)	54.710 (0.93×)
AdaSignSGD (uint8)	Overlapped VoteReduce	52.915 (0.84×)	53.366 (0.85×)	54.310 (0.90×)	54.859 (0.92×)
1B					
AdamW	Dense DDP + fused update	87.043	108.704	140.510	151.166
AdaSignSGD (q-bit)	Overlapped VoteReduce	83.814 (1.04×)	84.136 (1.29×)	90.207 (1.56×)	107.439 (1.41×)
AdaSignSGD (uint8)	Overlapped VoteReduce	79.386 (1.10×)	84.261 (1.29×)	87.621 (1.60×)	103.216 (1.46×)

Both VoteReduce encodings reduce communication-path latency in all 16 H200 blocks (Table 41), but complete-step latency improves only for 1B (Table 44). For 60M–350M, fused AdamW remains faster, consistent with compression, routing, and launch overhead not being fully amortized once model compute and native dense overlap are restored. For 1B, both AdaSignSGD encodings beat fused AdamW at every world size, with 1.04–$1.60\times$ speedups. At $N = 32$, uint8 and q-bit require 103.216 and 107.439 ms versus 151.166 ms for fused AdamW, reducing complete-step latency by 31.7% and 28.9%, respectively. An order-reversed repeat on the same allocation changes each of these three means by less than 0.1%.

I.4 H200 Implementation and Workload Controls

Table 45 shows the bucket-level backward-overlap schedule. Table 46 reports the overlap–update-fusion factorial and sequence-length control.

Table 45: H200 backward-overlap schedule for 1B, $N = 32$, `uint8`. With final-update fusion disabled, overlap hides 17.924 ms of VoteReduce and reduces the complete step from 121.200 to 103.276 ms. The forward/backward/update partition is schematic; communication durations are constrained by the isolated path and complete-step difference.

Lane	Wall-clock ⟶					Mean step
No backward overlap						
Compute	**Forward**	**Backward**			**Update**	**121.200 ms** reference
Communication		*await backward end*	**VoteReduce: 52.776 ms**			
Bucket-level backward overlap						
Compute	**Forward**	**Backward**		**Update**		**103.276 ms** −14.8%
Communication		*await first bucket*	**17.924 ms**	**Tail: 34.852 ms**		

Table 46: H200 implementation and workload controls at 1B and $N = 32$. All entries except ratio columns are mean complete-step milliseconds. Panel (a) crosses backward overlap with final parameter-update fusion at sequence length 256; its final column reports the latency ratio Neither/Both. Panel (b) compares both ADASIGNSGD encodings with fused AdamW as sequence length increases; each speedup is the corresponding AdamW/ADASIGNSGD latency ratio.

(a) Backward overlap × final parameter-update fusion

Encoding	Neither	Fusion only	Overlap only	Both	Speedup
q-bit	125.727	125.594	107.594	107.551	1.169×
`uint8`	121.200	121.125	103.276	103.242	1.174×

(b) Sequence-length sensitivity

Sequence length	Fused AdamW	ADASIGNSGD `uint8`	Speedup	ADASIGNSGD q-bit	Speedup
256	151.166	103.216	1.465×	107.439	1.407×
512	175.229	139.781	1.254×	141.592	1.238×
1024	261.324	228.492	1.144×	231.253	1.130×

Across the four matched comparisons in panel (a), backward overlap reduces complete-step latency by 14.4–14.8%; final parameter-update fusion changes latency by at most 0.11%. With both controls disabled, the q-bit and `uint8` paths are 1.20× and 1.25× faster than fused AdamW in Table 44. All eight factorial configurations pass same-gradient bitwise validation.

At sequence length 1024 in panel (b), the `uint8` and q-bit paths are 1.144× and 1.130× faster than fused AdamW, corresponding to latency reductions of 12.6% and 11.5%. No additional isolated communication-path measurements are reported for the sequence-length control.

J LANGUAGE-MODEL HYPERPARAMETER SWEEPS

All sweeps follow the model, data, batching, precision, and endpoint-validation protocol in Appendix B; only the displayed method-specific hyperparameters vary. Budgets denote the Chinchilla multiples defined there. "Val. loss" is endpoint validation negative log-likelihood; WU and WD denote warmup updates and weight decay, clip is the gradient-clipping threshold, and comp. indicates Birder's residual compensation. Bold hyperparameters identify the swept coordinate; bold losses mark the reported minimum. Numbered run entries link to the corresponding experiment records.

Double-blind artifacts. Identity-bearing components of the tracking links are temporarily redacted; configurations, metrics, run identifiers, and model-selection decisions are unchanged. Each table row corresponds to a concrete experiment record, with retained run identifiers mapping one-to-one to the tracking backend. The protocol, evaluated hyperparameters, endpoint losses, and selection procedure remain specified here and in Appendix B. Upon acceptance, the camera-ready version will restore complete links and release the corresponding run records.

J.1 AdamW

Table 47: AdamW sweep: 60M, 1× budget.

β_1	β_2	ϵ	η	clip	WU	WD	Val. loss	Run
0.87	0.98	1e-10	0.004	1.0	1000	0.1	3.549	1
0.90	0.98	1e-10	0.004	1.0	1000	0.1	3.564	2
0.95	0.98	1e-10	0.004	1.0	1000	0.1	3.591	3
0.98	0.98	1e-10	0.004	1.0	1000	0.1	3.734	4
0.87	**0.90**	1e-10	0.004	1.0	1000	0.1	3.580	5
0.87	**0.95**	1e-10	0.004	1.0	1000	0.1	3.562	6
0.87	**0.98**	1e-10	0.004	1.0	1000	0.1	3.549	7
0.87	**0.999**	1e-10	0.004	1.0	1000	0.1	3.573	8
0.87	0.98	**1e-20**	0.004	1.0	1000	0.1	3.554	9
0.87	0.98	**1e-10**	0.004	1.0	1000	0.1	3.549	10
0.87	0.98	**1e-8**	0.004	1.0	1000	0.1	3.552	11
0.87	0.98	1e-10	**5e-4**	1.0	1000	0.1	3.705	12
0.87	0.98	1e-10	**0.001**	1.0	1000	0.1	3.584	13
0.87	0.98	1e-10	**0.002**	1.0	1000	0.1	3.545	14
0.87	0.98	1e-10	**0.004**	1.0	1000	0.1	3.549	15
0.87	0.98	1e-10	**0.008**	1.0	1000	0.1	3.555	16
0.87	0.98	1e-10	0.002	**0**	1000	0.1	3.547	17
0.87	0.98	1e-10	0.002	**0.5**	1000	0.1	3.544	18
0.87	0.98	1e-10	0.002	**1.0**	1000	0.1	3.545	19
0.87	0.98	1e-10	0.002	**2.0**	1000	0.1	3.547	20
0.87	0.98	1e-10	0.002	1.0	**1000**	0.1	3.545	21
0.87	0.98	1e-10	0.002	1.0	**2000**	0.1	**3.543**	22
0.87	0.98	1e-10	0.002	1.0	1000	**0**	3.545	23
0.87	0.98	1e-10	0.002	1.0	1000	**0.1**	3.545	24
0.87	0.98	1e-10	0.002	1.0	1000	**0.5**	3.545	25

Table 48: AdamW sweep: 60M, 2× budget.

β_1	β_2	ϵ	η	clip	WU	WD	Val. loss	Run
0.87	0.98	1e-10	0.004	1.0	1000	0.1	3.486	1
0.90	0.98	1e-10	0.004	1.0	1000	0.1	3.501	2
0.95	0.98	1e-10	0.004	1.0	1000	0.1	3.528	3
0.98	0.98	1e-10	0.004	1.0	1000	0.1	3.637	4
0.87	**0.90**	1e-10	0.004	1.0	1000	0.1	3.526	5
0.87	**0.95**	1e-10	0.004	1.0	1000	0.1	3.499	6
0.87	**0.98**	1e-10	0.004	1.0	1000	0.1	3.486	7
0.87	**0.999**	1e-10	0.004	1.0	1000	0.1	3.506	8
0.87	0.98	**1e-8**	0.004	1.0	1000	0.1	3.486	9
0.87	0.98	**1e-10**	0.004	1.0	1000	0.1	3.486	10
0.87	0.98	**1e-20**	0.004	1.0	1000	0.1	3.488	11
0.87	0.98	1e-10	**5e-4**	1.0	1000	0.1	3.580	12
0.87	0.98	1e-10	**0.001**	1.0	1000	0.1	3.496	13
0.87	0.98	1e-10	**0.002**	1.0	1000	0.1	3.475	14
0.87	0.98	1e-10	**0.004**	1.0	1000	0.1	3.486	15
0.87	0.98	1e-10	**0.008**	1.0	1000	0.1	3.487	16
0.87	0.98	1e-10	0.002	**0**	1000	0.1	3.476	17
0.87	0.98	1e-10	0.002	**0.5**	1000	0.1	3.474	18
0.87	0.98	1e-10	0.002	**1.0**	1000	0.1	3.475	19
0.87	0.98	1e-10	0.002	**2.0**	1000	0.1	3.476	20
0.87	0.98	1e-10	0.002	1.0	**1000**	0.1	3.475	21
0.87	0.98	1e-10	0.002	1.0	**2000**	0.1	**3.471**	22
0.87	0.98	1e-10	0.002	1.0	2000	**0**	3.471	23
0.87	0.98	1e-10	0.002	1.0	2000	**0.1**	3.471	24
0.87	0.98	1e-10	0.002	1.0	2000	**0.5**	3.471	25

Table 49: AdamW sweep: 60M, $4\times$ budget.

β_1	β_2	ϵ	η	clip	WU	WD	Val. loss	Run
0.87	0.98	1e-8	0.001	1.0	2000	0.1	3.439	1
0.90	0.98	1e-8	0.001	1.0	2000	0.1	3.442	2
0.95	0.98	1e-8	0.001	1.0	2000	0.1	3.471	3
0.98	0.98	1e-8	0.001	1.0	2000	0.1	3.541	4
0.87	**0.90**	1e-8	0.001	1.0	2000	0.1	3.499	5
0.87	**0.95**	1e-8	0.001	1.0	2000	0.1	3.450	6
0.87	**0.98**	1e-8	0.001	1.0	2000	0.1	3.439	7
0.87	**0.999**	1e-8	0.001	1.0	2000	0.1	3.539	8
0.87	0.98	**1e-8**	0.001	1.0	2000	0.1	3.439	9
0.87	0.98	**1e-10**	0.001	1.0	2000	0.1	3.438	10
0.87	0.98	**1e-20**	0.001	1.0	2000	0.1	3.438	11
0.87	0.98	1e-8	**5e-4**	1.0	2000	0.1	3.494	12
0.87	0.98	1e-8	**0.001**	1.0	2000	0.1	3.439	13
0.87	0.98	1e-8	**0.002**	1.0	2000	0.1	3.422	14
0.87	0.98	1e-8	**0.004**	1.0	2000	0.1	3.423	15
0.87	0.98	1e-8	**0.008**	1.0	2000	0.1	3.419	16
0.87	0.98	1e-8	0.008	**0**	2000	0.1	**3.415**	17
0.87	0.98	1e-8	0.008	**0.5**	2000	0.1	3.418	18
0.87	0.98	1e-8	0.008	**1.0**	2000	0.1	3.419	19
0.87	0.98	1e-8	0.008	**2.0**	2000	0.1	3.422	20
0.87	0.98	1e-8	0.008	0	**1000**	0.1	3.446	21
0.87	0.98	1e-8	0.008	0	**2000**	0.1	3.415	22
0.87	0.98	1e-8	0.008	0	2000	**0**	3.415	23
0.87	0.98	1e-8	0.008	0	2000	**0.1**	3.415	24
0.87	0.98	1e-8	0.008	0	2000	**0.5**	3.418	25

Table 50: AdamW sweep: 60M, $8\times$ budget.

β_1	β_2	ϵ	η	clip	WU	WD	Val. loss	Run
0.87	0.98	1e-8	0.008	0	2000	0.1	3.387	1
0.90	0.98	1e-8	0.008	0	2000	0.1	3.392	2
0.95	0.98	1e-8	0.008	0	2000	0.1	3.422	3
0.98	0.98	1e-8	0.008	0	2000	0.1	3.552	4
0.87	**0.90**	1e-8	0.008	0	2000	0.1	3.471	5
0.87	**0.95**	1e-8	0.008	0	2000	0.1	3.398	6
0.87	**0.98**	1e-8	0.008	0	2000	0.1	3.387	7
0.87	**0.999**	1e-8	0.008	0	2000	0.1	5.740	8
0.87	0.98	**1e-20**	0.008	0	2000	0.1	3.386	9
0.87	0.98	**1e-10**	0.008	0	2000	0.1	3.383	10
0.87	0.98	**1e-8**	0.008	0	2000	0.1	3.387	11
0.87	0.98	1e-10	**5e-4**	0	2000	0.1	3.435	12
0.87	0.98	1e-10	**0.001**	0	2000	0.1	3.396	13
0.87	0.98	1e-10	**0.002**	0	2000	0.1	3.390	14
0.87	0.98	1e-10	**0.004**	0	2000	0.1	3.386	15
0.87	0.98	1e-10	**0.008**	0	2000	0.1	3.383	16
0.87	0.98	1e-10	0.008	**0**	2000	0.1	3.383	17
0.87	0.98	1e-10	0.008	**0.5**	2000	0.1	3.385	18
0.87	0.98	1e-10	0.008	**1.0**	2000	0.1	3.389	19
0.87	0.98	1e-10	0.008	**2.0**	2000	0.1	3.388	20
0.87	0.98	1e-10	0.008	0	**1000**	0.1	3.402	21
0.87	0.98	1e-10	0.008	0	**2000**	0.1	3.383	22
0.87	0.98	1e-10	0.008	0	2000	**0**	3.383	23
0.87	0.98	1e-10	0.008	0	2000	**0.1**	3.383	24
0.87	0.98	1e-10	0.008	0	2000	**0.5**	**3.369**	25

Table 51: AdamW sweep: 130M, 1× budget.

β_1	β_2	ϵ	η	clip	WU	WD	Val. loss	Run
0.87	0.98	1e-8	0.001	1.0	2000	0.1	3.290	1
0.90	0.98	1e-8	0.001	1.0	2000	0.1	3.297	2
0.95	0.98	1e-8	0.001	1.0	2000	0.1	3.322	3
0.98	0.98	1e-8	0.001	1.0	2000	0.1	3.382	4
0.87	**0.90**	1e-8	0.001	1.0	2000	0.1	3.332	5
0.87	**0.95**	1e-8	0.001	1.0	2000	0.1	3.302	6
0.87	**0.98**	1e-8	0.001	1.0	2000	0.1	3.290	7
0.87	**0.999**	1e-8	0.001	1.0	2000	0.1	3.373	8
0.87	0.98	**1e-20**	0.001	1.0	2000	0.1	3.290	9
0.87	0.98	**1e-10**	0.001	1.0	2000	0.1	3.291	10
0.87	0.98	**1e-8**	0.001	1.0	2000	0.1	3.290	11
0.87	0.98	1e-8	**5e-4**	1.0	2000	0.1	3.352	12
0.87	0.98	1e-8	**0.001**	1.0	2000	0.1	3.290	13
0.87	0.98	1e-8	**0.002**	1.0	2000	0.1	**3.284**	14
0.87	0.98	1e-8	**0.004**	1.0	2000	0.1	4.082	15
0.87	0.98	1e-8	**0.008**	1.0	2000	0.1	5.586	16
0.87	0.98	1e-8	0.002	**0**	2000	0.1	3.286	17
0.87	0.98	1e-8	0.002	**0.5**	2000	0.1	3.284	18
0.87	0.98	1e-8	0.002	**1.0**	2000	0.1	3.284	19
0.87	0.98	1e-8	0.002	**2.0**	2000	0.1	3.288	20
0.87	0.98	1e-8	0.002	1.0	**1000**	0.1	3.298	21
0.87	0.98	1e-8	0.002	1.0	**2000**	0.1	3.284	22
0.87	0.98	1e-8	0.002	1.0	2000	**0**	3.284	23
0.87	0.98	1e-8	0.002	1.0	2000	**0.1**	3.284	24
0.87	0.98	1e-8	0.002	1.0	2000	**0.5**	3.284	25

Table 52: AdamW sweep: 130M, 2× budget (scaling-sensitive).

β_1	β_2	ϵ	η	clip	WU	WD	Val. loss	Run
0.87	0.98	1e-8	**0.002**	1.0	2000	0.1	3.225	1
0.87	0.98	1e-8	**0.001**	1.0	2000	0.1	**3.222**	2
0.87	0.98	1e-8	0.001	**0**	2000	0.1	3.224	3
0.87	0.98	1e-8	0.001	1.0	2000	**0.5**	3.222	4

Table 53: AdamW sweep: 130M, 4× budget (scaling-sensitive).

