Normalized Fixed-Stride Descent: Divergence-Free Optimization and a Learning-Rate Calibration Error in a Frozen Readout Instrument

A Preprint

Caitlyn Meeks
AuxiLab Tenerife — research@auxi.cafe

August 11, 2026

registered 2026-08-11 · internal run log #8–#9 · 2 seeds × 2 feature sets, every registered inequality · replicated before publication · promoted 2026-08-11

Abstract

This study originated in an open-ended design brief. The resulting method — normalized fixed-stride descent (NFSD): a unit-norm momentum direction, a fixed step length, and each candidate step accepted or rejected on the current minibatch — is benchmarked here against the research program's frozen training instrument on two feature sets. Three results, all replicated on a second seed, plus one single-seed observation recorded but *not* promoted (the 8.6-bpc unvoided divergence in (1)). (1) NFSD **never triggered the divergence tripwire**: zero divergence-tripwire events across four step sizes spanning three orders of magnitude, 2 seeds × 2 feature sets, in a sweep where the reference Adam configuration at η=0.1 voided three times out of four and, in the fourth cell, diverged to 8.6 bits per character (bpc) *without tripping the tripwire at all*. (2) NFSD exposes a calibration error in the instrument: the reference logistic readout's fixed learning rate (Adam, 1e-3) is off-optimum in a conditioning-dependent way — by 0.017–0.018 bits on the ill-conditioned feature set, 0.013–0.014 on the well-conditioned one — so every cross-feature-set comparison of levels the program publishes now carries a ~0.02-bit error bar. (3) The step-acceptance test is dead weight: probing each candidate step before accepting it is **harmful** at large stride in 4/4 cells, though one of the four margins (+0.0038) sits inside the ±0.005 tie band. The safety comes from normalization, not from step probing.

Keywords stochastic optimization · normalized gradient methods · fixed step size · momentum · Adam · learning-rate calibration · pre-registration

§1 Motivation

This study originated in an open-ended design brief. Underneath the brief sits a question the research program needed answered: every representation the program scores passes through one frozen instrument — a logistic readout trained by Adam [1] at learning rate 1e-3. That instrument's validation loss is treated as *the* value of a representation. Is its learning rate well set on every feature set it is asked to score? And what does it cost to train with an optimizer whose step length is bounded by construction?

Divergence has a precise operational meaning here. The program applies a *divergence tripwire*: a fit-invalidating rule that voids any run whose validation loss worsens by more than 0.5 bpc in a single epoch. In this study that rule is the primary measurement.

---

§2 Method

Normalized fixed-stride descent (NFSD) is an optimizer for the program's reference logistic readout (27-way softmax, batch 8192, gradient clip 1.0, exponentially weighted moving-average momentum [2] β=0.9, validation early stopping). Take the momentum direction, normalize it to unit norm, and step a *fixed stride* s. Before a step is accepted, the candidate is evaluated on the current minibatch; if it fails to hold the minibatch loss within a 0.001-nat slack of its pre-step value, the stride is halved and the candidate re-evaluated (up to 3 halvings, after which the step is refused and the persistent stride is itself halved, floored at η/1024); 25 consecutive accepted steps without backtracking grow the persistent stride by 1.25×, capped at 4η. "Fixed stride" is therefore exact for the probe-free ablation below and *nominal* for the full machine, whose persistent stride adapts within [η/1024, 4η] and was observed to reach 3.81η at η=0.1. The full machine is NFSD with minibatch step acceptance.

```
# unit-norm direction: fixed step, independent of gradient magnitude
d = m / ‖m‖
# candidate step evaluated on the current minibatch before acceptance
W ← W − s·d
```

Baselines: the reference Adam configuration (the instrument's own optimizer class, verbatim) and momentum SGD, all on identical features and identical seeded batch order, over the program's standard text8 splits [4]. Two ablations isolate the two components: `noprobe` (probe-free NFSD — normalized fixed stride only, carrying the same single momentum buffer as every other arm but with no step acceptance test and no stride adaptation, so its stride is literally fixed at η) and `nonorm` (unnormalized probed descent — step acceptance without normalization). Two feature sets are used. The **well-conditioned** set is the program's reference sparse-tanh echo state network (ESN) features [3] at their best operating point (spectral radius ρ=0.8); the **ill-conditioned** set is the addressed-memory ESN's *linear-activation* variant at ρ=0.95 (permutation recurrence with Hadamard-coded inputs) — the worst-conditioned features in the program's inventory (see whitepaper permutation-recurrence ESN whitepaper for that construction). The three optimizer arms were swept over four step sizes spanning three orders of magnitude, η ∈ {1e-4, 1e-3, 1e-2, 1e-1}; the two ablation arms were run at η ∈ {1e-3, 1e-1} only. 24 sweep + 8 ablation = 32 fits per seed. This restriction matters when the probe-free ablation is quoted as an instrument elsewhere, and is revisited in §5.

