An absolute tolerance for the write

> The write° tolerance of *ex-2.2.1*°, which we call the relative tolerance here, counts a site as over when the write is more than a quarter over the bound°. So where the bound is a fraction of a degree, a tiny turn counts. We looked for an absolute tolerance° in degrees that the data could set, and found one in how much the bound varies from run to run: about a degree. With that absolute tolerance, every site at a symbol or separator drops out. What stays are the operands at the first slice or two, chiefly on `pool 1–3, soft τ`.

Write. The same quantile, over the states as they arrive under the full edit, after the preceding block has read the edited states of the slices below.

ex-2.2.1. Set the write bound of H4, with a tolerance of a quarter over the bound.

Bound. From the clean pass…

Absolute tolerance. An excess in degrees that a site must also pass to count….

The paper review of the *recipe-compare report*° asked whether the tolerance should ignore small shifts. In that report and the *fixed-effects report*°, every site over the relative tolerance° has a bound of a few degrees or less, so a ratio over 1.25 can be a turn of a fraction of a degree. With an absolute tolerance, a site would count only when the write is past both the ratio and the bound plus some number of degrees. The backlog item behind this report asked for that number to come from the measurement rather than be picked by eye.

recipe-compare report. `pool 1–3, soft τ` removes about as much as `falloff` and costs the task least, but its edit spills over the criterion on two seeds, and that spill enters with the edit at block 1.

fixed-effects report. On `τ_ℓ fixed` the write stays within the clean bound past the embedding, as on `falloff`, with a few dozen sites over the relative tolerance of ex-2.2.1, all where the bound is a few degrees.

Over the relative tolerance. A site where the write is more than 1.25× the bound, the tolerance of H4 in ex-2.2.1, each run taken on its own….

This report needs no new runs. It uses the stored alignment° with e_1 at every slice and position, clean and under the full edit, on the eight runs of each of four recipes: `τ_ℓ fixed` from the *fixed confirmation*°, `falloff` from the *edit-effects report*°, and `step, soft τ` and `pool 1–3, soft τ` from the *recipe-compare report*.

Alignment. How far a state points along e_1…

fixed confirmation. `τ_ℓ fixed` keeps the anchor on all eight runs at seeds 712 to 719 and is the recipe of record for D2.2.

edit-effects report. On `falloff`, the write stays within the clean bound past the embedding, measured as the ratio of the write to the bound at the 99% quantile.

Observations

- The spread from the choice of contexts (E1): drawing the held-out contexts again moves the excess° of the write° over the bound° by a few hundredths of a degree, far less than any excess past the relative tolerance. So that spread cannot set the absolute tolerance°. The excess at the sites over the relative tolerance° is a real feature of each run.

Excess. The write less the bound, in degrees….

Write. The same quantile, over the states as they arrive under the full edit, after the preceding block has read the edited states of the slices below.

Bound. From the clean pass…

Absolute tolerance. An excess in degrees that a site must also pass to count….

Over the relative tolerance. A site where the write is more than 1.25× the bound, the tolerance of H4 in ex-2.2.1, each run taken on its own….

- The spread of the bound between runs (E2): where the bound is small, it differs from run to run by about half a degree, and by about a degree at the 95% quantile. We take that degree as the absolute tolerance.
- What the absolute tolerance keeps (E3): every site at a `?`, `,` or ⏎ falls under the absolute tolerance. The sites that clear it are all at operands, at slice 1 (and slice 2 on `step, soft τ`): one to three runs of the other recipes, and every run of the pool. A rule taken site by site (an excess over twice the seed spread° at that site) flags about as many sites. But some of them are at symbols, where the bound hardly varies between runs.

Seed spread. The standard deviation of the bound at one site over the eight runs of one recipe.

Scope and method

This is an exploratory re-analysis, with no preregistration and no gate. It covers 32 runs: `τ_ℓ fixed` at seeds 712 to 719, and `falloff`, `step, soft τ` and `pool 1–3, soft τ` at seeds 704 to 711. Each run is scored on the 14,000 held-out contexts of ex-2.2.21, under the full edit, which takes the e_1 component out of the state at every position and every slice.

The bound° and the write° are computed as in the *fixed-effects report*°, and give the same counts over the relative tolerance° for its two recipes. A site is one slice past the embedding and one position, 96 per run. At the embedding the write equals the bound by construction, so it is left out.