β_1	β_2	ϵ	η	clip	WU	WD	Val. loss	Run
0.87	0.98	1e-8	**0.002**	1.0	2000	0.1	3.181	1
0.87	0.98	1e-8	**0.001**	1.0	2000	0.1	**3.174**	2
0.87	0.98	1e-8	0.001	**0**	2000	0.1	3.177	3
0.87	0.98	1e-8	0.001	1.0	2000	**0.5**	3.174	4

Table 54: AdamW sweep: 130M, 8× budget (scaling-sensitive).

β_1	β_2	ϵ	η	clip	WU	WD	Val. loss	Run
0.87	0.98	1e-8	**0.002**	1.0	2000	0.1	3.154	1
0.87	0.98	1e-8	**0.001**	1.0	2000	0.1	**3.143**	2
0.87	0.98	1e-8	0.001	**0**	2000	0.1	3.144	3
0.87	0.98	1e-8	0.001	1.0	2000	**0.5**	3.143	4

Table 55: AdamW sweep: 350M, 1× budget.

β_1	β_2	ϵ	η	clip	WU	WD	Val. loss	Run
0.87	0.98	1e-8	0.001	1.0	2000	0.1	3.017	1
0.90	0.98	1e-8	0.001	1.0	2000	0.1	3.020	2
0.95	0.98	1e-8	0.001	1.0	2000	0.1	3.039	3
0.98	0.98	1e-8	0.001	1.0	2000	0.1	3.103	4
0.87	**0.90**	1e-8	0.001	1.0	2000	0.1	3.066	5
0.87	**0.95**	1e-8	0.001	1.0	2000	0.1	3.026	6
0.87	**0.98**	1e-8	0.001	1.0	2000	0.1	3.017	7
0.87	**0.999**	1e-8	0.001	1.0	2000	0.1	3.079	8
0.87	0.98	**1e-20**	0.001	1.0	2000	0.1	3.012	9
0.87	0.98	**1e-10**	0.001	1.0	2000	0.1	3.012	10
0.87	0.98	**1e-8**	0.001	1.0	2000	0.1	3.017	11
0.87	0.98	1e-20	**5e-4**	1.0	2000	0.1	3.022	12
0.87	0.98	1e-20	**0.001**	1.0	2000	0.1	3.012	13
0.87	0.98	1e-20	**0.002**	1.0	2000	0.1	3.721	14
0.87	0.98	1e-20	**0.004**	1.0	2000	0.1	4.703	15
0.87	0.98	1e-20	**0.008**	1.0	2000	0.1	5.059	16
0.87	0.98	1e-20	0.001	**0**	2000	0.1	3.009	17
0.87	0.98	1e-20	0.001	**0.5**	2000	0.1	**3.008**	18
0.87	0.98	1e-20	0.001	**1.0**	2000	0.1	3.012	19
0.87	0.98	1e-20	0.001	**2.0**	2000	0.1	3.016	20
0.87	0.98	1e-20	0.001	0.5	**1000**	0.1	3.016	21
0.87	0.98	1e-20	0.001	0.5	**2000**	0.1	3.008	22
0.87	0.98	1e-20	0.001	0.5	2000	**0**	3.008	23
0.87	0.98	1e-20	0.001	0.5	2000	**0.1**	3.008	24
0.87	0.98	1e-20	0.001	0.5	2000	**0.5**	3.008	25

Table 56: AdamW sweep: 350M, 2× budget (scaling-sensitive).

β_1	β_2	ϵ	η	clip	WU	WD	Val. loss	Run
0.87	0.98	1e-20	0.001	0.5	2000	0.1	2.959	1
0.90	0.98	1e-20	0.001	0.5	2000	0.1	2.966	2
0.87	0.98	1e-20	0.001	**0**	2000	0.1	**2.958**	3

Table 57: AdamW sweep: 350M, 4× budget (scaling-sensitive).

β_1	β_2	ϵ	η	clip	WU	WD	Val. loss	Run
0.87	0.98	1e-20	0.001	0.5	2000	0.1	**2.921**	1
0.90	0.98	1e-20	0.001	0.5	2000	0.1	2.931	2
0.87	0.98	1e-20	0.001	**0**	2000	0.1	2.922	3

Table 58: AdamW sweep: 350M, 8× budget (scaling-sensitive).

β_1	β_2	ϵ	η	clip	WU	WD	Val. loss	Run
0.87	0.98	1e-20	0.001	0.5	2000	0.1	2.891	1
0.90	0.98	1e-20	0.001	0.5	2000	0.1	2.902	2
0.87	0.98	1e-20	0.001	**0**	2000	0.1	**2.890**	3

Table 59: AdamW configuration: 1B, 1× budget.

β_1	β_2	ϵ	η	clip	WU	WD	Val. loss	Run
0.87	0.98	1e-10	5e-4	0.5	2000	0.1	**2.838**	1
0.87	0.98	1e-10	6e-4	0.5	2000	0.1	2.885	3
0.87	0.98	1e-10	7e-4	0.5	2000	0.1	2.915	4
0.87	0.98	1e-10	0.001	0.5	2000	0.1	3.592	2

J.2 SIGNUM-MV

Table 60: SIGNUM-MV sweep: 60M, 1× budget.

β	η	clip	WU	WD	Val. loss	Run
0.87	5e-4	0	2000	0.1	6.393	1
0.90	5e-4	0	2000	0.1	5.001	2
0.95	5e-4	0	2000	0.1	5.124	3
0.98	5e-4	0	2000	0.1	**4.287**	4
0.98	**1.25e-4**	0	2000	0.1	4.311	5
0.98	**2.5e-4**	0	2000	0.1	5.763	6
0.98	**5e-4**	0	2000	0.1	4.287	7
0.98	**0.001**	0	2000	0.1	5.537	8
0.98	**0.002**	0	2000	0.1	5.883	9
0.98	**0.004**	0	2000	0.1	5.813	10
0.98	5e-4	**0**	2000	0.1	4.287	11
0.98	5e-4	**0.5**	2000	0.1	10.490	12
0.98	5e-4	**1.0**	2000	0.1	10.490	13
0.98	5e-4	**2.0**	2000	0.1	10.489	14
0.98	5e-4	0	**1000**	0.1	5.237	15
0.98	5e-4	0	**2000**	0.1	4.287	16
0.98	5e-4	0	2000	**0**	4.375	17
0.98	5e-4	0	2000	**0.1**	4.287	18
0.98	5e-4	0	2000	**0.5**	5.505	19

Table 61: SIGNUM-MV sweep: 60M, 2× budget.

β	η	clip	WU	WD	Val. loss	Run
0.87	5e-4	0	2000	0.1	5.499	1
0.90	5e-4	0	2000	0.1	5.853	2
0.95	5e-4	0	2000	0.1	6.569	3
0.98	5e-4	0	2000	0.1	4.737	4
0.98	**1.25e-4**	0	2000	0.1	**4.197**	5
0.98	**2.5e-4**	0	2000	0.1	4.319	6
0.98	**5e-4**	0	2000	0.1	4.737	7
0.98	**0.001**	0	2000	0.1	5.366	8
0.98	**0.002**	0	2000	0.1	5.074	9
0.98	**0.004**	0	2000	0.1	5.274	10
0.98	1.25e-4	**0**	2000	0.1	4.197	11
0.98	1.25e-4	**0.5**	2000	0.1	10.490	12
0.98	1.25e-4	**1.0**	2000	0.1	10.490	13
0.98	1.25e-4	**2.0**	2000	0.1	10.490	14
0.98	1.25e-4	0	**1000**	0.1	4.319	15
0.98	1.25e-4	0	**2000**	0.1	4.197	16
0.98	1.25e-4	0	2000	**0**	4.307	17
0.98	1.25e-4	0	2000	**0.1**	4.197	18
0.98	1.25e-4	0	2000	**0.5**	4.485	19

Table 62: SIGNUM-MV sweep: 60M, 4× budget.

β	η	clip	WU	WD	Val. loss	Run
0.87	1.25e-4	0	2000	0.1	4.576	1
0.90	1.25e-4	0	2000	0.1	4.514	2
0.95	1.25e-4	0	2000	0.1	4.494	3
0.98	1.25e-4	0	2000	0.1	6.057	4
0.95	**1.25e-4**	0	2000	0.1	4.494	5
0.95	**2.5e-4**	0	2000	0.1	4.786	6
0.95	**5e-4**	0	2000	0.1	4.806	7
0.95	**0.001**	0	2000	0.1	4.900	8
0.95	**0.002**	0	2000	0.1	5.178	9
0.95	**0.004**	0	2000	0.1	5.512	10
0.95	1.25e-4	**0**	2000	0.1	4.494	11
0.95	1.25e-4	**0.5**	2000	0.1	10.490	12
0.95	1.25e-4	**1.0**	2000	0.1	10.490	13
0.95	1.25e-4	**2.0**	2000	0.1	10.490	14
0.95	1.25e-4	0	**1000**	0.1	4.404	15
0.95	1.25e-4	0	**2000**	0.1	4.494	16
0.95	1.25e-4	0	1000	**0**	**4.262**	17
0.95	1.25e-4	0	1000	**0.1**	4.404	18
0.95	1.25e-4	0	1000	**0.5**	4.765	19

Table 63: SIGNUM-MV sweep: 60M, 8× budget.

β	η	clip	WU	WD	Val. loss	Run
0.87	1.25e-4	0	2000	0.1	4.594	1
0.90	1.25e-4	0	2000	0.1	4.647	2
0.95	1.25e-4	0	2000	0.1	4.554	3
0.98	1.25e-4	0	2000	0.1	4.412	4
0.98	**1.25e-4**	0	2000	0.1	4.412	5
0.98	**2.5e-4**	0	2000	0.1	4.566	6
0.98	**5e-4**	0	2000	0.1	5.688	7
0.98	**0.001**	0	2000	0.1	4.766	8
0.98	**0.002**	0	2000	0.1	4.447	9
0.98	**0.004**	0	2000	0.1	5.406	10
0.98	1.25e-4	**0**	2000	0.1	4.412	11
0.98	1.25e-4	**0.5**	2000	0.1	10.490	12
0.98	1.25e-4	**1.0**	2000	0.1	10.490	13
0.98	1.25e-4	**2.0**	2000	0.1	10.490	14
0.98	1.25e-4	0	**1000**	0.1	**4.304**	15
0.98	1.25e-4	0	**2000**	0.1	4.412	16
0.98	1.25e-4	0	1000	**0**	5.051	17
0.98	1.25e-4	0	1000	**0.1**	4.304	18
0.98	1.25e-4	0	1000	**0.5**	4.646	19

Table 64: SIGNUM-MV sweep: 130M, 1× budget.

β	η	clip	WU	WD	Val. loss	Run
0.87	5e-4	0	2000	0.1	5.602	1
0.90	5e-4	0	2000	0.1	6.015	2
0.95	5e-4	0	2000	0.1	5.102	3
0.98	5e-4	0	2000	0.1	4.768	4
0.98	**1.25e-4**	0	2000	0.1	4.446	5
0.98	**2.5e-4**	0	2000	0.1	4.747	6
0.98	**5e-4**	0	2000	0.1	4.768	7
0.98	**0.001**	0	2000	0.1	5.419	8
0.98	**0.002**	0	2000	0.1	5.839	9
0.98	**0.004**	0	2000	0.1	5.742	10
0.98	1.25e-4	**0**	2000	0.1	4.446	11
0.98	1.25e-4	**0.5**	2000	0.1	10.505	12
0.98	1.25e-4	**1.0**	2000	0.1	10.505	13
0.98	1.25e-4	**2.0**	2000	0.1	10.505	14
0.98	1.25e-4	0	**1000**	0.1	7.237	15
0.98	1.25e-4	0	**2000**	0.1	4.446	16
0.98	1.25e-4	0	2000	**0**	4.284	17
0.98	1.25e-4	0	2000	**0.1**	4.446	18
0.98	1.25e-4	0	2000	**0.5**	**4.246**	19

Table 65: SIGNUM-MV sweep: 130M, 2× budget (scaling-sensitive).

β	η	clip	WU	WD	Val. loss	Run
0.98	1.25e-4	0	2000	0.5	4.613	1
0.95	1.25e-4	0	2000	0.5	5.532	2
0.98	**5e-4**	0	2000	0.5	4.701	3
0.98	1.25e-4	0	**1000**	0.5	4.464	4
0.98	1.25e-4	0	1000	**0**	**4.255**	5
0.98	1.25e-4	0	1000	**0.1**	4.383	6

Table 66: SIGNUM-MV sweep: 130M, 4× budget (scaling-sensitive).

β	η	clip	WU	WD	Val. loss	Run
0.98	1.25e-4	0	2000	0.5	4.775	1
0.95	1.25e-4	0	2000	0.5	4.896	2
0.98	**5e-4**	0	2000	0.5	4.582	3
0.98	5e-4	0	**1000**	0.5	**4.579**	4
0.98	5e-4	0	1000	**0**	5.346	5
0.98	5e-4	0	1000	**0.1**	5.350	6

Table 67: SIGNUM-MV sweep: 130M, 8× budget (scaling-sensitive).

β	η	clip	WU	WD	Val. loss	Run
0.98	1.25e-4	0	2000	0.5	4.916	1
0.95	1.25e-4	0	2000	0.5	5.089	2
0.98	**5e-4**	0	2000	0.5	**4.591**	3
0.98	5e-4	0	**1000**	0.5	4.668	4
0.98	5e-4	0	2000	**0**	4.796	5
0.98	5e-4	0	2000	**0.1**	7.073	6

Table 68: SIGNUM-MV sweep: 350M, 1× budget.

β	η	clip	WU	WD	Val. loss	Run
0.87	0.000125	0	2000	0.5	5.355	1
0.90	0.000125	0	2000	0.5	5.303	2
0.95	0.000125	0	2000	0.5	5.079	3
0.98	0.000125	0	2000	0.5	4.755	4
0.98	**0.000125**	0	2000	0.5	4.755	5
0.98	**0.00025**	0	2000	0.5	**4.673**	6
0.98	**5e-4**	0	2000	0.5	6.745	7
0.98	**0.001**	0	2000	0.5	5.645	8
0.98	**0.002**	0	2000	0.5	5.094	9
0.98	**0.004**	0	2000	0.5	5.570	10
0.98	0.00025	**0**	2000	0.5	4.673	11
0.98	0.00025	**0.5**	2000	0.5	nonfinite	12
0.98	0.00025	**1.0**	2000	0.5	nonfinite	13
0.98	0.00025	**2.0**	2000	0.5	nonfinite	14
0.98	0.00025	0	**1000**	0.5	4.804	15
0.98	0.00025	0	**2000**	0.5	4.673	16
0.98	0.00025	0	2000	**0**	5.822	17
0.98	0.00025	0	2000	**0.1**	4.769	18
0.98	0.00025	0	2000	**0.5**	4.673	19

Table 69: SIGNUM-MV sweep: 350M, 2× budget (scaling-sensitive).

β	η	clip	WU	WD	Val. loss	Run
0.98	**0.00025**	0	2000	0.5	5.138	1
0.98	**0.000125**	0	2000	0.5	5.158	2
0.98	0.00025	0	**1000**	0.5	**4.940**	3
0.98	0.00025	0	1000	**0.1**	5.548	4

Table 70: SIGNUM-MV sweep: 350M, 4× budget (scaling-sensitive).

β	η	clip	WU	WD	Val. loss	Run
0.98	**0.00025**	0	2000	0.5	4.845	1
0.98	**0.000125**	0	2000	0.5	4.697	2
0.98	0.000125	0	**1000**	0.5	4.608	3
0.98	0.000125	0	1000	**0.1**	**4.589**	4

Table 71: SIGNUM-MV sweep: 350M, 8× budget (scaling-sensitive).

β	η	clip	WU	WD	Val. loss	Run
0.98	**0.00025**	0	2000	0.5	4.857	1
0.98	**0.000125**	0	2000	0.5	4.606	2
0.98	0.000125	0	**1000**	0.5	**4.575**	3

Table 72: SIGNUM-MV sweep: 1B, 1× budget.

β	η	clip	WU	WD	Val. loss	Run
0.98	1.25e-4	0	2000	0.5	**4.853**	1

J.3 SIGNUM-MV+EF

Table 73: SIGNUM-MV+EF sweep: 60M, 1× budget.