---

§3 Pre-registered predictions

Six ordinal predictions (G1–G6) and four quantitative bands were registered before the run. The program registers *named alternatives* — each prediction is recorded together with the reading that would follow from its failure — and that discipline matters here, because the principal finding arrived through the misses.

- **G1**, the headline: NFSD voids in 0/8 sweep cells while at least one baseline voids. (Hit.)

- **G2**, a null expected to hold: NFSD should *not* beat the instrument's operating point by more than 0.01 bits. Named alternative, written in advance: if it does, the instrument's learning rate is off-optimum — provisional, replication automatically queued. (The alternative fired on the ill-conditioned set.)
- **G3**: the cost of the bounded step is small — per feature set, best-over-η NFSD ≤ best-over-η Adam + 0.05 bits. (Hit, with room to spare: the tax measured ≤ 0 on both feature sets.)
- **G4**: the optimizer ordering Adam < NFSD < SGD transfers across feature sets. (Miss on both, informatively: on the well-conditioned set the Adam–NFSD leg tied inside the ±0.005 band and inverted (Δ 0.0017 in NFSD's favour), and on the ill-conditioned set optimizer choice at each method's best η stopped mattering at all.)
- **G5**: mechanism claims — where NFSD's step refusals fire and where its worst cell sits. (Two misses: refusals were essentially absent — 3 in 9,680 steps in one cell, zero in every other — so the mechanism operated one level up, in stride-halving backtracks; and on the ill-conditioned set the worst cell was the *largest* step, not the smallest, which is thrashing rather than the predicted undertraining.)
- **G6**: step acceptance should earn its keep against `noprobe`. Named alternative: step probing is redundant, and the safety comes from normalization rather than from step probing. (The alternative fired, strongly.)

The rerun registration (seed 1) then promoted the surprises to registered claims R1–R4 with a promotion rule stated before running: R1 (ill-conditioned set: NFSD beats the instrument's learning rate by >0.01) *and* R2 (well-conditioned set: `noprobe`@0.1 beats the instrument's learning rate) both holding on the fresh seed would promote the finding — that is, enter it in the program's confirmed-findings record; any split result would be recorded without promotion.

---

§4 Results

Table 1. Seed 0, logistic validation bpc (lower is better). VOID = the divergence tripwire fired. Reference instrument setting = adam@1e-3 (bold row).

η	well-cond. Adam	well-cond. SGD	well-cond. NFSD	ill-cond. Adam	ill-cond. SGD	ill-cond. NFSD
1e-4	2.8319	3.7530	3.1719	3.1650	3.4968	3.1979
1e-3	**2.6519**	**3.2842**	**2.8184**	**3.1830**	**3.2267**	**3.1647**
1e-2	2.6577	3.0634	2.6502	3.3763	3.1704	3.1907
1e-1	**VOID**	2.8064	2.6558	**VOID**	3.1693	3.3473

Ablations at η = 1e-3 / 1e-1 — well-conditioned set: `noprobe` 3.0165 / **2.6384**, `nonorm` 3.1516 / 2.6937; ill-conditioned set: `noprobe` 3.1749 / 3.2441, `nonorm` 3.1828 / 3.2035. As a consistency check, the verbatim adam@1e-3 arm reproduced the archived instrument values (2.6519 well-conditioned, 3.1830 ill-conditioned) to all four decimals.