We left out one source the item suggested for the absolute tolerance°: the write that an edit along a reference direction gives at the same sites. The stored arrays hold the alignment° with e_1 only, so that would need a new measurement pass.

Bound. From the clean pass…

Write. The same quantile, over the states as they arrive under the full edit, after the preceding block has read the edited states of the slices below.

***fixed-effects report*.** On `τ_ℓ fixed` the write stays within the clean bound past the embedding, as on `falloff`, with a few dozen sites over the relative tolerance of ex-2.2.1, all where the bound is a few degrees.

Over the relative tolerance. A site where the write is more than 1.25× the bound, the tolerance of H4 in ex-2.2.1, each run taken on its own.…

Absolute tolerance. An excess in degrees that a site must also pass to count.…

Alignment. How far a state points along e_1…

The measurements

Bound

: From the clean pass: arcsin of the 99% quantile of |α| over the states of the other ops at a slice and position, as an angle. It is how far the edit turns those states there. At an answer position it counts only the answers that `difference` could not give from the same operands.

Write

: The same quantile, over the states as they arrive under the full edit, after the preceding block has read the edited states of the slices below.

Excess

: The write less the bound, in degrees. Positive when the blocks after the edit put more e_1 on the other ops than the clean model has there.

Over the relative tolerance

: A site where the write is more than 1.25× the bound, the tolerance of H4 in ex-2.2.1, each run taken on its own. We call this the relative tolerance.

Absolute tolerance

: An excess in degrees that a site must also pass to count. A site over both counts; one over the relative tolerance alone does not.

Seed spread

: The standard deviation of the bound at one site over the eight runs of one recipe.

The spread from the choice of contexts (E1)

The bound° and the write° are each a 99% quantile over about 12,000 states, so they would move a little on another set of contexts. If that movement were as large as the excess° at the sites over the relative tolerance°, it would be the natural absolute tolerance°: an excess within it could be an accident of which contexts were held out. We drew the contexts again with replacement 200 times per run, took the bound and the write on the same draw, and measured the spread of the excess at each site.

Bound. From the clean pass…

Write. The same quantile, over the states as they arrive under the full edit, after the preceding block has read the edited states of the slices below.

Excess. The write less the bound, in degrees….

Over the relative tolerance. A site where the write is more than 1.25× the bound, the tolerance of H4 in ex-2.2.1, each run taken on its own….

Absolute tolerance. An excess in degrees that a site must also pass to count….

recipe	sites over	excess there	spread there, median	spread there, largest
`τ_ℓ fixed`	34 of 768	0.04° to 1.20°	0.01°	0.07°
`falloff`	28 of 768	0.04° to 1.72°	0.01°	0.10°
`step, soft τ`	23 of 768	0.03° to 2.21°	0.02°	0.87°
`pool 1–3, soft τ`	48 of 768	0.03° to 4.09°	0.04°	0.22°

The excess at the sites over the relative tolerance, against its spread from the choice of contexts. *The sites over count slice and position pairs past the embedding over the eight runs of each recipe, each run taken on its own. The spread is the standard deviation of the excess over 200 draws of the held-out contexts with replacement, at each of those sites.*

The spread is tiny next to the excess. At the sites over the relative tolerance it is a few hundredths of a degree at the median, and the excess there is about 25 times its own spread at the median. The closest case is an operand site with a bound of about 6°, where the spread is nearly a degree and the excess 2.5 times that. The spread is wider on `step, soft τ` and the pool, which have more of their sites over the relative tolerance at operands, where the bound is a few degrees, but it stays at hundredths of a degree at the median there too. So the excess at every site over the relative tolerance would survive another set of contexts drawn the same way.

Whether a site stays over the relative tolerance is a narrower question. The tolerance line is at an excess of a quarter of the bound, not at zero. 27 of the 133 sites are within twice their spread of that line, and another draw could put those under it.