β	η	clip	WU	WD	Val. loss	Run
0.87	4.096	0	2000	0	**4.143**	1
0.90	4.096	0	2000	0	4.288	2
0.95	4.096	0	2000	0	10.374	3
0.98	4.096	0	2000	0	5.654	4
0.87	**0.128**	0	2000	0	6.424	5
0.87	**0.256**	0	2000	0	6.052	6
0.87	**0.512**	0	2000	0	5.253	7
0.87	**1.024**	0	2000	0	5.049	8
0.87	**2.048**	0	2000	0	4.467	9
0.87	**4.096**	0	2000	0	4.143	10
0.87	**8.192**	0	2000	0	non-finite	11
0.87	4.096	**0**	2000	0	4.143	12
0.87	4.096	**0.5**	2000	0	4.688	13
0.87	4.096	**1.0**	2000	0	4.185	14
0.87	4.096	**2.0**	2000	0	4.143	15
0.87	4.096	0	**1000**	0	10.374	16
0.87	4.096	0	**2000**	0	4.143	17
0.87	4.096	0	2000	**0**	4.143	18
0.87	4.096	0	2000	**0.1**	13.737	19
0.87	4.096	0	2000	**0.5**	non-finite	20

Table 74: SIGNUM-MV+EF sweep: 60M, 2× budget.

β	η	clip	WU	WD	Val. loss	Run
0.87	4.096	0	2000	0	**3.866**	1
0.90	4.096	0	2000	0	3.925	2
0.95	4.096	0	2000	0	non-finite	3
0.98	4.096	0	2000	0	4.462	4
0.87	**0.128**	0	2000	0	6.170	5
0.87	**0.256**	0	2000	0	5.730	6
0.87	**0.512**	0	2000	0	4.848	7
0.87	**1.024**	0	2000	0	4.623	8
0.87	**2.048**	0	2000	0	4.122	9
0.87	**4.096**	0	2000	0	3.866	10
0.87	**8.192**	0	2000	0	non-finite	11
0.87	4.096	**0**	2000	0	3.866	12
0.87	4.096	**0.5**	2000	0	4.360	13
0.87	4.096	**1.0**	2000	0	3.880	14
0.87	4.096	**2.0**	2000	0	3.866	15
0.87	4.096	0	**1000**	0	10.374	16
0.87	4.096	0	**2000**	0	3.866	17
0.87	4.096	0	2000	**0**	3.866	18
0.87	4.096	0	2000	**0.1**	15.361	19
0.87	4.096	0	2000	**0.5**	non-finite	20

Table 75: SIGNUM-MV+EF sweep: 60M, 4× budget.

β	η	clip	WU	WD	Val. loss	Run
0.87	4.096	0	2000	0	**3.683**	1
0.90	4.096	0	2000	0	3.703	2
0.95	4.096	0	2000	0	non-finite	3
0.98	4.096	0	2000	0	3.818	4
0.87	**0.128**	0	2000	0	5.888	5
0.87	**0.256**	0	2000	0	4.905	6
0.87	**0.512**	0	2000	0	4.515	7
0.87	**1.024**	0	2000	0	4.331	8
0.87	**2.048**	0	2000	0	3.898	9
0.87	**4.096**	0	2000	0	3.683	10
0.87	**8.192**	0	2000	0	10.374	11
0.87	4.096	**0**	2000	0	3.683	12
0.87	4.096	**0.5**	2000	0	4.171	13
0.87	4.096	**1.0**	2000	0	3.687	14
0.87	4.096	**2.0**	2000	0	3.683	15
0.87	4.096	0	**1000**	0	10.374	16
0.87	4.096	0	**2000**	0	3.683	17
0.87	4.096	0	2000	**0**	3.683	18
0.87	4.096	0	2000	**0.1**	10.105	19
0.87	4.096	0	2000	**0.5**	non-finite	20

Table 76: SIGNUM-MV+EF sweep: 60M, 8× budget.

β	η	clip	WU	WD	Val. loss	Run
0.87	4.096	0	2000	0	3.551	1
0.90	4.096	0	2000	0	3.555	2
0.95	4.096	0	2000	0	3.569	3
0.98	4.096	0	2000	0	10.374	4
0.87	**0.128**	0	2000	0	4.863	5
0.87	**0.256**	0	2000	0	4.485	6
0.87	**0.512**	0	2000	0	4.271	7
0.87	**1.024**	0	2000	0	4.089	8
0.87	**2.048**	0	2000	0	3.744	9
0.87	**4.096**	0	2000	0	3.551	10
0.87	**8.192**	0	2000	0	non-finite	11
0.87	4.096	**0**	2000	0	3.551	12
0.87	4.096	**0.5**	2000	0	4.056	13
0.87	4.096	**1.0**	2000	0	**3.550**	14
0.87	4.096	**2.0**	2000	0	3.551	15
0.87	4.096	0	**1000**	0	10.374	16
0.87	4.096	0	**2000**	0	3.551	17
0.87	4.096	0	2000	**0**	3.551	18
0.87	4.096	0	2000	**0.1**	11.596	19
0.87	4.096	0	2000	**0.5**	non-finite	20

Table 77: SIGNUM-MV+EF sweep: 130M, 1× budget.

β	η	clip	WU	WD	Val. loss	Run
0.87	4.096	0	2000	0	3.807	1
0.90	4.096	0	2000	0	3.800	2
0.95	4.096	0	2000	0	3.857	3
0.98	4.096	0	2000	0	3.897	4
0.87	**0.128**	0	2000	0	5.771	5
0.87	**0.256**	0	2000	0	5.551	6
0.87	**0.512**	0	2000	0	4.572	7
0.87	**1.024**	0	2000	0	4.306	8
0.87	**2.048**	0	2000	0	4.052	9
0.87	**4.096**	0	2000	0	3.807	10
0.87	4.096	**0**	2000	0	3.807	11
0.87	4.096	**0.5**	2000	0	4.880	12
0.87	4.096	**1.0**	2000	0	3.879	13
0.87	4.096	**2.0**	2000	0	3.807	14
0.87	4.096	0	**1000**	0	**3.795**	15
0.87	4.096	0	**2000**	0	3.807	16
0.87	4.096	0	1000	**0**	3.795	17
0.87	4.096	0	1000	**0.1**	9.319	18
0.87	4.096	0	1000	**0.5**	nonfinite	19

Table 78: SIGNUM-MV+EF sweep: 130M, 2× budget (scaling-sensitive).

β	η	clip	WU	WD	Val. loss	Run
0.87	**4.096**	0	2000	0	**3.630**	1

Table 79: SIGNUM-MV+EF sweep: 130M, 4× budget (scaling-sensitive).

β	η	clip	WU	WD	Val. loss	Run
0.87	**4.096**	0	2000	0	**3.488**	1

Table 80: SIGNUM-MV+EF sweep: 130M, 8× budget (scaling-sensitive).

β	η	clip	WU	WD	Val. loss	Run
0.87	**4.096**	0	2000	0	**3.375**	1

Table 81: SIGNUM-MV+EF sweep: 350M, 1× budget.

β	η	clip	WU	WD	Val. loss	Run
0.87	4.096	0	2000	0	3.397	1
0.90	4.096	0	2000	0	3.396	2
0.95	4.096	0	2000	0	3.367	3
0.98	4.096	0	2000	0	3.383	4
0.95	**0.128**	0	2000	0	6.098	5
0.95	**0.256**	0	2000	0	5.446	6
0.95	**0.512**	0	2000	0	4.802	7
0.95	**1.024**	0	2000	0	3.704	8
0.95	**2.048**	0	2000	0	3.496	9
0.95	**4.096**	0	2000	0	3.367	10
0.95	4.096	**0**	2000	0	3.367	11
0.95	4.096	**0.5**	2000	0	5.823	12
0.95	4.096	**1.0**	2000	0	4.065	13
0.95	4.096	**2.0**	2000	0	3.367	14
0.95	4.096	0	**2000**	0	3.367	15
0.95	4.096	0	**1000**	0	**3.356**	16
0.95	4.096	0	1000	**0**	3.356	17
0.95	4.096	0	1000	**0.1**	14.033	18
0.95	4.096	0	1000	**0.5**	nonfinite	19

Table 82: SIGNUM-MV+EF sweep: 350M, 2× budget (scaling-sensitive).

β	η	clip	WU	WD	Val. loss	Run
0.95	**4.096**	0	2000	0	**3.217**	1

Table 83: SIGNUM-MV+EF sweep: 350M, 4× budget (scaling-sensitive).

β	η	clip	WU	WD	Val. loss	Run
0.95	**4.096**	0	2000	0	**3.106**	1

Table 84: SIGNUM-MV+EF sweep: 350M, 8× budget (scaling-sensitive).

β	η	clip	WU	WD	Val. loss	Run
0.95	**4.096**	0	2000	0	**3.030**	1

Table 85: SIGNUM-MV+EF sweep: 1B, 1× budget.

β	η	clip	WU	WD	Val. loss	Run
0.95	4.096	0	2000	0	**3.384**	1

J.4 BIRDER

Table 86: Birder sweep: 60M, 1× budget.

β	η	clip	WU	WD	comp.	Val. loss	Run
0.87	0.001	0	1000	0	off	3.830	1
0.90	0.001	0	1000	0	off	3.792	2
0.95	0.001	0	1000	0	off	3.731	3
0.98	0.001	0	1000	0	off	3.836	4
0.95	**5e-4**	0	1000	0	off	3.852	5
0.95	**0.001**	0	1000	0	off	3.731	6
0.95	**0.002**	0	1000	0	off	3.694	7
0.95	**0.004**	0	1000	0	off	3.701	8
0.95	**0.008**	0	1000	0	off	3.725	9
0.95	0.002	**0**	1000	0	off	3.694	10
0.95	0.002	**0.5**	1000	0	off	10.489	11
0.95	0.002	**1.0**	1000	0	off	10.488	12
0.95	0.002	**2.0**	1000	0	off	10.481	13
0.95	0.002	0	**1000**	0	off	3.694	14
0.95	0.002	0	**2000**	0	off	3.676	15
0.95	0.002	0	2000	**0**	off	3.676	16
0.95	0.002	0	2000	**0.1**	off	3.667	17
0.95	0.002	0	2000	**0.5**	off	3.681	18
0.95	0.002	0	2000	0.1	**off**	3.667	19
0.95	0.002	0	2000	0.1	**on**	**3.647**	20

Table 87: Birder sweep: 60M, 2× budget.

β	η	clip	WU	WD	comp.	Val. loss	Run
0.87	0.001	0	1000	0	off	3.735	1
0.90	0.001	0	1000	0	off	3.696	2
0.95	0.001	0	1000	0	off	3.627	3
0.98	0.001	0	1000	0	off	3.674	4
0.95	**5e-4**	0	1000	0	off	3.718	5
0.95	**0.001**	0	1000	0	off	3.627	6
0.95	**0.002**	0	1000	0	off	3.607	7
0.95	**0.004**	0	1000	0	off	3.617	8
0.95	**0.008**	0	1000	0	off	3.637	9
0.95	0.002	**0**	1000	0	off	3.607	10
0.95	0.002	**0.5**	1000	0	off	10.489	11
0.95	0.002	**1.0**	1000	0	off	10.488	12
0.95	0.002	**2.0**	1000	0	off	10.481	13
0.95	0.002	0	**1000**	0	off	3.607	14
0.95	0.002	0	**2000**	0	off	3.592	15
0.95	0.002	0	2000	**0**	off	3.592	16
0.95	0.002	0	2000	**0.1**	off	3.580	17
0.95	0.002	0	2000	**0.5**	off	3.602	18
0.95	0.002	0	2000	0.1	**on**	**3.553**	19

Table 88: Birder sweep: 60M, 4× budget.

β	η	clip	WU	WD	comp.	Val. loss	Run
0.87	0.001	0	1000	0	off	3.664	1
0.90	0.001	0	1000	0	off	3.625	2
0.95	0.001	0	1000	0	off	3.552	3
0.98	0.001	0	1000	0	off	3.569	4
0.95	**5e-4**	0	1000	0	off	3.619	5
0.95	**0.001**	0	1000	0	off	3.552	6
0.95	**0.002**	0	1000	0	off	3.545	7
0.95	**0.004**	0	1000	0	off	3.552	8
0.95	**0.008**	0	1000	0	off	3.572	9
0.95	0.002	**0**	1000	0	off	3.545	10
0.95	0.002	**0.5**	1000	0	off	10.489	11
0.95	0.002	**1.0**	1000	0	off	10.488	12
0.95	0.002	**2.0**	1000	0	off	10.481	13
0.95	0.002	0	**1000**	0	off	3.545	14
0.95	0.002	0	**2000**	0	off	3.532	15
0.95	0.002	0	2000	**0**	off	3.532	16
0.95	0.002	0	2000	**0.1**	off	3.523	17
0.95	0.002	0	2000	**0.5**	off	3.549	18
0.95	0.002	0	2000	0.1	**off**	3.523	19
0.95	0.002	0	2000	0.1	**on**	**3.489**	20

Table 89: Birder sweep: 60M, 8× budget.

β	η	clip	WU	WD	comp.	Val. loss	Run
0.87	0.001	0	1000	0	off	3.615	1
0.90	0.001	0	1000	0	off	3.573	2
0.95	0.001	0	1000	0	off	3.500	3
0.98	0.001	0	1000	0	off	3.496	4
0.98	**5e-4**	0	1000	0	off	3.514	5
0.98	**0.001**	0	1000	0	off	3.496	6
0.98	**0.002**	0	1000	0	off	3.520	7
0.98	**0.004**	0	1000	0	off	3.530	8
0.98	**0.008**	0	1000	0	off	4.756	9
0.98	0.001	**0**	1000	0	off	3.496	10
0.98	0.001	**0.5**	1000	0	off	10.490	11
0.98	0.001	**1.0**	1000	0	off	10.489	12
0.98	0.001	**2.0**	1000	0	off	10.488	13
0.98	0.001	0	**1000**	0	off	3.496	14
0.98	0.001	0	**2000**	0	off	3.458	15
0.98	0.001	0	2000	**0**	off	3.458	16
0.98	0.001	0	2000	**0.1**	off	3.446	17
0.98	0.001	0	2000	**0.5**	off	3.482	18
0.98	0.001	0	2000	0.1	**off**	3.446	19
0.98	0.001	0	2000	0.1	**on**	**3.418**	20

Table 90: Birder sweep: 130M, 1× budget.

β	η	clip	WU	WD	comp.	Val. loss	Run
0.87	0.001	0	1000	0	off	3.513	1
0.90	0.001	0	1000	0	off	3.477	2
0.95	0.001	0	1000	0	off	3.416	3
0.98	0.001	0	1000	0	off	3.505	4
0.95	**5e-4**	0	1000	0	off	3.488	5
0.95	**0.001**	0	1000	0	off	3.416	6
0.95	**0.002**	0	1000	0	off	3.420	7
0.95	**0.004**	0	1000	0	off	3.442	8
0.95	**0.008**	0	1000	0	off	3.814	9
0.95	0.001	**0**	1000	0	off	3.416	10
0.95	0.001	**0.5**	1000	0	off	10.505	11
0.95	0.001	**1.0**	1000	0	off	10.504	12
0.95	0.001	**2.0**	1000	0	off	10.501	13
0.95	0.001	0	**1000**	0	off	3.416	14
0.95	0.001	0	**2000**	0	off	3.413	15
0.95	0.001	0	1000	**0**	off	3.416	16
0.95	0.001	0	1000	**0.1**	off	3.402	17
0.95	0.001	0	1000	**0.5**	off	3.408	18
0.95	0.001	0	1000	0.1	**off**	3.402	19
0.95	0.001	0	1000	0.1	**on**	**3.378**	20

Table 91: Birder sweep: 130M, 2× budget (scaling-sensitive).

β	η	clip	WU	WD	comp.	Val. loss	Run
0.95	0.001	0	1000	0.1	off	3.318	1
0.98	0.001	0	1000	0.1	off	3.338	2
0.95	**0.002**	0	1000	0.1	off	3.320	3
0.95	0.001	0	**2000**	0.1	off	3.316	4
0.95	0.001	0	1000	0.1	**on**	**3.285**	5

Table 92: Birder sweep: 130M, 4× budget (scaling-sensitive).

β	η	clip	WU	WD	comp.	Val. loss	Run
0.95	0.001	0	1000	0.1	off	3.261	1
0.98	0.001	0	1000	0.1	off	3.256	2
0.98	**0.002**	0	1000	0.1	off	3.265	3
0.98	0.001	0	**2000**	0.1	off	3.238	4
0.98	0.001	0	2000	0.1	**on**	**3.206**	5

Table 93: Birder sweep: 130M, 8× budget (scaling-sensitive).