Reading the grid: NFSD voided nowhere and never exceeded 3.3473 bpc, whereas Adam at η=0.1 tripped the divergence tripwire on *both* feature sets. On the ill-conditioned set, NFSD@1e-3 (3.1647) beat the reference instrument setting (3.1830) by 0.0183 — firing G2's named alternative — while tuned Adam at 1e-4 reached 3.1650, so NFSD *ties tuned Adam* to within 0.0003; the discrepancy is with the instrument's fixed learning rate, not with Adam's attainable optimum. On the well-conditioned set, `noprobe`@0.1 (2.6384) was the best cell in the entire grid, 0.0135 below the reference instrument setting. Step acceptance was worse than useless at large stride: full NFSD lost to its own `noprobe` ablation by +0.0174 (well-conditioned) and +0.1032 (ill-conditioned). Step refusal — the terminal path, reached only after three halvings all fail — fired 3 times in 9,680 steps in one cell (well-conditioned, η=0.1, seed 1) and never in the other three η=0.1 cells. A single halving sufficed for the large majority of backtracking steps but not all: on the well-conditioned set at seed 0, 203 of 2,619 backtracking steps needed two or three. Backtrack rates at η=0.1 were 47% well-conditioned and 80% ill-conditioned at seed 0, rising to 69% and 89% at seed 1.

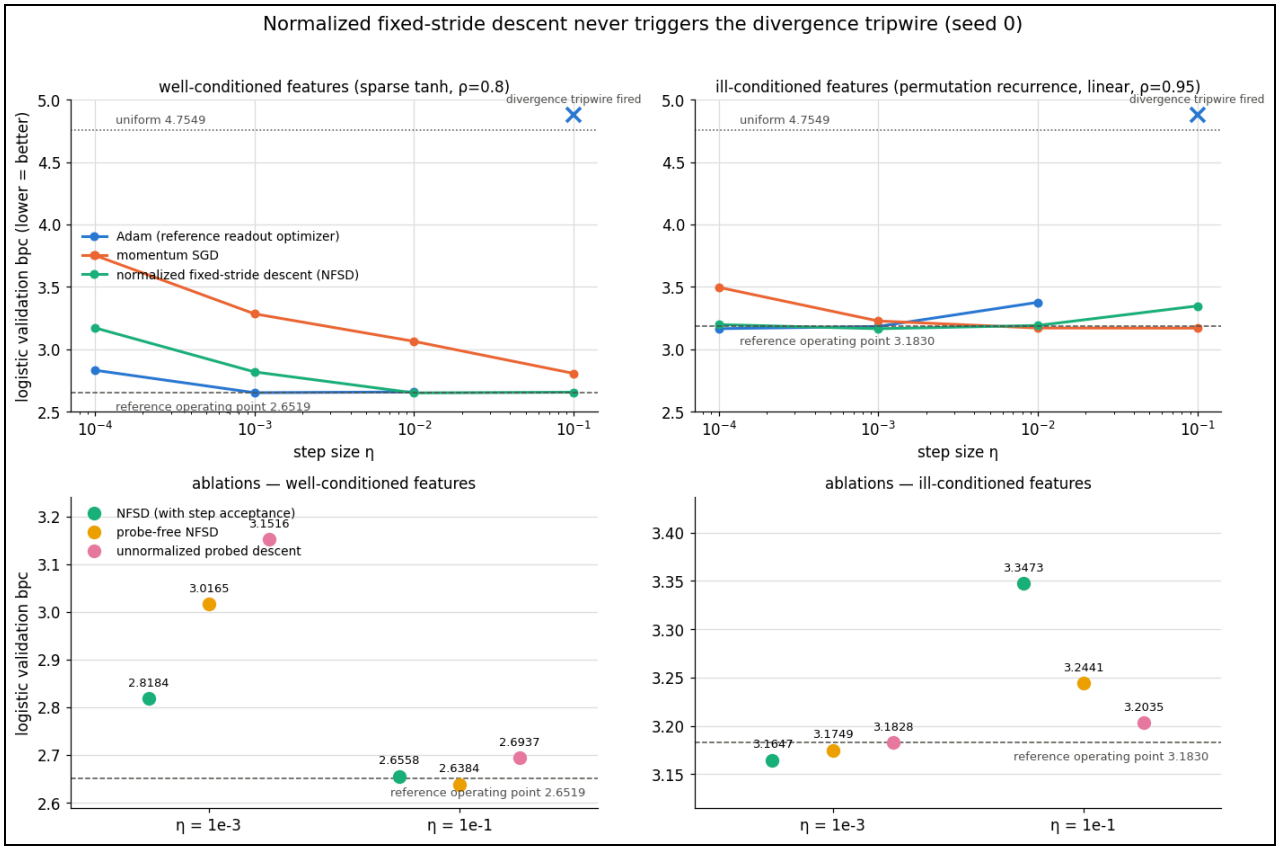

Figure 1. Seed 0 summary (internal run log #8). Top row: logistic validation bpc vs step size η (log scale) for the reference Adam configuration, momentum SGD, and NFSD on the well-conditioned (left) and ill-conditioned (right) feature sets; × marks a divergence-tripwire event (reference Adam at η=0.1, both feature sets); dashed lines: the reference instrument anchors 2.6519 / 3.1830; dotted: the uniform-prediction baseline 4.755. Bottom row: ablations at η = 1e-3 and 1e-1 — full NFSD vs `noprobe` vs `nonorm`; the well-conditioned `noprobe`@0.1 (2.638) is the best cell in the grid, below the reference anchor.