The spread of the bound between runs (E2)

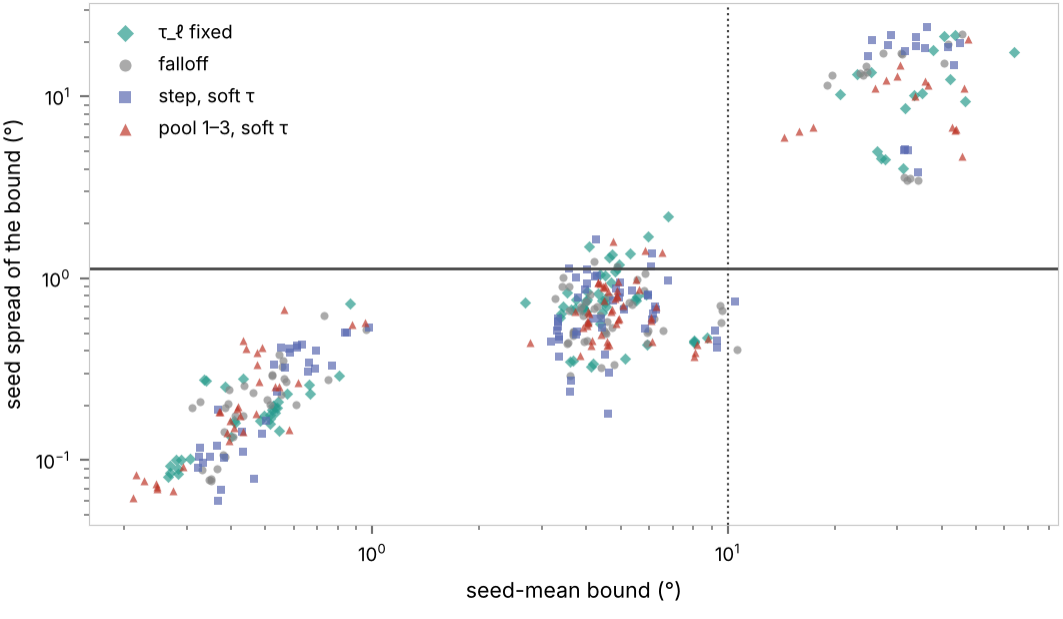
The bound° differs from one trained run to the next, by much more than it does from one set of contexts to another. An excess° smaller than that seed spread° is within what training alone moves the bound by, so the seed spread is a candidate for the absolute tolerance°. The figure shows it at every site, against how wide the bound is there.

Bound. From the clean pass…

Excess. The write less the bound, in degrees….

Seed spread. The standard deviation of the bound at one site over the eight runs of one recipe.

Absolute tolerance. An excess in degrees that a site must also pass to count….



How much the bound varies between runs, by how wide it is. *Each point is one site (a slice past the embedding and a position) of one recipe: the seed mean of the bound on the horizontal axis, and its standard deviation over the eight runs on the vertical. Both axes are logarithmic. The solid line is the absolute tolerance, the 95% quantile of the spread over the sites with a seed-mean bound under 10° (left of the dotted line).*

The spread grows with the bound, and the sites fall into three groups. At the symbols and separators the bound is under a degree, and it spreads by about 0.2° between runs. At the operands it is a few degrees, and spreads by about half a degree to a degree. At the answers it is tens of degrees and spreads by about 12°, but no site there goes over the relative tolerance°: every site that does has a bound under 6°.

Over the relative tolerance. A site where the write is more than 1.25× the bound, the tolerance of H4 in ex-2.2.1, each run taken on its own….

Over the sites with a bound under 10°, the seed spread is 0.47° at the median and 1.13° at the 95% quantile. We take the second as the absolute tolerance: an excess larger than that is more than the bound moves between runs at 95% of the small-bound sites. The operand sites set it, since the bound at every symbol and separator spreads by less.

The absolute tolerance depends a little on which sites go into the pool. The symbol sites are two fifths of it (the `=` sites, whose bound is a few degrees, are most of the rest). They are well under the absolute tolerance and pull the quantile down: over the operand sites alone it is 1.39°. Taken one recipe at a time, the median is the same on all four, and the 95% quantile runs from 0.97° to 1.35°.

The cut at 10° leaves out the answers. Their spread comes from the tail of states found in the *fixed-effects report*°, and including them would lift the absolute tolerance past anything the small-bound sites need.

fixed-effects report. On `τ_ℓ fixed` the write stays within the clean bound past the embedding, as on `falloff`, with a few dozen sites over the relative tolerance of ex-2.2.1, all where the bound is a few degrees.