β	η	clip	WU	WD	comp.	Val. loss	Run
0.95	0.001	0	1000	0.1	off	3.221	1
0.98	0.001	0	1000	0.1	off	3.201	2
0.98	**0.002**	0	1000	0.1	off	3.220	3
0.98	0.001	0	**2000**	0.1	off	3.191	4
0.98	0.001	0	2000	0.1	**on**	**3.156**	5

Table 94: Birder sweep: 350M, 1× budget.

β	η	clip	WU	WD	comp.	Val. loss	Run
0.87	0.001	0	1000	0	off	3.195	1
0.90	0.001	0	1000	0	off	3.162	2
0.95	0.001	0	1000	0	off	3.110	3
0.98	0.001	0	1000	0	off	3.161	4
0.95	**5e-4**	0	1000	0	off	3.127	5
0.95	**0.001**	0	1000	0	off	3.110	6
0.95	**0.002**	0	1000	0	off	3.146	7
0.95	**0.004**	0	1000	0	off	3.186	8
0.95	**0.008**	0	1000	0	off	5.117	9
0.95	0.001	**0**	1000	0	off	3.110	10
0.95	0.001	**0.5**	1000	0	off	10.586	11
0.95	0.001	**1.0**	1000	0	off	10.585	12
0.95	0.001	**2.0**	1000	0	off	10.584	13
0.95	0.001	0	**1000**	0	off	3.110	14
0.95	0.001	0	**2000**	0	off	3.100	15
0.95	0.001	0	2000	**0**	off	3.100	16
0.95	0.001	0	2000	**0.1**	off	3.084	17
0.95	0.001	0	2000	**0.5**	off	3.096	18
0.95	0.001	0	2000	0.1	**off**	3.084	19
0.95	0.001	0	2000	0.1	**on**	**3.036**	20

Table 95: Birder sweep: 350M, 2× budget (scaling-sensitive).

β	η	clip	WU	WD	comp.	Val. loss	Run
0.95	0.001	0	2000	0.1	off	3.030	1
0.98	0.001	0	2000	0.1	off	3.021	2
0.98	**0.002**	0	2000	0.1	off	3.054	3
0.98	0.001	0	**1000**	0.1	off	3.043	4
0.98	0.001	0	2000	0.1	**on**	**2.977**	5

Table 96: Birder sweep: 350M, 4× budget (scaling-sensitive).

β	η	clip	WU	WD	comp.	Val. loss	Run
0.95	0.001	0	2000	0.1	off	2.991	1
0.98	0.001	0	2000	0.1	off	2.976	2
0.98	**0.002**	0	2000	0.1	off	3.002	3
0.98	0.001	0	**1000**	0.1	off	2.993	4
0.98	0.001	0	2000	0.1	**on**	**2.934**	5

Table 97: Birder sweep: 350M, 8× budget (scaling-sensitive).

β	η	clip	WU	WD	comp.	Val. loss	Run
0.95	0.001	0	2000	0.1	off	2.956	1
0.95	**0.002**	0	2000	0.1	off	2.963	2
0.95	0.001	0	2000	0.1	**on**	**2.921**	3

Table 98: Birder sweep: 1B, 1× budget (scaling-sensitive).

β	η	clip	WU	WD	comp.	Val. loss	Run
0.98	0.001	0	2000	0.1	**off**	**2.860**	1
0.98	0.001	0	2000	0.1	**on**	2.891	2

J.5 D-LION-MAVO

Table 99: D-Lion-MaVo sweep: 60M, 1× budget.

β_1	β_2	η	clip	WU	WD	Val. loss	Run
0.90	0.98	5e-4	0	1000	1.0	4.820	1
0.87	0.98	5e-4	0	1000	1.0	4.552	2
0.95	0.98	5e-4	0	1000	1.0	4.859	3
0.98	0.98	5e-4	0	1000	1.0	4.348	4
0.98	**0.98**	5e-4	0	1000	1.0	4.348	5
0.98	**0.999**	5e-4	0	1000	1.0	6.017	6
0.98	**0.90**	5e-4	0	1000	1.0	6.574	7
0.98	**0.95**	5e-4	0	1000	1.0	4.587	8
0.98	0.98	**5e-4**	0	1000	1.0	4.348	9
0.98	0.98	**1.25e-4**	0	1000	1.0	4.436	10
0.98	0.98	**2.5e-4**	0	1000	1.0	4.319	11
0.98	0.98	**0.001**	0	1000	1.0	6.082	12
0.98	0.98	**0.002**	0	1000	1.0	5.520	13
0.98	0.98	2.5e-4	0	**1000**	1.0	4.319	14
0.98	0.98	2.5e-4	0	**2000**	1.0	4.912	15
0.98	0.98	2.5e-4	0	1000	**1.0**	4.319	16
0.98	0.98	2.5e-4	0	1000	**0**	**4.279**	17
0.98	0.98	2.5e-4	0	1000	**0.1**	4.291	18
0.98	0.98	2.5e-4	0	1000	**0.5**	4.337	19

Table 100: D-Lion-MaVo sweep: 60M, 2× budget.

β_1	β_2	η	clip	WU	WD	Val. loss	Run
0.98	0.98	2.5e-4	0	1000	0	4.383	1
0.90	0.98	2.5e-4	0	1000	0	6.410	2
0.87	0.98	2.5e-4	0	1000	0	5.811	3
0.95	0.98	2.5e-4	0	1000	0	4.513	4
0.98	**0.98**	2.5e-4	0	1000	0	4.383	5
0.98	**0.999**	2.5e-4	0	1000	0	7.093	6
0.98	**0.90**	2.5e-4	0	1000	0	4.537	7
0.98	**0.95**	2.5e-4	0	1000	0	4.459	8
0.98	0.98	**2.5e-4**	0	1000	0	4.383	9
0.98	0.98	**1.25e-4**	0	1000	0	4.314	10
0.98	0.98	**5e-4**	0	1000	0	4.767	11
0.98	0.98	**0.001**	0	1000	0	5.176	12
0.98	0.98	**0.002**	0	1000	0	5.324	13
0.98	0.98	1.25e-4	0	**1000**	0	4.314	14
0.98	0.98	1.25e-4	0	**2000**	0	**4.141**	15
0.98	0.98	1.25e-4	0	2000	**0**	4.141	16
0.98	0.98	1.25e-4	0	2000	**1.0**	4.699	17
0.98	0.98	1.25e-4	0	2000	**0.1**	4.225	18
0.98	0.98	1.25e-4	0	2000	**0.5**	4.406	19

Table 101: D-Lion-MaVo sweep: 60M, 4× budget.

β_1	β_2	η	clip	WU	WD	Val. loss	Run
0.98	0.98	2.5e-4	0	1000	0	4.316	1
0.90	0.98	2.5e-4	0	1000	0	6.367	2
0.87	0.98	2.5e-4	0	1000	0	6.111	3
0.95	0.98	2.5e-4	0	1000	0	6.825	4
0.98	**0.98**	2.5e-4	0	1000	0	4.316	5
0.98	**0.999**	2.5e-4	0	1000	0	6.935	6
0.98	**0.90**	2.5e-4	0	1000	0	4.715	7
0.98	**0.95**	2.5e-4	0	1000	0	4.595	8
0.98	0.98	**2.5e-4**	0	1000	0	4.316	9
0.98	0.98	**1.25e-4**	0	1000	0	4.207	10
0.98	0.98	**5e-4**	0	1000	0	4.860	11
0.98	0.98	**0.001**	0	1000	0	5.296	12
0.98	0.98	**0.002**	0	1000	0	5.166	13
0.98	0.98	1.25e-4	0	**1000**	0	4.207	14
0.98	0.98	1.25e-4	0	**2000**	0	**4.079**	15
0.98	0.98	1.25e-4	0	2000	**0**	4.079	16
0.98	0.98	1.25e-4	0	2000	**1.0**	4.675	17
0.98	0.98	1.25e-4	0	2000	**0.1**	4.178	18
0.98	0.98	1.25e-4	0	2000	**0.5**	4.737	19

Table 102: D-Lion-MaVo sweep: 60M, 8× budget.

β_1	β_2	η	clip	WU	WD	Val. loss	Run
0.98	0.98	2.5e-4	0	1000	0	**4.225**	1
0.90	0.98	2.5e-4	0	1000	0	6.338	2
0.87	0.98	2.5e-4	0	1000	0	5.691	3
0.95	0.98	2.5e-4	0	1000	0	6.101	4
0.98	**0.98**	2.5e-4	0	1000	0	4.225	5
0.98	**0.999**	2.5e-4	0	1000	0	6.788	6
0.98	**0.90**	2.5e-4	0	1000	0	5.151	7
0.98	**0.95**	2.5e-4	0	1000	0	4.593	8
0.98	0.98	**2.5e-4**	0	1000	0	4.225	9
0.98	0.98	**1.25e-4**	0	1000	0	4.334	10
0.98	0.98	**5e-4**	0	1000	0	4.854	11
0.98	0.98	**0.001**	0	1000	0	5.256	12
0.98	0.98	**0.002**	0	1000	0	5.646	13
0.98	0.98	2.5e-4	0	**1000**	0	4.225	14
0.98	0.98	2.5e-4	0	**2000**	0	4.801	15
0.98	0.98	2.5e-4	0	1000	**0**	4.225	16
0.98	0.98	2.5e-4	0	1000	**1.0**	4.687	17
0.98	0.98	2.5e-4	0	1000	**0.1**	4.626	18
0.98	0.98	2.5e-4	0	1000	**0.5**	5.684	19

Table 103: D-Lion-MaVo sweep: 130M, 1× budget.

β_1	β_2	η	clip	WU	WD	Val. loss	Run
0.98	0.98	2.5e-4	0	1000	0	4.662	1
0.90	0.98	2.5e-4	0	1000	0	5.178	2
0.87	0.98	2.5e-4	0	1000	0	5.710	3
0.95	0.98	2.5e-4	0	1000	0	**4.597**	4
0.95	**0.98**	2.5e-4	0	1000	0	4.597	5
0.95	**0.999**	2.5e-4	0	1000	0	6.915	6
0.95	**0.90**	2.5e-4	0	1000	0	5.371	7
0.95	**0.95**	2.5e-4	0	1000	0	5.227	8
0.95	0.98	**2.5e-4**	0	1000	0	4.597	9
0.95	0.98	**1.25e-4**	0	1000	0	5.984	10
0.95	0.98	**5e-4**	0	1000	0	5.250	11
0.95	0.98	**0.001**	0	1000	0	5.485	12
0.95	0.98	**0.002**	0	1000	0	5.773	13
0.95	0.98	2.5e-4	0	**1000**	0	4.597	14
0.95	0.98	2.5e-4	0	**2000**	0	5.306	15
0.95	0.98	2.5e-4	0	1000	**0**	4.597	16
0.95	0.98	2.5e-4	0	1000	**1.0**	6.294	17
0.95	0.98	2.5e-4	0	1000	**0.1**	4.898	18
0.95	0.98	2.5e-4	0	1000	**0.5**	5.407	19

Table 104: D-Lion-MaVo sweep: 130M, 2× budget (scaling-sensitive).

β_1	β_2	η	clip	WU	WD	Val. loss	Run
0.95	0.98	2.5e-4	0	1000	0	**4.853**	1
0.95	0.98	**1.25e-4**	0	1000	0	5.094	2
0.95	0.98	2.5e-4	0	**2000**	0	5.321	3

Table 105: D-Lion-MaVo sweep: 130M, 4× budget (scaling-sensitive).

β_1	β_2	η	clip	WU	WD	Val. loss	Run
0.95	0.98	2.5e-4	0	1000	0	**5.081**	1
0.95	0.98	**1.25e-4**	0	1000	0	5.196	2
0.95	0.98	2.5e-4	0	**2000**	0	5.441	3

Table 106: D-Lion-MaVo sweep: 130M, 8× budget (scaling-sensitive).

β_1	β_2	η	clip	WU	WD	Val. loss	Run
0.95	0.98	2.5e-4	0	1000	0	**5.085**	1

Table 107: D-Lion-MaVo sweep: 350M, 1× budget.

β_1	β_2	η	clip	WU	WD	Val. loss	Run
0.98	0.98	2.5e-4	0	1000	0	5.263153	1
0.90	0.98	2.5e-4	0	1000	0	5.568975	2
0.87	0.98	2.5e-4	0	1000	0	5.871058	3
0.95	0.98	2.5e-4	0	1000	0	5.462575	4
0.98	**0.98**	2.5e-4	0	1000	0	5.263153	5
0.98	**0.999**	2.5e-4	0	1000	0	6.799395	6
0.98	**0.90**	2.5e-4	0	1000	0	5.721149	7
0.98	**0.95**	2.5e-4	0	1000	0	5.923743	8
0.98	0.98	**2.5e-4**	0	1000	0	5.263153	9
0.98	0.98	**1.25e-4**	0	1000	0	5.010968	10
0.98	0.98	**5e-4**	0	1000	0	5.614230	11
0.98	0.98	**0.001**	0	1000	0	6.003282	12
0.98	0.98	**0.002**	0	1000	0	6.150418	13
0.98	0.98	1.25e-4	0	**1000**	0	5.010968	14
0.98	0.98	1.25e-4	0	**2000**	0	5.425168	15
0.98	0.98	1.25e-4	0	1000	**0**	5.010968	16
0.98	0.98	1.25e-4	0	1000	**1.0**	**4.825097**	17
0.98	0.98	1.25e-4	0	1000	**0.1**	5.185792	18
0.98	0.98	1.25e-4	0	1000	**0.5**	4.831936	19

Table 108: D-Lion-MaVo sweep: 350M, 2× budget.

β_1	β_2	η	clip	WU	WD	Val. loss	Run
0.98	0.98	1.25e-4	0	1000	1.0	**4.747**	1
0.98	0.98	6.25e-5	0	1000	1.0	4.769	2
0.98	0.98	1.25e-4	0	2000	1.0	5.164	3
0.98	0.98	1.25e-4	0	1000	0.5	4.973	4

Table 109: D-Lion-MaVo sweep: 350M, 4× budget (scaling-sensitive).

β_1	β_2	η	clip	WU	WD	Val. loss	Run
0.98	0.98	1.25e-4	0	1000	1.0	4.849	1
0.98	0.98	**6.25e-5**	0	1000	1.0	**4.764**	2
0.98	0.98	6.25e-5	0	**2000**	1.0	4.908	3
0.98	0.98	6.25e-5	0	1000	**0.5**	4.787	4

Table 110: D-Lion-MaVo sweep: 350M, 8× budget (scaling-sensitive).

β_1	β_2	η	clip	WU	WD	Val. loss	Run
0.98	0.98	1.25e-4	0	1000	1.0	5.193	1
0.98	0.98	1.25e-4	0	1000	**0.5**	**4.711**	2

Table 111: D-Lion-MaVo sweep: 1B, 1× budget.

β_1	β_2	η	clip	WU	WD	Val. loss	Run
0.98	0.98	6.25e-5	0	1000	1.0	**4.765**	1

J.6 1-BIT ADAM

Table 112: 1-Bit Adam sweep: 60M, 1× budget.

β_1	β_2	ϵ	η	clip	WU	WD	T_c/T	Val. loss	Run
0.87	0.98	1e-8	0.002	1.0	1000	0.1	0.10	3.860	1
0.90	0.98	1e-8	0.002	1.0	1000	0.1	0.10	3.809	2
0.95	0.98	1e-8	0.002	1.0	1000	0.1	0.10	3.805	3
0.98	0.98	1e-8	0.002	1.0	1000	0.1	0.10	5.707	4
0.95	**0.90**	1e-8	0.002	1.0	1000	0.1	0.10	3.990	5
0.95	**0.95**	1e-8	0.002	1.0	1000	0.1	0.10	3.947	6
0.95	**0.98**	1e-8	0.002	1.0	1000	0.1	0.10	3.805	7
0.95	0.98	**1e-20**	0.002	1.0	1000	0.1	0.10	nonfinite	8
0.95	0.98	**1e-10**	0.002	1.0	1000	0.1	0.10	5.916	9
0.95	0.98	**1e-8**	0.002	1.0	1000	0.1	0.10	3.805	10
0.95	0.98	1e-8	**5e-4**	1.0	1000	0.1	0.10	4.191	11
0.95	0.98	1e-8	**0.001**	1.0	1000	0.1	0.10	3.792	12
0.95	0.98	1e-8	**0.002**	1.0	1000	0.1	0.10	3.805	13
0.95	0.98	1e-8	**0.004**	1.0	1000	0.1	0.10	3.872	14
0.95	0.98	1e-8	0.001	**0**	1000	0.1	0.10	nonfinite	15
0.95	0.98	1e-8	0.001	**0.5**	1000	0.1	0.10	3.883	16
0.95	0.98	1e-8	0.001	**1.0**	1000	0.1	0.10	3.792	17
0.95	0.98	1e-8	0.001	**2.0**	1000	0.1	0.10	3.851	18
0.95	0.98	1e-8	0.001	1.0	**1000**	0.1	0.10	3.792	19
0.95	0.98	1e-8	0.001	1.0	**2000**	0.1	0.10	3.855	20
0.95	0.98	1e-8	0.001	1.0	1000	**0**	0.10	3.792	21
0.95	0.98	1e-8	0.001	1.0	1000	**0.1**	0.10	3.792	22
0.95	0.98	1e-8	0.001	1.0	1000	**0.5**	0.10	3.792	23
0.95	0.98	1e-8	0.001	1.0	1000	0.1	**0.05**	**3.760**	24
0.95	0.98	1e-8	0.001	1.0	1000	0.1	**0.10**	3.792	25

Table 113: 1-Bit Adam sweep: 60M, 2× budget.