Table 2. The registered replication, seed 1 (internal run log #9): 4/4 ordinals, 2/2 bands.

claim	seed 0	seed 1
R1 ill-cond.: best NFSD vs instrument setting	3.1647 vs 3.1830 (Δ 0.0183)	3.1651 vs 3.1823 (Δ 0.0172)

R2 well-cond.: noprobe@0.1 vs instrument setting	2.6384 vs 2.6519 (Δ 0.0135)	2.6325 vs 2.6467 (Δ 0.0142)
R3 step probing harmful at η=0.1 (both feature sets)	+0.0174 well-cond. · +0.1032 ill-cond.	+0.0038 well-cond. (thin) · +0.095 ill-cond.
R4 NFSD voids in sweep	0/8; adam@0.1 voided ×2	0/8; adam@0.1 voided (well-cond.), 8.5978 bpc unvoided (ill-cond.)

Two disclosures from the rerun, recorded as registered. The well-conditioned margin in R3 (0.0038) is thin — inside the tie band other claims in this program use; it was judged as registered (a strict inequality) and flagged in the run journal. Second, seed 1 produced a new single-seed observation, *not* promoted: on the ill-conditioned set, Adam at η=0.1 diverged to 8.5978 bpc — nearly double the uniform-prediction baseline (4.7549) — without ever tripping the divergence tripwire, because the loss oscillated rather than rising monotonically and no single epoch's increase reached the 0.5-bpc bar (the largest was 0.4656). The tripwire detects monotone blow-up, not sustained oscillation. Across both seeds, the reference Adam configuration at η=0.1 therefore voided in 3/4 cells, with unvoided divergence in the 4th. NFSD: 0 voids in all 16 sweep cells.

§5 Discussion

First, the instrument finding, which is the principal result. The program's frozen readout instrument is mis-calibrated in a conditioning-dependent way: its fixed learning rate is off-optimum by 0.017–0.018 bits on the ill-conditioned feature set (recovered by adam@1e-4 or a normalized fixed stride at 1e-3) and by 0.013–0.014 bits on the well-conditioned feature set (best cell in the grid: normalized fixed stride 0.1). The program action, entered in the program's confirmed-findings record, is that cross-feature-set comparisons of logistic *levels* carry a ~0.02-bit learning-rate × conditioning error bar. The instrument nevertheless stays frozen — comparability across the existing archive outranks 0.02 bits — and separate evidence (internal run log #5) indicates that matched-parametrization *gaps* remain robust even where levels are biased.

Second, the optimizer — stated separately for the two variants, because they were not run over the same grid and an earlier wording of this program's summary conflated them. *Full NFSD*, with step acceptance, was swept over all four step sizes (1e-4 to 1e-1, three orders of magnitude) on both seeds and both feature sets: it needs no per-parameter *adaptive learning rates* — only the single momentum buffer every arm here carries — never triggered the divergence tripwire in any of the 16 sweep cells, and either ties tuned Adam to within 0.001 or beats it (−0.0017 / −0.0104 well-conditioned, −0.0003 / +0.0002 ill-conditioned). The *probe-free* variant — normalized fixed stride with no step acceptance — was run at only two step sizes, 1e-3 and 1e-1 — 8 cells, not the four-rung sweep. Within that coverage it never voided either, 8 of 8. Its tie-with-tuned-Adam tolerance is *not* 0.001: on the ill-conditioned feature set it is +0.0099 / +0.0097 worse than tuned Adam, so the correct figure for the probe-free variant is 0.010. Both variants derive their stability from the normalization — the step *direction* is unit-norm, so no step can be scaled up by a large gradient — but only the probe-free variant holds its stride literally fixed; under step acceptance the persistent stride adapts within [η/1024, 4η] and was observed at 3.81η. The probe-free variant's evidence base is narrower than the full method's, and elsewhere in this program it is relied on as a trusted instrument, which is why the distinction is drawn here explicitly.