What the absolute tolerance keeps (E3)

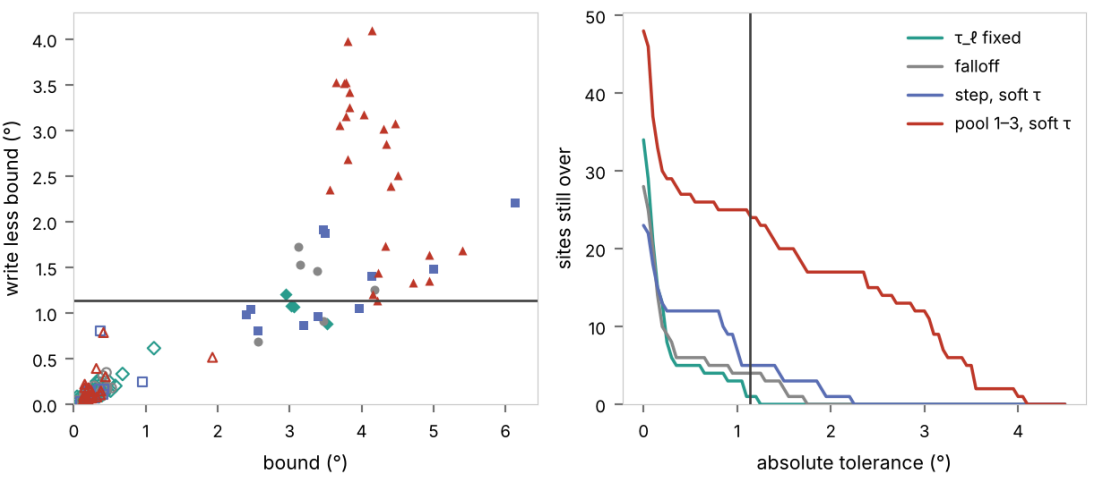
The figure places each site over the relative tolerance° by its bound° and its excess°, and counts how many stay over as the absolute tolerance° rises.

Over the relative tolerance. A site where the write is more than 1.25× the bound, the tolerance of H4 in ex-2.2.1, each run taken on its own....

Bound. From the clean pass...

Excess. The write less the bound, in degrees....

Absolute tolerance. An excess in degrees that a site must also pass to count....



***Left:** the sites over the relative tolerance, by the bound there and the excess of the write over it, on all four recipes, each run taken on its own. Hollow markers are at a `?`, `,` or ⏎, filled ones at an operand. **Right:** how many of them stay over as the absolute tolerance rises, per recipe over its eight runs. The solid line in both panels is the absolute tolerance of E2, 1.13°.*

The sites over the relative tolerance are of two kinds. The bound separates them cleanly, and the absolute tolerance does so only in part. At the symbols and separators (`?`, `,` and ⏎) the bound is under a degree at all but 2 of them and at most 1.9°, the excess is at most 0.8°, and every one of them falls under the absolute tolerance. They make up most of the sites over the relative tolerance on `τ_ℓ fixed` and `falloff` (the table below). At the operands the bound is a few degrees, and the excess ranges from 0.7° to several degrees. So the absolute tolerance takes out every symbol site, and 12 of the 46 operand sites, those with the smallest excess. No site at `=` or at an answer goes over the relative tolerance. Every site that clears the absolute tolerance is at an operand, at slice 1 or 2.

recipe	over the relative tolerance	at a symbol	over both	runs	slices	twice the seed spread	1° absolute
`τ_ℓ fixed`	34	30	1	1 of 8	1	2 (2)	3
`falloff`	28	22	4	1 of 8	1	4 (1)	4
`step, soft τ`	23	12	5	3 of 8	1, 2	5 (1)	7
`pool 1–3, soft τ`	48	23	24	8 of 8	1	24 (3)	25

***The sites over the relative tolerance, and those over the absolute tolerance as well**, per recipe over its eight runs, each run taken on its own. "At a symbol" counts those at a `?`, `,` or ⏎. "Over both" counts those whose excess is also over the absolute tolerance of E2 (1.13°), with the runs and slices they fall on. The last two columns give the count over both under two other rules: an excess over twice the seed spread at that site (in brackets, how many of those are at a symbol), and an absolute tolerance of a whole degree.*