β_1	β_2	ϵ	η	clip	WU	WD	T_c/T	Val. loss	Run
0.87	0.98	1e-8	0.002	1.0	1000	0.1	0.10	3.574	1
0.90	0.98	1e-8	0.002	1.0	1000	0.1	0.10	3.584	2
0.95	0.98	1e-8	0.002	1.0	1000	0.1	0.10	3.607	3
0.98	0.98	1e-8	0.002	1.0	1000	0.1	0.10	5.221	4
0.87	**0.90**	1e-8	0.002	1.0	1000	0.1	0.10	3.750	5
0.87	**0.95**	1e-8	0.002	1.0	1000	0.1	0.10	3.654	6
0.87	**0.98**	1e-8	0.002	1.0	1000	0.1	0.10	3.574	7
0.87	0.98	**1e-20**	0.002	1.0	1000	0.1	0.10	nonfinite	8
0.87	0.98	**1e-10**	0.002	1.0	1000	0.1	0.10	nonfinite	9
0.87	0.98	**1e-8**	0.002	1.0	1000	0.1	0.10	3.574	10
0.87	0.98	1e-8	**5e-4**	1.0	1000	0.1	0.10	3.711	11
0.87	0.98	1e-8	**0.001**	1.0	1000	0.1	0.10	3.635	12
0.87	0.98	1e-8	**0.002**	1.0	1000	0.1	0.10	3.574	13
0.87	0.98	1e-8	**0.004**	1.0	1000	0.1	0.10	3.650	14
0.87	0.98	1e-8	0.002	**0**	1000	0.1	0.10	nonfinite	15
0.87	0.98	1e-8	0.002	**0.5**	1000	0.1	0.10	3.570	16
0.87	0.98	1e-8	0.002	**1.0**	1000	0.1	0.10	3.574	17
0.87	0.98	1e-8	0.002	**2.0**	1000	0.1	0.10	3.727	18
0.87	0.98	1e-8	0.002	0.5	**1000**	0.1	0.10	3.570	19
0.87	0.98	1e-8	0.002	0.5	**2000**	0.1	0.10	**3.561**	20
0.87	0.98	1e-8	0.002	0.5	2000	0.1	**0.05**	3.607	21
0.87	0.98	1e-8	0.002	0.5	2000	0.1	**0.10**	3.561	22

Table 114: 1-Bit Adam sweep: 60M, 4× budget.

β_1	β_2	ϵ	η	clip	WU	WD	T_c/T	Val. loss	Run
0.87	0.98	1e-8	0.002	1.0	1000	0.1	0.10	3.507	1
0.90	0.98	1e-8	0.002	1.0	1000	0.1	0.10	3.513	2
0.95	0.98	1e-8	0.002	1.0	1000	0.1	0.10	3.523	3
0.98	0.98	1e-8	0.002	1.0	1000	0.1	0.10	5.169	4
0.87	**0.90**	1e-8	0.002	1.0	1000	0.1	0.10	3.513	5
0.87	**0.95**	1e-8	0.002	1.0	1000	0.1	0.10	3.498	6
0.87	**0.98**	1e-8	0.002	1.0	1000	0.1	0.10	3.507	7
0.87	0.95	**1e-20**	0.002	1.0	1000	0.1	0.10	4.004	8
0.87	0.95	**1e-10**	0.002	1.0	1000	0.1	0.10	3.627	9
0.87	0.95	**1e-8**	0.002	1.0	1000	0.1	0.10	3.498	10
0.87	0.95	1e-8	**5e-4**	1.0	1000	0.1	0.10	3.587	11
0.87	0.95	1e-8	**0.001**	1.0	1000	0.1	0.10	**3.494**	12
0.87	0.95	1e-8	**0.002**	1.0	1000	0.1	0.10	3.498	13
0.87	0.95	1e-8	**0.004**	1.0	1000	0.1	0.10	3.499	14
0.87	0.95	1e-8	0.001	**0**	1000	0.1	0.10	nonfinite	15
0.87	0.95	1e-8	0.001	**0.5**	1000	0.1	0.10	3.502	16
0.87	0.95	1e-8	0.001	**1.0**	1000	0.1	0.10	3.494	17
0.87	0.95	1e-8	0.001	**2.0**	1000	0.1	0.10	3.540	18
0.87	0.95	1e-8	0.001	1.0	**1000**	0.1	0.10	3.494	19
0.87	0.95	1e-8	0.001	1.0	**2000**	0.1	0.10	3.507	20
0.87	0.95	1e-8	0.001	1.0	1000	0.1	**0.05**	3.584	21
0.87	0.95	1e-8	0.001	1.0	1000	0.1	**0.10**	3.494	22

Table 115: 1-Bit Adam sweep: 60M, 8× budget.

β_1	β_2	ϵ	η	clip	WU	WD	T_c/T	Val. loss	Run
0.87	0.98	1e-8	0.002	1.0	1000	0.1	0.10	3.482	1
0.90	0.98	1e-8	0.002	1.0	1000	0.1	0.10	3.503	2
0.95	0.98	1e-8	0.002	1.0	1000	0.1	0.10	3.523	3
0.98	0.98	1e-8	0.002	1.0	1000	0.1	0.10	4.915	4
0.87	**0.90**	1e-8	0.002	1.0	1000	0.1	0.10	3.443	5
0.87	**0.95**	1e-8	0.002	1.0	1000	0.1	0.10	3.471	6
0.87	**0.98**	1e-8	0.002	1.0	1000	0.1	0.10	3.482	7
0.87	0.90	**1e-20**	0.002	1.0	1000	0.1	0.10	3.544	8
0.87	0.90	**1e-10**	0.002	1.0	1000	0.1	0.10	3.491	9
0.87	0.90	**1e-8**	0.002	1.0	1000	0.1	0.10	3.443	10
0.87	0.90	1e-8	**5e-4**	1.0	1000	0.1	0.10	3.494	11
0.87	0.90	1e-8	**0.001**	1.0	1000	0.1	0.10	**3.411**	12
0.87	0.90	1e-8	**0.002**	1.0	1000	0.1	0.10	3.443	13
0.87	0.90	1e-8	**0.004**	1.0	1000	0.1	0.10	5.435	14
0.87	0.90	1e-8	0.001	**0**	1000	0.1	0.10	nonfinite	15
0.87	0.90	1e-8	0.001	**0.5**	1000	0.1	0.10	3.435	16
0.87	0.90	1e-8	0.001	**1.0**	1000	0.1	0.10	3.411	17
0.87	0.90	1e-8	0.001	**2.0**	1000	0.1	0.10	3.429	18
0.87	0.90	1e-8	0.001	1.0	**1000**	0.1	0.10	3.411	19
0.87	0.90	1e-8	0.001	1.0	**2000**	0.1	0.10	3.425	20
0.87	0.90	1e-8	0.001	1.0	1000	0.1	**0.05**	3.442	21
0.87	0.90	1e-8	0.001	1.0	1000	0.1	**0.10**	3.411	22

Table 116: 1-Bit Adam sweep: 130M, 1× budget.

β_1	β_2	ϵ	η	clip	WU	WD	T_c/T	Val. loss	Run
0.87	0.98	1e-8	0.002	1.0	1000	0.1	0.10	3.660	1
0.90	0.98	1e-8	0.002	1.0	1000	0.1	0.10	5.214	2
0.95	0.98	1e-8	0.002	1.0	1000	0.1	0.10	5.425	3
0.98	0.98	1e-8	0.002	1.0	1000	0.1	0.10	5.321	4
0.87	**0.90**	1e-8	0.002	1.0	1000	0.1	0.10	3.981	5
0.87	**0.95**	1e-8	0.002	1.0	1000	0.1	0.10	3.566	6
0.87	**0.98**	1e-8	0.002	1.0	1000	0.1	0.10	3.660	7
0.87	0.95	**1e-8**	0.002	1.0	1000	0.1	0.10	3.566	8
0.87	0.95	**1e-10**	0.002	1.0	1000	0.1	0.10	9.504	9
0.87	0.95	**1e-20**	0.002	1.0	1000	0.1	0.10	nonfinite	10
0.87	0.95	1e-8	**5e-4**	1.0	1000	0.1	0.10	3.550	11
0.87	0.95	1e-8	**0.001**	1.0	1000	0.1	0.10	3.540	12
0.87	0.95	1e-8	**0.002**	1.0	1000	0.1	0.10	3.566	13
0.87	0.95	1e-8	**0.004**	1.0	1000	0.1	0.10	8.420	14
0.87	0.95	1e-8	0.001	**0**	1000	0.1	0.10	nonfinite	15
0.87	0.95	1e-8	0.001	**0.5**	1000	0.1	0.10	**3.436**	16
0.87	0.95	1e-8	0.001	**1.0**	1000	0.1	0.10	3.540	17
0.87	0.95	1e-8	0.001	**2.0**	1000	0.1	0.10	5.313	18
0.87	0.95	1e-8	0.001	0.5	**1000**	0.1	0.10	3.436	19
0.87	0.95	1e-8	0.001	0.5	**2000**	0.1	0.10	3.469	20
0.87	0.95	1e-8	0.001	0.5	1000	0.1	**0.05**	3.526	21
0.87	0.95	1e-8	0.001	0.5	1000	0.1	**0.10**	3.436	22

Table 117: 1-Bit Adam sweep: 130M, 2× budget (scaling-sensitive).

β_1	β_2	ϵ	η	clip	WU	WD	T_c/T	Val. loss	Run
0.87	**0.95**	1e-8	0.001	0.5	1000	0.1	0.10	3.279	1
0.87	**0.90**	1e-8	0.001	0.5	1000	0.1	0.10	**3.278**	2
0.87	0.95	1e-8	0.001	**1.0**	1000	0.1	0.10	3.305	3

Table 118: 1-Bit Adam sweep: 130M, 4× budget (scaling-sensitive).

β_1	β_2	ϵ	η	clip	WU	WD	T_c/T	Val. loss	Run
0.87	**0.95**	1e-8	0.001	0.5	1000	0.1	0.10	**3.206**	1
0.87	**0.90**	1e-8	0.001	0.5	1000	0.1	0.10	3.207	2
0.87	0.95	1e-8	0.001	**1.0**	1000	0.1	0.10	3.217	3

Table 119: 1-Bit Adam sweep: 130M, 8× budget (scaling-sensitive).

β_1	β_2	ϵ	η	clip	WU	WD	T_c/T	Val. loss	Run
0.87	**0.95**	1e-8	0.001	0.5	1000	0.1	0.10	**3.183**	1
0.87	**0.90**	1e-8	0.001	0.5	1000	0.1	0.10	3.185	2
0.87	0.95	1e-8	0.001	**1.0**	1000	0.1	0.10	3.187	3

Table 120: 1-Bit Adam sweep: 350M, 1× budget.

β_1	β_2	ϵ	η	clip	WU	WD	T_c/T	Val. loss	Run
0.87	0.98	1e-8	0.002	1.0	1000	0.1	0.10	nonfinite	1
0.90	0.98	1e-8	0.002	1.0	1000	0.1	0.10	nonfinite	2
0.95	0.98	1e-8	0.002	1.0	1000	0.1	0.10	5.036	3
0.98	0.98	1e-8	0.002	1.0	1000	0.1	0.10	4.666	4
0.98	**0.90**	1e-8	0.002	1.0	1000	0.1	0.10	5.436	5
0.98	**0.95**	1e-8	0.002	1.0	1000	0.1	0.10	nonfinite	6
0.98	**0.98**	1e-8	0.002	1.0	1000	0.1	0.10	4.666	7
0.98	0.98	**1e-8**	0.002	1.0	1000	0.1	0.10	4.666	8
0.98	0.98	**1e-10**	0.002	1.0	1000	0.1	0.10	9.746	9
0.98	0.98	**1e-20**	0.002	1.0	1000	0.1	0.10	nonfinite	10
0.98	0.98	1e-8	**5e-4**	1.0	1000	0.1	0.10	4.215	11
0.98	0.98	1e-8	**0.001**	1.0	1000	0.1	0.10	4.278	12
0.98	0.98	1e-8	**0.002**	1.0	1000	0.1	0.10	4.666	13
0.98	0.98	1e-8	**0.004**	1.0	1000	0.1	0.10	nonfinite	14
0.98	0.98	1e-8	5e-4	**0**	1000	0.1	0.10	nonfinite	15
0.98	0.98	1e-8	5e-4	**0.5**	1000	0.1	0.10	**3.353**	16
0.98	0.98	1e-8	5e-4	**1.0**	1000	0.1	0.10	4.215	17
0.98	0.98	1e-8	5e-4	**2.0**	1000	0.1	0.10	4.822	18
0.98	0.98	1e-8	5e-4	0.5	**1000**	0.1	0.10	3.353	19
0.98	0.98	1e-8	5e-4	0.5	**2000**	0.1	0.10	3.358	20
0.98	0.98	1e-8	5e-4	0.5	1000	0.1	**0.05**	3.660	21
0.98	0.98	1e-8	5e-4	0.5	1000	0.1	**0.10**	3.353	22

Table 121: 1-Bit Adam sweep: 350M, 2× budget (scaling-sensitive).

β_1	β_2	ϵ	η	clip	WU	WD	T_c/T	Val. loss	Run
0.98	0.98	1e-8	5e-4	0.5	1000	0.1	0.10	**3.226**	1

Table 122: 1-Bit Adam sweep: 350M, 4× budget (scaling-sensitive).

β_1	β_2	ϵ	η	clip	WU	WD	T_c/T	Val. loss	Run
0.98	0.98	1e-8	5e-4	0.5	1000	0.1	0.10	**3.123**	1

Table 123: 1-Bit Adam sweep: 350M, 8× budget (scaling-sensitive).

β_1	β_2	ϵ	η	clip	WU	WD	T_c/T	Val. loss	Run
0.98	0.98	1e-8	5e-4	0.5	1000	0.1	0.10	**3.041**	1

Table 124: 1-Bit Adam sweep: 1B, 1× budget.

β_1	β_2	ϵ	η	clip	WU	WD	T_c/T	Val. loss	Run
0.98	0.98	1e-8	2.5e-4	0.5	1000	0.1	0.10	**3.297**	1

J.7 0/1 Adam

Table 125: 0/1 Adam sweep: 60M, 1× budget.