Third, the attribution. Minibatch step acceptance — the component the design brief made most distinctive — was harmful at large stride, in all four cells and on both seeds, though the well-conditioned seed-1 margin of +0.0038 sits inside the ±0.005 tie band this program uses and so supports the sign but not the word "strictly". The safety of the method comes from normalization, not from step probing. The original run report's summary stands in substance: several registered ordinal predictions missed onto their pre-named alternatives, the quantitative bands held, and the step-acceptance component proved dispensable.

---

§5.1 Amendment: the independent audit

This study was independently audited on 2026-08-12. The audit's verdict on the principal result is unambiguous and, if anything, makes it stronger. Both off-optimum gaps re-derive exactly from the raw result files (0.0183 / 0.0172 ill-conditioned, 0.0135 / 0.0142 well-conditioned); the probe-free variant at η=0.1 really is the argmin of the entire well-conditioned grid at both seeds; and all six gaps keep their sign beyond 0.005 bits on the *untouched test split*, so none of the instrument finding is an artifact of selecting the best epoch on the same split early stopping uses. The auditor checked the reference-Adam arm and the fixed-stride arms line by line for asymmetry and found the same skeleton throughout — same zero initialization, same seeded batch order, same gradient clipping, same divergence rule, same patience, same best-epoch restore, one shared metric path. The ill-conditioned half of the instrument finding requires no fixed-stride code at all: it is the program's own logistic readout at η=1e-4 against the same readout at η=1e-3, on identical features. An independent re-measurement on a fresh hardware-drawn seed at reduced budget reproduces the headline ordinals.

The audit recorded a **concern**, and it is against the optimizer summary rather than the instrument finding: an earlier program wording described the probe-free variant while quoting the full method's numbers, making two claims that are false of the variant they were attached to — the step-size coverage and the 0.001 tie tolerance. Both are corrected in §5's second paragraph above, and the auditor's own re-measurement independently reproduced the tolerance concern. A third, minor observation — that "strictly harmful" overstates a +0.0038 margin inside the tie band — is corrected in §5's third paragraph and in the abstract. A repair run is registered in the program's queue. No number in this paper's tables changed.

Three further corrections were made at the same time, found while checking the above rather than by the audit itself. First, the claim that the step-refusal path "never once fired" was a seed-0 statement carried into a two-seed paper; on seed 1 it fired three times in 9,680 steps, and §4 now reports that, along with the backtracking steps that needed more than one halving. Second, the ablation arms were described as though they shared the main sweep's four step sizes; they were run at two, which §2 now states where a replicator will meet it. Third, "fixed stride" was stated without qualification: it is exact for the probe-free ablation but nominal for the full method, whose persistent stride adapts within [η/1024, 4η]. §2 and §5 now draw that distinction.

---

§6 References

1. D. P. Kingma, J. Ba, "Adam: a method for stochastic optimization," *ICLR* 2015 (arXiv:1412.6980).

2. B. T. Polyak, "Some methods of speeding up the convergence of iteration methods," *USSR Computational Mathematics and Mathematical Physics* 4(5), 1964.

3. H. Jaeger, "The 'echo state' approach to analysing and training recurrent neural networks," GMD Report 148, German National Research Center for Information Technology, 2001.

4. M. Mahoney, "Large text compression benchmark" (text8), mattmahoney.net/dc/textdata.

---

§7 Provenance

- **32caa262** — the original design brief and main grid, 32 fits, seed 0 (internal run log #8). Registered before running; the pilot run was disclosed in the registration, including a known pilot-scale inversion artifact against which the predictions were held.
- **8a0a29b4** — the registered replication: full 32-fit grid on seed 1, fresh reservoir draws and batch order (internal run log #9). Promotion by the pre-stated R1+R2 rule on two independent passes.

Audit status: independently audited 2026-08-12, with a recorded concern. The concern, in one line: the program's summary of this finding described the probe-free variant while carrying the full method's numbers, overstating that variant's step-size coverage and its tie tolerance against tuned Adam, plus a third, minor overstatement in the word "strictly". The instrument finding — this paper's principal result — re-derived exactly, survived on the untouched test split, and was independently re-measured. The corrections are made in §5 and §5.1 above; no table on this page changed. A repair run is registered in the program's queue.

All claims are judged strictly against the registered wording, with misses included above; 2 seeds × 2 feature sets; replicated before publication. Internal designation: goat descent.