On `τ_ℓ fixed`, `falloff` and `step, soft τ` a few operand sites stay, on one to three runs each. On `pool 1–3, soft τ` every run keeps some, 24 sites in all, with a median excess of 2.9° on bounds of about 4°. At those sites, block 1 puts e_1 back on the operands of the other ops, up to about twice what the clean model has there. That is the slice where the *recipe-compare report*° found the pool spilling, and the one slice the pool pulls and the step leaves alone.

recipe-compare report. `pool 1–3, soft τ` removes about as much as `falloff` and costs the task least, but its edit spills over the criterion on two seeds, and that spill enters with the edit at block 1.

An absolute tolerance of one whole degree keeps the same kind of sites, all at operands, with a few more of them: two more on `τ_ℓ fixed` (whose excess at the operands of one run is just over a degree), two more on `step, soft τ`, and one more on the pool. The count on the recipe of record shifts when the absolute tolerance moves by a few tenths of a degree, but the split between the two kinds does not.

A rule taken site by site counts a site when the excess there is over twice the seed spread° of the bound at that site. It keeps 21 of the 24 pool sites that the absolute tolerance keeps, and about as many sites as the absolute tolerance on the other recipes, though on `τ_ℓ fixed` the two rules share none.

Seed spread. The standard deviation of the bound at one site over the eight runs of one recipe.

It also keeps 7 sites at symbols, with excesses of up to 0.8°. The bound at a symbol spreads by only about 0.2° between runs, so a small excess there is large for that site. In turn, the rule drops some operand sites, where the spread is wider. The two rules ask different questions: whether the turn is large in degrees, or unusual for that site.

Discussion

The tolerance of *ex-2.2.1*°, taken as a ratio alone, mixes two kinds of site: symbols and separators, where the clean model puts almost no e1 and the write° adds a fraction of a degree, and operands at the first slice or two, where the write adds a degree or more to a bound° of a few degrees. An absolute tolerance° of about a degree removes the first kind and keeps most of the second.

ex-2.2.1. Set the write bound of H4, with a tolerance of a quarter over the bound.

Write. The same quantile, over the states as they arrive under the full edit, after the preceding block has read the edited states of the slices below.

Bound. From the clean pass…

Absolute tolerance. An excess in degrees that a site must also pass to count….

We are fairly sure of the split under an absolute tolerance. It holds for any absolute tolerance above the largest excess° at a symbol, 0.8°, and the two kinds are apart in bound on the left of the E3 figure, though their excesses overlap a little below the absolute tolerance. A rule relative to each site would draw the line elsewhere, flagging a few symbol sites whose small excess is unusual for that site. Which of the two a gate wants is a choice about what the relative tolerance is for.

Excess. The write less the bound, in degrees….

We are less sure of the exact absolute tolerance. It rests on eight runs per recipe, and on the choice of the 95% quantile and the cut at 10°. It also rests on taking the quantile over the symbol and operand sites together. The operand sites alone would put it a few tenths of a degree higher (E2), and a few more sites would fall under it. A reference direction, edited at the same sites, would give a second scale for it that this report could not measure from stored data.

If the absolute tolerance were adopted, the reports that quote the relative tolerance would count one site over both on the recipe of record and 4 on `falloff`, all at operands of a single run on each recipe. They would also count sites on every run of `pool 1–3, soft τ`. The pool would then stand apart on the write as it already does on spill°, at the same slice.

Spill. The side effect of an edit on the other ops…

References

recipe-compare report

`pool 1–3, soft τ` removes about as much as `falloff` and costs the task least, but its edit spills over the criterion on two seeds, and that spill enters with the edit at block 1.

fixed-effects report

On `τ_ℓ fixed` the write stays within the clean bound past the embedding, as on `falloff`, with a few dozen sites over the relative tolerance of ex-2.2.1, all where the bound is a few degrees.

fixed confirmation

`τ_ℓ fixed` keeps the anchor on all eight runs at seeds 712 to 719 and is the recipe of record for D2.2.

edit-effects report

On `falloff`, the write stays within the clean bound past the embedding, measured as the ratio of the write to the bound at the 99% quantile.

ex-2.2.1

Set the write bound of H4, with a tolerance of a quarter over the bound.