β_1	β_2	ϵ	η	clip	WU	WD	L	T_v	Val. loss	Run
0.87	0.999	1e-8	0.008	2.0	2000	0.1	4	8	nonfinite	1
0.90	0.999	1e-8	0.008	2.0	2000	0.1	4	8	7.201	2
0.95	0.999	1e-8	0.008	2.0	2000	0.1	4	8	5.917	3
0.98	0.999	1e-8	0.008	2.0	2000	0.1	4	8	3.831	4
0.98	**0.90**	1e-8	0.008	2.0	2000	0.1	4	8	nonfinite	5
0.98	**0.95**	1e-8	0.008	2.0	2000	0.1	4	8	nonfinite	6
0.98	**0.98**	1e-8	0.008	2.0	2000	0.1	4	8	4.039	7
0.98	**0.999**	1e-8	0.008	2.0	2000	0.1	4	8	3.831	8
0.98	0.999	**1e-8**	0.008	2.0	2000	0.1	4	8	3.831	9
0.98	0.999	**1e-10**	0.008	2.0	2000	0.1	4	8	4.028	10
0.98	0.999	**1e-20**	0.008	2.0	2000	0.1	4	8	nonfinite	11
0.98	0.999	1e-8	**5e-4**	2.0	2000	0.1	4	8	5.219	12
0.98	0.999	1e-8	**0.001**	2.0	2000	0.1	4	8	4.746	13
0.98	0.999	1e-8	**0.002**	2.0	2000	0.1	4	8	4.332	14
0.98	0.999	1e-8	**0.004**	2.0	2000	0.1	4	8	4.024	15
0.98	0.999	1e-8	**0.008**	2.0	2000	0.1	4	8	3.831	16
0.98	0.999	1e-8	0.008	**0**	2000	0.1	4	8	nonfinite	17
0.98	0.999	1e-8	0.008	**0.5**	2000	0.1	4	8	**3.744**	18
0.98	0.999	1e-8	0.008	**1.0**	2000	0.1	4	8	3.768	19
0.98	0.999	1e-8	0.008	**2.0**	2000	0.1	4	8	3.831	20
0.98	0.999	1e-8	0.008	0.5	**1000**	0.1	4	8	4.289	21
0.98	0.999	1e-8	0.008	0.5	**2000**	0.1	4	8	3.744	22
0.98	0.999	1e-8	0.008	0.5	2000	**0**	4	8	3.744	23
0.98	0.999	1e-8	0.008	0.5	2000	**0.1**	4	8	3.744	24
0.98	0.999	1e-8	0.008	0.5	2000	**0.5**	4	8	3.744	25
0.98	0.999	1e-8	0.008	0.5	2000	0.1	**1**	8	nonfinite	26
0.98	0.999	1e-8	0.008	0.5	2000	0.1	**2**	8	4.062	27
0.98	0.999	1e-8	0.008	0.5	2000	0.1	**4**	8	3.744	28
0.98	0.999	1e-8	0.008	0.5	2000	0.1	**8**	8	5.391	29
0.98	0.999	1e-8	0.008	0.5	2000	0.1	4	**1**	nonfinite	30
0.98	0.999	1e-8	0.008	0.5	2000	0.1	4	**8**	3.744	32
0.98	0.999	1e-8	0.008	0.5	2000	0.1	4	**16**	3.911	33

Table 126: 0/1 Adam sweep: 60M, 2× budget.

β_1	β_2	ϵ	η	clip	WU	WD	L	T_v	Val. loss	Run
0.87	0.999	1e-8	0.008	2.0	2000	0.1	4	8	nonfinite	1
0.90	0.999	1e-8	0.008	2.0	2000	0.1	4	8	6.381	2
0.95	0.999	1e-8	0.008	2.0	2000	0.1	4	8	5.416	3
0.98	0.999	1e-8	0.008	2.0	2000	0.1	4	8	3.688	4
0.98	**0.90**	1e-8	0.008	2.0	2000	0.1	4	8	nonfinite	5
0.98	**0.95**	1e-8	0.008	2.0	2000	0.1	4	8	nonfinite	6
0.98	**0.98**	1e-8	0.008	2.0	2000	0.1	4	8	3.766	7
0.98	**0.999**	1e-8	0.008	2.0	2000	0.1	4	8	3.688	8
0.98	0.999	**1e-20**	0.008	2.0	2000	0.1	4	8	nonfinite	9
0.98	0.999	**1e-10**	0.008	2.0	2000	0.1	4	8	3.849	10
0.98	0.999	**1e-8**	0.008	2.0	2000	0.1	4	8	3.688	11
0.98	0.999	1e-8	**5e-4**	2.0	2000	0.1	4	8	4.823	12
0.98	0.999	1e-8	**0.001**	2.0	2000	0.1	4	8	4.371	13
0.98	0.999	1e-8	**0.002**	2.0	2000	0.1	4	8	4.059	14
0.98	0.999	1e-8	**0.004**	2.0	2000	0.1	4	8	3.830	15
0.98	0.999	1e-8	**0.008**	2.0	2000	0.1	4	8	3.688	16
0.98	0.999	1e-8	0.008	**0**	2000	0.1	4	8	nonfinite	17
0.98	0.999	1e-8	0.008	**0.5**	2000	0.1	4	8	**3.609**	18
0.98	0.999	1e-8	0.008	**1.0**	2000	0.1	4	8	3.630	19
0.98	0.999	1e-8	0.008	**2.0**	2000	0.1	4	8	3.688	20
0.98	0.999	1e-8	0.008	0.5	**1000**	0.1	4	8	3.956	21
0.98	0.999	1e-8	0.008	0.5	**2000**	0.1	4	8	3.609	22
0.98	0.999	1e-8	0.008	0.5	2000	**0**	4	8	3.609	23
0.98	0.999	1e-8	0.008	0.5	2000	**0.1**	4	8	3.609	24
0.98	0.999	1e-8	0.008	0.5	2000	**0.5**	4	8	3.609	25
0.98	0.999	1e-8	0.008	0.5	2000	0.1	**1**	8	nonfinite	26
0.98	0.999	1e-8	0.008	0.5	2000	0.1	**2**	8	3.873	27
0.98	0.999	1e-8	0.008	0.5	2000	0.1	**4**	8	3.609	28
0.98	0.999	1e-8	0.008	0.5	2000	0.1	**8**	8	4.954	29
0.98	0.999	1e-8	0.008	0.5	2000	0.1	4	**1**	nonfinite	30
0.98	0.999	1e-8	0.008	0.5	2000	0.1	4	**8**	3.609	32
0.98	0.999	1e-8	0.008	0.5	2000	0.1	4	**16**	3.705	33

Table 127: 0/1 Adam sweep: 60M, 4× budget.

β_1	β_2	ϵ	η	clip	WU	WD	L	T_v	Val. loss	Run
0.87	0.999	1e-8	0.008	2.0	2000	0.1	4	8	nonfinite	1
0.90	0.999	1e-8	0.008	2.0	2000	0.1	4	8	5.836	2
0.95	0.999	1e-8	0.008	2.0	2000	0.1	4	8	4.932	3
0.98	0.999	1e-8	0.008	2.0	2000	0.1	4	8	3.572	4
0.98	**0.90**	1e-8	0.008	2.0	2000	0.1	4	8	nonfinite	5
0.98	**0.95**	1e-8	0.008	2.0	2000	0.1	4	8	nonfinite	6
0.98	**0.98**	1e-8	0.008	2.0	2000	0.1	4	8	3.586	7
0.98	**0.999**	1e-8	0.008	2.0	2000	0.1	4	8	3.572	8
0.98	0.999	**1e-20**	0.008	2.0	2000	0.1	4	8	nonfinite	9
0.98	0.999	**1e-10**	0.008	2.0	2000	0.1	4	8	3.706	10
0.98	0.999	**1e-8**	0.008	2.0	2000	0.1	4	8	3.572	11
0.98	0.999	1e-8	**5e-4**	2.0	2000	0.1	4	8	4.486	12
0.98	0.999	1e-8	**0.001**	2.0	2000	0.1	4	8	4.092	13
0.98	0.999	1e-8	**0.002**	2.0	2000	0.1	4	8	3.862	14
0.98	0.999	1e-8	**0.004**	2.0	2000	0.1	4	8	3.683	15
0.98	0.999	1e-8	**0.008**	2.0	2000	0.1	4	8	3.572	16
0.98	0.999	1e-8	0.008	**0**	2000	0.1	4	8	8.886	17
0.98	0.999	1e-8	0.008	**0.5**	2000	0.1	4	8	**3.511**	18
0.98	0.999	1e-8	0.008	**1.0**	2000	0.1	4	8	3.526	19
0.98	0.999	1e-8	0.008	**2.0**	2000	0.1	4	8	3.572	20
0.98	0.999	1e-8	0.008	0.5	**1000**	0.1	4	8	3.769	21
0.98	0.999	1e-8	0.008	0.5	**2000**	0.1	4	8	3.511	22
0.98	0.999	1e-8	0.008	0.5	2000	**0**	4	8	3.511	23
0.98	0.999	1e-8	0.008	0.5	2000	**0.1**	4	8	3.511	24
0.98	0.999	1e-8	0.008	0.5	2000	**0.5**	4	8	3.511	25
0.98	0.999	1e-8	0.008	0.5	2000	0.1	**1**	8	nonfinite	26
0.98	0.999	1e-8	0.008	0.5	2000	0.1	**2**	8	3.729	27
0.98	0.999	1e-8	0.008	0.5	2000	0.1	**4**	8	3.511	28
0.98	0.999	1e-8	0.008	0.5	2000	0.1	**8**	8	4.391	29
0.98	0.999	1e-8	0.008	0.5	2000	0.1	4	**1**	nonfinite	30
0.98	0.999	1e-8	0.008	0.5	2000	0.1	4	**8**	3.511	32
0.98	0.999	1e-8	0.008	0.5	2000	0.1	4	**16**	3.564	33

Table 128: 0/1 Adam sweep: 60M, 8× budget.

β_1	β_2	ϵ	η	clip	WU	WD	L	T_v	Val. loss	Run
0.87	0.999	1e-8	0.008	2.0	2000	0.1	4	8	51.048	1
0.90	0.999	1e-8	0.008	2.0	2000	0.1	4	8	8.640	2
0.95	0.999	1e-8	0.008	2.0	2000	0.1	4	8	3.665	3
0.98	0.999	1e-8	0.008	2.0	2000	0.1	4	8	3.478	4
0.98	**0.90**	1e-8	0.008	2.0	2000	0.1	4	8	nonfinite	5
0.98	**0.95**	1e-8	0.008	2.0	2000	0.1	4	8	nonfinite	6
0.98	**0.98**	1e-8	0.008	2.0	2000	0.1	4	8	3.491	7
0.98	**0.999**	1e-8	0.008	2.0	2000	0.1	4	8	3.478	8
0.98	0.999	**1e-20**	0.008	2.0	2000	0.1	4	8	10.374	9
0.98	0.999	**1e-10**	0.008	2.0	2000	0.1	4	8	3.715	10
0.98	0.999	**1e-8**	0.008	2.0	2000	0.1	4	8	3.478	11
0.98	0.999	1e-8	**5e-4**	2.0	2000	0.1	4	8	4.215	12
0.98	0.999	1e-8	**0.001**	2.0	2000	0.1	4	8	3.897	13
0.98	0.999	1e-8	**0.002**	2.0	2000	0.1	4	8	3.709	14
0.98	0.999	1e-8	**0.004**	2.0	2000	0.1	4	8	3.560	15
0.98	0.999	1e-8	**0.008**	2.0	2000	0.1	4	8	3.478	16
0.98	0.999	1e-8	0.008	**0**	2000	0.1	4	8	53.966	17
0.98	0.999	1e-8	0.008	**0.5**	2000	0.1	4	8	**3.437**	18
0.98	0.999	1e-8	0.008	**1.0**	2000	0.1	4	8	3.440	19
0.98	0.999	1e-8	0.008	**2.0**	2000	0.1	4	8	3.478	20
0.98	0.999	1e-8	0.008	0.5	**1000**	0.1	4	8	3.525	21
0.98	0.999	1e-8	0.008	0.5	**2000**	0.1	4	8	3.437	22
0.98	0.999	1e-8	0.008	0.5	2000	**0**	4	8	3.437	23
0.98	0.999	1e-8	0.008	0.5	2000	**0.1**	4	8	3.437	24
0.98	0.999	1e-8	0.008	0.5	2000	**0.5**	4	8	3.437	25
0.98	0.999	1e-8	0.008	0.5	2000	0.1	**1**	8	10.490	26
0.98	0.999	1e-8	0.008	0.5	2000	0.1	**2**	8	3.615	27
0.98	0.999	1e-8	0.008	0.5	2000	0.1	**4**	8	3.437	28
0.98	0.999	1e-8	0.008	0.5	2000	0.1	**8**	8	4.229	29
0.98	0.999	1e-8	0.008	0.5	2000	0.1	4	**1**	10.490	30
0.98	0.999	1e-8	0.008	0.5	2000	0.1	4	**8**	3.437	32
0.98	0.999	1e-8	0.008	0.5	2000	0.1	4	**16**	3.462	33

Table 129: 0/1 Adam sweep: 130M, 1× budget.

β_1	β_2	ϵ	η	clip	WU	WD	L	T_v	Val. loss	Run
0.87	0.999	1e-8	0.008	2.0	2000	0.1	4	8	110.190	1
0.90	0.999	1e-8	0.008	2.0	2000	0.1	4	8	41.271	2
0.95	0.999	1e-8	0.008	2.0	2000	0.1	4	8	5.500	3
0.98	0.999	1e-8	0.008	2.0	2000	0.1	4	8	3.485	4
0.98	**0.90**	1e-8	0.008	2.0	2000	0.1	4	8	969.147	5
0.98	**0.95**	1e-8	0.008	2.0	2000	0.1	4	8	nonfinite	6
0.98	**0.98**	1e-8	0.008	2.0	2000	0.1	4	8	3.492	7
0.98	**0.999**	1e-8	0.008	2.0	2000	0.1	4	8	3.485	8
0.98	0.999	**1e-20**	0.008	2.0	2000	0.1	4	8	10.374	9
0.98	0.999	**1e-10**	0.008	2.0	2000	0.1	4	8	8.781	10
0.98	0.999	**1e-8**	0.008	2.0	2000	0.1	4	8	3.485	11
0.98	0.999	1e-8	**5e-4**	2.0	2000	0.1	4	8	4.457	12
0.98	0.999	1e-8	**0.001**	2.0	2000	0.1	4	8	4.086	13
0.98	0.999	1e-8	**0.002**	2.0	2000	0.1	4	8	3.777	14
0.98	0.999	1e-8	**0.004**	2.0	2000	0.1	4	8	3.581	15
0.98	0.999	1e-8	**0.008**	2.0	2000	0.1	4	8	3.485	16
0.98	0.999	1e-8	0.008	**0**	2000	0.1	4	8	323.355	17
0.98	0.999	1e-8	0.008	**0.5**	2000	0.1	4	8	**3.475**	18
0.98	0.999	1e-8	0.008	**1.0**	2000	0.1	4	8	3.491	19
0.98	0.999	1e-8	0.008	**2.0**	2000	0.1	4	8	3.485	20
0.98	0.999	1e-8	0.008	0.5	**1000**	0.1	4	8	3.701	21
0.98	0.999	1e-8	0.008	0.5	**2000**	0.1	4	8	3.475	22
0.98	0.999	1e-8	0.008	0.5	2000	**0**	4	8	3.475	23
0.98	0.999	1e-8	0.008	0.5	2000	**0.1**	4	8	3.475	24
0.98	0.999	1e-8	0.008	0.5	2000	**0.5**	4	8	3.475	25
0.98	0.999	1e-8	0.008	0.5	2000	0.1	**1**	8	10.505	26
0.98	0.999	1e-8	0.008	0.5	2000	0.1	**2**	8	3.615	27
0.98	0.999	1e-8	0.008	0.5	2000	0.1	**4**	8	3.475	28
0.98	0.999	1e-8	0.008	0.5	2000	0.1	**8**	8	5.566	29
0.98	0.999	1e-8	0.008	0.5	2000	0.1	4	**1**	10.505	30
0.98	0.999	1e-8	0.008	0.5	2000	0.1	4	**8**	3.475	32
0.98	0.999	1e-8	0.008	0.5	2000	0.1	4	**16**	3.686	33

Table 130: 0/1 Adam sweep: 130M, 2× budget (scaling-sensitive).

β_1	β_2	ϵ	η	clip	WU	WD	L	T_v	Val. loss	Run
0.98	0.999	1e-8	0.008	0.5	2000	0.1	4	**8**	**3.389**	2

Table 131: 0/1 Adam sweep: 130M, 4× budget (scaling-sensitive).

β_1	β_2	ϵ	η	clip	WU	WD	L	T_v	Val. loss	Run
0.98	0.999	1e-8	0.008	0.5	2000	0.1	4	**8**	**3.279**	2

Table 132: 0/1 Adam sweep: 130M, 8× budget (scaling-sensitive).

β_1	β_2	ϵ	η	clip	WU	WD	L	T_v	Val. loss	Run
0.98	0.999	1e-8	0.008	0.5	2000	0.1	4	**8**	**3.193**	2

Table 133: 0/1 Adam sweep: 350M, 1× budget.

β_1	β_2	ϵ	η	clip	WU	WD	L	T_v	Val. loss	Run
0.87	0.999	1e-8	0.008	2.0	2000	0.1	4	8	nonfinite	1
0.90	0.999	1e-8	0.008	2.0	2000	0.1	4	8	nonfinite	2
0.95	0.999	1e-8	0.008	2.0	2000	0.1	4	8	5.283	3
0.98	0.999	1e-8	0.008	2.0	2000	0.1	4	8	3.182	4
0.98	**0.90**	1e-8	0.008	2.0	2000	0.1	4	8	nonfinite	5
0.98	**0.95**	1e-8	0.008	2.0	2000	0.1	4	8	nonfinite	6
0.98	**0.98**	1e-8	0.008	2.0	2000	0.1	4	8	3.144	7
0.98	**0.999**	1e-8	0.008	2.0	2000	0.1	4	8	3.182	8
0.98	0.98	**1e-8**	0.008	2.0	2000	0.1	4	8	3.144	9
0.98	0.98	**1e-10**	0.008	2.0	2000	0.1	4	8	nonfinite	10
0.98	0.98	**1e-20**	0.008	2.0	2000	0.1	4	8	nonfinite	11
0.98	0.98	1e-8	**5e-4**	2.0	2000	0.1	4	8	3.746	12
0.98	0.98	1e-8	**0.001**	2.0	2000	0.1	4	8	3.466	13
0.98	0.98	1e-8	**0.002**	2.0	2000	0.1	4	8	3.271	14
0.98	0.98	1e-8	**0.004**	2.0	2000	0.1	4	8	3.146	15
0.98	0.98	1e-8	**0.008**	2.0	2000	0.1	4	8	3.144	16
0.98	0.98	1e-8	0.008	**0**	2000	0.1	4	8	nonfinite	17
0.98	0.98	1e-8	0.008	**0.5**	2000	0.1	4	8	nonfinite	18
0.98	0.98	1e-8	0.008	**1.0**	2000	0.1	4	8	nonfinite	19
0.98	0.98	1e-8	0.008	**2.0**	2000	0.1	4	8	3.144	20
0.98	0.98	1e-8	0.008	2.0	**1000**	0.1	4	8	3.156	21
0.98	0.98	1e-8	0.008	2.0	**2000**	0.1	4	8	3.144	22
0.98	0.98	1e-8	0.008	2.0	2000	**0**	4	8	3.144	23
0.98	0.98	1e-8	0.008	2.0	2000	**0.1**	4	8	3.144	24
0.98	0.98	1e-8	0.008	2.0	2000	**0.5**	4	8	3.144	25
0.98	0.98	1e-8	0.008	2.0	2000	0.1	**1**	8	nonfinite	26
0.98	0.98	1e-8	0.008	2.0	2000	0.1	**2**	8	3.274	27
0.98	0.98	1e-8	0.008	2.0	2000	0.1	**4**	8	3.144	28
0.98	0.98	1e-8	0.008	2.0	2000	0.1	**8**	8	nonfinite	29
0.98	0.98	1e-8	0.008	2.0	2000	0.1	4	**1**	nonfinite	30
0.98	0.98	1e-8	0.008	2.0	2000	0.1	4	**4**	nonfinite	31
0.98	0.98	1e-8	0.008	2.0	2000	0.1	4	**8**	3.144	32
0.98	0.98	1e-8	0.008	2.0	2000	0.1	4	**16**	**3.128**	33

Table 134: 0/1 Adam sweep: 350M, 2× budget (scaling-sensitive).

β_1	β_2	ϵ	η	clip	WU	WD	L	T_v	Val. loss	Run
0.98	0.98	1e-8	0.008	2.0	2000	0.1	4	**16**	**3.047**	1
0.98	0.98	1e-8	0.008	2.0	2000	0.1	4	**8**	nonfinite	2

Table 135: 0/1 Adam sweep: 350M, 4× budget (scaling-sensitive).

β_1	β_2	ϵ	η	clip	WU	WD	L	T_v	Val. loss	Run
0.98	0.98	1e-8	0.008	2.0	2000	0.1	4	16	nonfinite	1
0.98	0.98	1e-8	0.008	2.0	2000	0.1	4	8	nonfinite	2
0.98	0.98	1e-8	0.004	2.0	2000	0.1	4	16	2.920	3
0.98	0.98	1e-8	0.004	2.0	2000	0.1	4	**8**	**2.918**	4

Table 136: 0/1 Adam sweep: 350M, 8× budget (scaling-sensitive).

β_1	β_2	ϵ	η	clip	WU	WD	L	T_v	Val. loss	Run
0.98	0.98	1e-8	0.008	2.0	2000	0.1	4	16	10.374	1
0.98	0.98	1e-8	0.008	2.0	2000	0.1	4	8	10.374	2
0.98	0.98	1e-8	0.004	2.0	2000	0.1	4	16	2.862	3
0.98	0.98	1e-8	0.004	2.0	2000	0.1	4	**8**	**2.862**	4

Table 137: 0/1 Adam configuration: 1B, 1× budget.

β_1	β_2	ϵ	η	clip	WU	WD	L	T_v	Val. loss	Run
0.98	0.98	1e-8	0.002	2.0	2000	0.1	4	16	**2.971**	1

J.8 ADASIGNSGD

Table 138: ADASIGNSGD sweep: 60M, 1× budget.

β	ε	ϵ_p	η	clip	WU	WD	Val. loss	Run
0.87	1e-8	1e-4	0.001	0	2000	0.5	3.731	1
0.90	1e-8	1e-4	0.001	0	2000	0.5	3.689	2
0.95	1e-8	1e-4	0.001	0	2000	0.5	3.624	3
0.98	1e-8	1e-4	0.001	0	2000	0.5	3.663	4
0.95	**1e-20**	1e-4	0.001	0	2000	0.5	3.623	5
0.95	**1e-10**	1e-4	0.001	0	2000	0.5	3.623	6
0.95	**1e-8**	1e-4	0.001	0	2000	0.5	3.624	7
0.95	1e-8	**1e-4**	0.001	0	2000	0.5	3.624	8
0.95	1e-8	**1e-3**	0.001	0	2000	0.5	3.624	9
0.95	1e-8	**5e-3**	0.001	0	2000	0.5	**3.622**	10
0.95	1e-8	**1e-2**	0.001	0	2000	0.5	3.623	11
0.95	1e-8	**5e-2**	0.001	0	2000	0.5	3.638	12
0.95	1e-8	1e-4	**5e-4**	0	2000	0.5	3.657	13
0.95	1e-8	1e-4	**0.001**	0	2000	0.5	3.624	14
0.95	1e-8	1e-4	**0.002**	0	2000	0.5	3.641	15
0.95	1e-8	1e-4	**0.004**	0	2000	0.5	3.648	16
0.95	1e-8	1e-4	**0.008**	0	2000	0.5	3.710	17
0.95	1e-8	1e-4	0.001	**0**	2000	0.5	3.624	18
0.95	1e-8	1e-4	0.001	**0.5**	2000	0.5	10.490	19
0.95	1e-8	1e-4	0.001	**1.0**	2000	0.5	10.489	20
0.95	1e-8	1e-4	0.001	**2.0**	2000	0.5	10.488	21
0.95	1e-8	1e-4	0.001	0	**1000**	0.5	3.643	22
0.95	1e-8	1e-4	0.001	0	**2000**	0.5	3.624	23
0.95	1e-8	1e-4	0.001	0	2000	**0**	3.661	24
0.95	1e-8	1e-4	0.001	0	2000	**0.1**	3.640	25
0.95	1e-8	1e-4	0.001	0	2000	**0.5**	3.624	26

Table 139: ADASIGNSGD sweep: 60M, 2× budget.

β	ε	ϵ_p	η	clip	WU	WD	Val. loss	Run
0.87	1e-10	1e-4	0.001	0	2000	0.5	3.661	1
0.90	1e-10	1e-4	0.001	0	2000	0.5	3.619	2
0.95	1e-10	1e-4	0.001	0	2000	0.5	**3.551**	3
0.98	1e-10	1e-4	0.001	0	2000	0.5	3.563	4
0.95	**1e-8**	1e-4	0.001	0	2000	0.5	3.554	5
0.95	**1e-10**	1e-4	0.001	0	2000	0.5	3.551	6
0.95	**1e-20**	1e-4	0.001	0	2000	0.5	3.551	7
0.95	1e-10	**1e-4**	0.001	0	2000	0.5	3.551	8
0.95	1e-10	**5e-3**	0.001	0	2000	0.5	3.552	9
0.95	1e-10	**1e-3**	0.001	0	2000	0.5	3.553	10
0.95	1e-10	**1e-2**	0.001	0	2000	0.5	3.552	11
0.95	1e-10	**5e-2**	0.001	0	2000	0.5	3.640	12
0.95	1e-10	1e-4	**5e-4**	0	2000	0.5	3.569	13
0.95	1e-10	1e-4	**0.001**	0	2000	0.5	3.551	14
0.95	1e-10	1e-4	**0.002**	0	2000	0.5	3.567	15
0.95	1e-10	1e-4	**0.004**	0	2000	0.5	3.569	16
0.95	1e-10	1e-4	**0.008**	0	2000	0.5	3.635	17
0.95	1e-10	1e-4	0.001	**0**	2000	0.5	3.551	18
0.95	1e-10	1e-4	0.001	**0.5**	2000	0.5	10.490	19
0.95	1e-10	1e-4	0.001	**1.0**	2000	0.5	10.489	20
0.95	1e-10	1e-4	0.001	**2.0**	2000	0.5	10.488	21
0.95	1e-10	1e-4	0.001	0	**1000**	0.5	3.563	22
0.95	1e-10	1e-4	0.001	0	**2000**	0.5	3.551	23
0.95	1e-10	1e-4	0.001	0	2000	**0**	3.600	24
0.95	1e-10	1e-4	0.001	0	2000	**0.1**	3.568	25
0.95	1e-10	1e-4	0.001	0	2000	**0.5**	3.551	26

Table 140: ADASIGNSGD sweep: 60M, 4× budget.

β	ε	ϵ_p	η	clip	WU	WD	Val. loss	Run
0.87	1e-8	1e-4	0.001	0	2000	0.5	3.616	1
0.90	1e-8	1e-4	0.001	0	2000	0.5	3.574	2
0.95	1e-8	1e-4	0.001	0	2000	0.5	3.510	3
0.98	1e-8	1e-4	0.001	0	2000	0.5	3.505	4
0.98	**1e-8**	1e-4	0.001	0	2000	0.5	3.505	5
0.98	**1e-10**	1e-4	0.001	0	2000	0.5	3.505	6
0.98	**1e-20**	1e-4	0.001	0	2000	0.5	3.505	7
0.98	1e-8	**1e-4**	0.001	0	2000	0.5	3.505	8
0.98	1e-8	**5e-3**	0.001	0	2000	0.5	3.504	9
0.98	1e-8	**1e-3**	0.001	0	2000	0.5	3.508	10
0.98	1e-8	**1e-2**	0.001	0	2000	0.5	3.503	11
0.98	1e-8	**5e-2**	0.001	0	2000	0.5	3.509	12
0.98	1e-8	1e-4	**5e-4**	0	2000	0.5	**3.486**	13
0.98	1e-8	1e-4	**0.001**	0	2000	0.5	3.505	14
0.98	1e-8	1e-4	**0.002**	0	2000	0.5	3.530	15
0.98	1e-8	1e-4	**0.004**	0	2000	0.5	3.547	16
0.98	1e-8	1e-4	**0.008**	0	2000	0.5	5.427	17
0.98	1e-8	1e-4	5e-4	**0**	2000	0.5	3.486	18
0.98	1e-8	1e-4	5e-4	**0.5**	2000	0.5	10.490	19
0.98	1e-8	1e-4	5e-4	**1.0**	2000	0.5	10.490	20
0.98	1e-8	1e-4	5e-4	**2.0**	2000	0.5	10.489	21
0.98	1e-8	1e-4	5e-4	0	**1000**	0.5	3.508	22
0.98	1e-8	1e-4	5e-4	0	**2000**	0.5	3.486	23
0.98	1e-8	1e-4	5e-4	0	2000	**0**	3.541	24
0.98	1e-8	1e-4	5e-4	0	2000	**0.1**	3.513	25
0.98	1e-8	1e-4	5e-4	0	2000	**0.5**	3.486	26

Table 141: ADASIGNSGD sweep: 60M, 8× budget.

β	ε	ϵ_p	η	clip	WU	WD	Val. loss	Run
0.87	1e-8	1e-4	5e-4	0	2000	0.5	3.612	1
0.90	1e-8	1e-4	5e-4	0	2000	0.5	3.566	2
0.95	1e-8	1e-4	5e-4	0	2000	0.5	3.475	3
0.98	1e-8	1e-4	5e-4	0	2000	0.5	3.436	4
0.98	**1e-20**	1e-4	5e-4	0	2000	0.5	**3.433**	5
0.98	**1e-10**	1e-4	5e-4	0	2000	0.5	3.433	6
0.98	**1e-8**	1e-4	5e-4	0	2000	0.5	3.436	7
0.98	1e-8	**1e-4**	5e-4	0	2000	0.5	3.436	8
0.98	1e-8	**1e-3**	5e-4	0	2000	0.5	3.437	9
0.98	1e-8	**5e-3**	5e-4	0	2000	0.5	3.439	10
0.98	1e-8	**1e-2**	5e-4	0	2000	0.5	3.441	11
0.98	1e-8	**5e-2**	5e-4	0	2000	0.5	3.463	12
0.98	1e-8	1e-4	**5e-4**	0	2000	0.5	3.436	13
0.98	1e-8	1e-4	**0.001**	0	2000	0.5	3.460	14
0.98	1e-8	1e-4	**0.002**	0	2000	0.5	3.492	15
0.98	1e-8	1e-4	**0.004**	0	2000	0.5	3.500	16
0.98	1e-8	1e-4	**0.008**	0	2000	0.5	4.716	17
0.98	1e-8	1e-4	5e-4	**0**	2000	0.5	3.436	18
0.98	1e-8	1e-4	5e-4	**0.5**	2000	0.5	10.490	19
0.98	1e-8	1e-4	5e-4	**1.0**	2000	0.5	10.490	20
0.98	1e-8	1e-4	5e-4	**2.0**	2000	0.5	10.489	21
0.98	1e-8	1e-4	5e-4	0	**1000**	0.5	3.450	22
0.98	1e-8	1e-4	5e-4	0	**2000**	0.5	3.436	23
0.98	1e-8	1e-4	5e-4	0	2000	**0**	3.472	24
0.98	1e-8	1e-4	5e-4	0	2000	**0.1**	3.450	25
0.98	1e-8	1e-4	5e-4	0	2000	**0.5**	3.436	26

Table 142: ADASIGNSGD sweep: 130M, 1× budget.

β	ε	ϵ_p	η	clip	WU	WD	Val. loss	Run
0.87	1e-8	1e-4	0.001	0	2000	0.5	3.422	1
0.90	1e-8	1e-4	0.001	0	2000	0.5	3.389	2
0.95	1e-8	1e-4	0.001	0	2000	0.5	3.349	3
0.98	1e-8	1e-4	0.001	0	2000	0.5	3.437	4
0.95	**1e-8**	1e-4	0.001	0	2000	0.5	3.349	5
0.95	**1e-10**	1e-4	0.001	0	2000	0.5	3.348	6
0.95	**1e-20**	1e-4	0.001	0	2000	0.5	3.348	7
0.95	1e-8	**1e-4**	0.001	0	2000	0.5	3.349	8
0.95	1e-8	**5e-3**	0.001	0	2000	0.5	3.351	9
0.95	1e-8	**1e-3**	0.001	0	2000	0.5	3.350	10
0.95	1e-8	**1e-2**	0.001	0	2000	0.5	3.347	11
0.95	1e-8	**5e-2**	0.001	0	2000	0.5	3.467	12
0.95	1e-8	1e-4	**5e-4**	0	2000	0.5	**3.344**	13
0.95	1e-8	1e-4	**0.001**	0	2000	0.5	3.349	14
0.95	1e-8	1e-4	**0.002**	0	2000	0.5	3.401	15
0.95	1e-8	1e-4	**0.004**	0	2000	0.5	3.629	16
0.95	1e-8	1e-4	**0.008**	0	2000	0.5	5.385	17
0.95	1e-8	1e-4	5e-4	**0**	2000	0.5	3.344	18
0.95	1e-8	1e-4	5e-4	**0.5**	2000	0.5	10.505	19
0.95	1e-8	1e-4	5e-4	**1.0**	2000	0.5	10.505	20
0.95	1e-8	1e-4	5e-4	**2.0**	2000	0.5	10.505	21
0.95	1e-8	1e-4	5e-4	0	**1000**	0.5	3.355	22
0.95	1e-8	1e-4	5e-4	0	**2000**	0.5	3.344	23

Table 143: ADASIGNSGD sweep: 130M, 2× budget (scaling-sensitive).

β	ε	ϵ_p	η	clip	WU	WD	Val. loss	Run
0.95	1e-8	1e-4	5e-4	0	2000	0.5	**3.281**	1
0.98	1e-8	1e-4	5e-4	0	2000	0.5	3.291	2
0.95	1e-8	1e-4	**0.001**	0	2000	0.5	3.297	3

Table 144: ADASIGNSGD sweep: 130M, 4× budget (scaling-sensitive).

β	ε	ϵ_p	η	clip	WU	WD	Val. loss	Run
0.95	1e-8	1e-4	5e-4	0	2000	0.5	3.235	1
0.98	1e-8	1e-4	5e-4	0	2000	0.5	**3.232**	2
0.95	1e-8	1e-4	**0.001**	0	2000	0.5	3.254	3

Table 145: ADASIGNSGD sweep: 130M, 8× budget (scaling-sensitive).

β	ε	ϵ_p	η	clip	WU	WD	Val. loss	Run
0.95	1e-8	1e-4	5e-4	0	2000	0.5	3.198	1
0.98	1e-8	1e-4	5e-4	0	2000	0.5	**3.191**	2
0.98	1e-8	1e-4	**0.001**	0	2000	0.5	3.265	3

Table 146: ADASIGNSGD sweep: 350M, 1× budget.

β	ε	ϵ_p	η	clip	WU	WD	Val. loss	Run
0.87	1e-10	1e-4	5e-4	0	2000	0.5	3.129	1
0.90	1e-10	1e-4	5e-4	0	2000	0.5	3.100	2
0.95	1e-10	1e-4	5e-4	0	2000	0.5	**3.050**	3
0.98	1e-10	1e-4	5e-4	0	2000	0.5	3.102	4
0.95	**1e-8**	1e-4	5e-4	0	2000	0.5	3.050	5
0.95	**1e-10**	1e-4	5e-4	0	2000	0.5	3.050	6
0.95	**1e-20**	1e-4	5e-4	0	2000	0.5	3.050	7
0.95	1e-10	**1e-4**	5e-4	0	2000	0.5	3.050	8
0.95	1e-10	**5e-3**	5e-4	0	2000	0.5	3.052	9
0.95	1e-10	**1e-3**	5e-4	0	2000	0.5	3.051	10
0.95	1e-10	**1e-2**	5e-4	0	2000	0.5	3.052	11
0.95	1e-10	**5e-2**	5e-4	0	2000	0.5	3.256	12
0.95	1e-10	1e-4	**2.5e-4**	0	2000	0.5	3.061	13
0.95	1e-10	1e-4	**5e-4**	0	2000	0.5	3.050	14
0.95	1e-10	1e-4	**0.001**	0	2000	0.5	3.110	15
0.95	1e-10	1e-4	**0.002**	0	2000	0.5	3.832	16
0.95	1e-10	1e-4	**0.004**	0	2000	0.5	3.272	17
0.95	1e-10	1e-4	5e-4	**0**	2000	0.5	3.050	18
0.95	1e-10	1e-4	5e-4	**0.5**	2000	0.5	nonfinite	19
0.95	1e-10	1e-4	5e-4	**1.0**	2000	0.5	nonfinite	20
0.95	1e-10	1e-4	5e-4	**2.0**	2000	0.5	nonfinite	21
0.95	1e-10	1e-4	5e-4	0	**1000**	0.5	3.074	22
0.95	1e-10	1e-4	5e-4	0	**2000**	0.5	3.050	23
0.95	1e-10	1e-4	5e-4	0	2000	**0**	3.124	24
0.95	1e-10	1e-4	5e-4	0	2000	**0.1**	3.067	25
0.95	1e-10	1e-4	5e-4	0	2000	**0.5**	3.050	26

Table 147: ADASIGNSGD sweep: 350M, 2× budget (scaling-sensitive).

β	ε	ϵ_p	η	clip	WU	WD	Val. loss	Run
0.95	1e-10	1e-4	5e-4	0	2000	0.5	**3.001**	1
0.98	1e-10	1e-4	5e-4	0	2000	0.5	3.055	2
0.95	1e-10	1e-4	**2.5e-4**	0	2000	0.5	3.002	3

Table 148: ADASIGNSGD sweep: 350M, 4× budget (scaling-sensitive).

β	ε	ϵ_p	η	clip	WU	WD	Val. loss	Run
0.95	1e-10	1e-4	5e-4	0	2000	0.5	2.967	1
0.98	1e-10	1e-4	5e-4	0	2000	0.5	3.096	2
0.95	1e-10	1e-4	**2.5e-4**	0	2000	0.5	**2.959**	3

Table 149: ADASIGNSGD sweep: 350M, 8× budget (scaling-sensitive).

β	ε	ϵ_p	η	clip	WU	WD	Val. loss	Run
0.95	1e-10	1e-4	5e-4	0	2000	0.5	2.935	1
0.98	1e-10	1e-4	5e-4	0	2000	0.5	3.101	2
0.95	1e-10	1e-4	**2.5e-4**	0	2000	0.5	**2.931**	3

Table 150: ADASIGNSGD configuration: 1B, 1× budget.

β	ε	ϵ_p	η	clip	WU	WD	Val. loss	Run
0.95	1e-8	1e-4	2.5e-4	0	2000	0.5	**2.830**	1

J.9 ADASIGNSGD (UNCOMPRESSED)

Table 151: ADASIGNSGD (uncompressed) sweep: 60M, 1× budget.

β	ε	η	clip	WU	WD	Val. loss	Run
0.87	1e-20	0.002	0	1000	0	3.595	1
0.90	1e-20	0.002	0	1000	0	3.591	2
0.95	1e-20	0.002	0	1000	0	3.597	3
0.98	1e-20	0.002	0	1000	0	3.731	4
0.90	**1e-20**	0.002	0	1000	0	3.591	5
0.90	**1e-10**	0.002	0	1000	0	3.593	6
0.90	**1e-8**	0.002	0	1000	0	3.592	7
0.90	1e-20	**5e-4**	0	1000	0	3.779	8
0.90	1e-20	**0.001**	0	1000	0	3.634	9
0.90	1e-20	**0.002**	0	1000	0	3.591	10
0.90	1e-20	**0.004**	0	1000	0	3.592	11
0.90	1e-20	**0.008**	0	1000	0	3.598	12
0.90	1e-20	0.002	**0**	1000	0	3.591	13
0.90	1e-20	0.002	**0.5**	1000	0	3.592	14
0.90	1e-20	0.002	**1.0**	1000	0	3.590	15
0.90	1e-20	0.002	**2.0**	1000	0	3.592	16
0.90	1e-20	0.002	0	**1000**	0	3.591	17
0.90	1e-20	0.002	0	**2000**	0	**3.582**	18
0.90	1e-20	0.002	0	2000	**0**	3.582	19
0.90	1e-20	0.002	0	2000	**0.1**	3.582	20
0.90	1e-20	0.002	0	2000	**0.5**	3.582	21

Table 152: ADASIGNSGD (uncompressed) sweep: 60M, 2× budget.

β	ε	η	clip	WU	WD	Val. loss	Run
0.87	1e-20	0.002	0	1000	0	3.542	1
0.90	1e-20	0.002	0	1000	0	3.536	2
0.95	1e-20	0.002	0	1000	0	3.535	3
0.98	1e-20	0.002	0	1000	0	3.627	4
0.90	**1e-8**	0.002	0	1000	0	3.536	5
0.90	**1e-10**	0.002	0	1000	0	3.536	6
0.90	**1e-20**	0.002	0	1000	0	3.536	7
0.90	1e-20	**5e-4**	0	1000	0	3.668	8
0.90	1e-20	**0.001**	0	1000	0	3.561	9
0.90	1e-20	**0.002**	0	1000	0	3.536	10
0.90	1e-20	**0.004**	0	1000	0	3.543	11
0.90	1e-20	**0.008**	0	1000	0	3.546	12
0.90	1e-20	0.002	**0**	1000	0	3.536	13
0.90	1e-20	0.002	**0.5**	1000	0	3.536	14
0.90	1e-20	0.002	**1.0**	1000	0	3.534	15
0.90	1e-20	0.002	**2.0**	1000	0	3.537	16
0.90	1e-20	0.002	0	**1000**	0	3.536	17
0.90	1e-20	0.002	0	**2000**	0	**3.526**	18

Table 153: ADASIGNSGD (uncompressed) sweep: 60M, 4× budget.

β	ε	η	clip	WU	WD	Val. loss	Run
0.87	1e-10	0.001	0	2000	0.5	3.525	1
0.90	1e-10	0.001	0	2000	0.5	3.514	2
0.95	1e-10	0.001	0	2000	0.5	3.511	3
0.98	1e-10	0.001	0	2000	0.5	3.540	4
0.95	**1e-8**	0.001	0	2000	0.5	3.511	5
0.95	**1e-10**	0.001	0	2000	0.5	3.511	6
0.95	**1e-20**	0.001	0	2000	0.5	3.511	7
0.95	1e-10	**5e-4**	0	2000	0.5	3.625	8
0.95	1e-10	**0.001**	0	2000	0.5	3.511	9
0.95	1e-10	**0.002**	0	2000	0.5	3.485	10
0.95	1e-10	**0.004**	0	2000	0.5	3.496	11
0.95	1e-10	**0.008**	0	2000	0.5	**3.455**	12
0.95	1e-10	0.008	**0**	2000	0.5	3.455	13
0.95	1e-10	0.008	**0.5**	2000	0.5	3.463	14
0.95	1e-10	0.008	**1.0**	2000	0.5	3.470	15
0.95	1e-10	0.008	**2.0**	2000	0.5	3.462	16
0.95	1e-10	0.008	0	**1000**	0.5	3.478	17
0.95	1e-10	0.008	0	**2000**	0.5	3.455	18
0.95	1e-10	0.008	0	2000	**0**	3.500	19
0.95	1e-10	0.008	0	2000	**0.1**	3.500	20
0.95	1e-10	0.008	0	2000	**0.5**	3.455	21

Table 154: ADASIGNSGD (uncompressed) sweep: 60M, 8× budget.

β	ε	η	clip	WU	WD	Val. loss	Run
0.87	1e-20	0.002	0	2000	0	3.500	1
0.90	1e-20	0.002	0	2000	0	3.481	2
0.95	1e-20	0.002	0	2000	0	3.457	3
0.98	1e-20	0.002	0	2000	0	3.476	4
0.95	**1e-20**	0.002	0	2000	0	3.457	5
0.95	**1e-10**	0.002	0	2000	0	**3.456**	6
0.95	**1e-8**	0.002	0	2000	0	3.457	7
0.95	1e-20	**5e-4**	0	2000	0	3.561	8
0.95	1e-20	**0.001**	0	2000	0	3.471	9
0.95	1e-20	**0.002**	0	2000	0	3.457	10
0.95	1e-20	**0.004**	0	2000	0	3.476	11
0.95	1e-20	**0.008**	0	2000	0	3.475	12
0.95	1e-20	0.002	**0**	2000	0	3.457	13
0.95	1e-20	0.002	**0.5**	2000	0	3.457	14
0.95	1e-20	0.002	**1.0**	2000	0	3.457	15
0.95	1e-20	0.002	**2.0**	2000	0	3.458	16
0.95	1e-20	0.002	0	**1000**	0	3.464	17
0.95	1e-20	0.002	0	**2000**	0	3.457	18
0.95	1e-20	0.002	0	2000	**0**	3.457	19
0.95	1e-20	0.002	0	2000	**0.1**	3.457	20
0.95	1e-20	0.002	0	2000	**0.5**	3.457	21

Table 155: ADASIGNSGD (uncompressed) sweep: 130M, 1× budget.

β	ε	η	clip	WU	WD	Val. loss	Run
0.87	1e-8	0.001	0	2000	0.5	3.350	1
0.90	1e-8	0.001	0	2000	0.5	3.347	2
0.95	1e-8	0.001	0	2000	0.5	3.349	3
0.98	1e-8	0.001	0	2000	0.5	3.384	4
0.95	**1e-8**	0.001	0	2000	0.5	3.349	5
0.95	**1e-10**	0.001	0	2000	0.5	3.349	6
0.95	**1e-20**	0.001	0	2000	0.5	3.348	7
0.95	1e-8	**5e-4**	0	2000	0.5	3.440	8
0.95	1e-8	**0.001**	0	2000	0.5	3.349	9
0.95	1e-8	**0.002**	0	2000	0.5	**3.334**	10
0.95	1e-8	**0.004**	0	2000	0.5	3.349	11
0.95	1e-8	**0.008**	0	2000	0.5	3.361	12

Table 156: ADASIGNSGD (uncompressed) sweep: 130M, 2× budget (scaling-sensitive).

β	ε	η	clip	WU	WD	Val. loss	Run
0.95	1e-8	0.002	0	2000	0.5	**3.284**	1
0.90	1e-8	0.002	0	2000	0.5	3.294	2
0.95	1e-8	**0.008**	0	2000	0.5	3.389	3
0.95	1e-8	0.002	0	2000	**0**	3.284	4

Table 157: ADASIGNSGD (uncompressed) sweep: 130M, 4× budget (scaling-sensitive).

β	ε	η	clip	WU	WD	Val. loss	Run
0.95	1e-8	0.002	0	2000	0.5	**3.249**	1
0.90	1e-8	0.002	0	2000	0.5	3.275	2
0.95	1e-8	**0.008**	0	2000	0.5	3.346	3
0.95	1e-8	0.002	0	2000	**0**	3.249	4

Table 158: ADASIGNSGD (uncompressed) sweep: 130M, 8× budget (scaling-sensitive).

β	ε	η	clip	WU	WD	Val. loss	Run
0.95	1e-8	0.002	0	2000	0.5	**3.223**	1
0.90	1e-8	0.002	0	2000	0.5	3.277	2
0.95	1e-8	**0.008**	0	2000	0.5	3.313	3
0.95	1e-8	0.002	0	2000	**0**	3.223	4

Table 159: ADASIGNSGD (uncompressed) sweep: 350M, 1× budget.

β	ε	η	clip	WU	WD	Val. loss	Run
0.87	1e-20	0.001	0	1000	0	3.100	1
0.90	1e-20	0.001	0	1000	0	3.089	2
0.95	1e-20	0.001	0	1000	0	3.089	3
0.98	1e-20	0.001	0	1000	0	3.200	4
0.95	**1e-8**	0.001	0	1000	0	3.090	5
0.95	**1e-10**	0.001	0	1000	0	3.089	6
0.95	**1e-20**	0.001	0	1000	0	3.089	7
0.95	1e-20	**5e-4**	0	1000	0	3.097	8
0.95	1e-20	**0.001**	0	1000	0	3.089	9
0.95	1e-20	**0.002**	0	1000	0	3.230	10
0.95	1e-20	**0.004**	0	1000	0	4.661	11
0.95	1e-20	**0.008**	0	1000	0	5.257	12
0.95	1e-20	0.001	**0**	1000	0	3.089	13
0.95	1e-20	0.001	**0.5**	1000	0	3.082	14
0.95	1e-20	0.001	**1.0**	1000	0	3.084	15
0.95	1e-20	0.001	**2.0**	1000	0	3.086	16
0.95	1e-20	0.001	0.5	**1000**	0	3.082	17
0.95	1e-20	0.001	0.5	**2000**	0	**3.073**	18
0.95	1e-20	0.001	0.5	2000	**0**	3.073	19
0.95	1e-20	0.001	0.5	2000	**0.1**	3.073	20
0.95	1e-20	0.001	0.5	2000	**0.5**	3.073	21

Table 160: ADASIGNSGD (uncompressed) sweep: 350M, 2× budget (scaling-sensitive).

β	ε	η	clip	WU	WD	Val. loss	Run
0.95	1e-20	**0.001**	0.5	2000	0	**3.023**	1
0.95	1e-20	**0.002**	0.5	2000	0	3.220	2
0.95	1e-20	0.001	**0**	2000	0	3.029	3
0.95	1e-20	0.001	0.5	2000	**0.5**	3.023	4

Table 161: ADASIGNSGD (uncompressed) sweep: 350M, 4× budget (scaling-sensitive).

β	ε	η	clip	WU	WD	Val. loss	Run
0.95	1e-20	**0.001**	0.5	2000	0	**2.984**	1
0.95	1e-20	**0.002**	0.5	2000	0	3.327	2
0.95	1e-20	0.001	**0**	2000	0	2.990	3
0.95	1e-20	0.001	0.5	2000	**0.5**	2.984	4

Table 162: ADASIGNSGD (uncompressed) sweep: 350M, 8× budget (scaling-sensitive).

β	ε	η	clip	WU	WD	Val. loss	Run
0.95	1e-20	**0.001**	0.5	2000	0	**2.965**	1
0.95	1e-20	**0.002**	0.5	2000	0	3.543	2
0.95	1e-20	0.001	**0**	2000	0	2.971	3

Table 163: ADASIGNSGD (uncompressed) configuration: 1B, 1× budget.

β	ε	η	clip	WU	WD	Val. loss	Run
0.95	1e-20	5e-4	0.5	2000	0.5	**2.882**	1