A baseline from the quiet axes

> Removal and spill° have been measured net of what the same edit does to the control at the same seed, so they depend on where the anchored axis, e_1, happens to sit in that control. We re-scored the stored runs of the recipe of record against a baseline that does not depend on it: the median drop over edits along the other quiet coordinate axes. The numbers quoted so far barely move, and on the rotated-basis runs the new baseline brings the short seeds back up beside the rest.

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

The question comes from rotated-basis, where Adam acted in a rotated basis.[1] On two seeds e_1 overlapped the syntax direction of the control, so the edit took half of the op from the control as well. Removal net of the control then fell short, although the anchored runs removed the op as before.

1 Adam and the basis…

That report and the D2.2 design proposed a different baseline: the median over reference edits along other coordinate axes, leaving out the syntax axes, as in the spill-timing experiment.

This report trains nothing. It scores stored runs: the eight `falloff` runs (the recipe of record) of split-anti-seeds, the eight controls of ex-2.2.23 at the same seeds, the sixteen runs of rotated-basis, and, for E4, the sixteen runs of two soft-pool arms. On each run it scores the edit along each of the 64 coordinate axes, at every dose.

Observations

- Where e_1 sits in the control (E1): it varies by seed. On some seeds e_1 is among the quietest axes of the control, and on one it is among the five costliest. The median over the reference axes is steady across seeds. Separately, on three seeds the anchored run keeps its spike of energy on a different coordinate from the control.
- Net removal and spill against the new baseline (E2): little changes. The mean removal rises by a hundredth or two, mostly on the one seed where e_1 was a busy axis of the control. Every run stays within the spill° criterion. The median taken in the anchored run itself, with no control, agrees with the one taken in the control.

  Spill. The side effect of an edit on the other ops…
- The rotated-basis runs (E3): the new baseline, taken along the rotated axes, brings the removal on the two seeds below half, and a third at six tenths, up to the level of the rest.
- Runs near the spill criterion (E4): on two arms of soft-pool that come close to the criterion, the median moves runs across it. `step, soft τ` goes from 2 runs over to none, and `pool 1–3, soft τ` keeps about as many, but not the same ones.

Scope and method

This is an exploratory re-scoring, with no preregistration and no gate. The model seeds are 704 to 711, and the controls are the 400-epoch ones. Each anchored run pairs with the control at its seed, which shares its initialization, batches, and label draws.

The edit is the dose edit of ex-2.2.22, unchanged. It projects a fixed direction out of the state at every position and slice, removing a quarter, a half, three quarters, or all of the component along it. It is scored on the whole held-out set of every op, as before.

Along e_1 the drops this edit causes reproduce the published dose scoring to rounding. New here is the same edit along every other coordinate axis, and, for the rotated-basis runs, along every axis of the rotated basis Adam acted in.

A reference axis is any coordinate axis other than e_1 and the two that hold the most energy[2] of the stream, averaged over slices. Those two are where the model puts its syntax tokens. For the rotated-basis runs the reference axes are the rotated axes, with the two highest in energy left out the same way.

2 Energy…

One coordinate holds about 0.4 of the energy in every control, and E1 shows what projecting it out costs.

Measurements

- Removal: the net drop in score on `difference` under the full edit along e_1, as a share of the way from the clean score to the target null°. Higher is better.
- Spill°: the largest net drop on another op under the e_1 edit, over the four doses. A run is within the criterion when it is at most 0.02. Lower is better.
- The cost of an axis: the drop under the full edit along it, averaged over the ops. A quiet axis is one that costs little.

Target null. Where a removal should land…

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

Each of these is net of a baseline. We compare three, each taken per op and per dose:

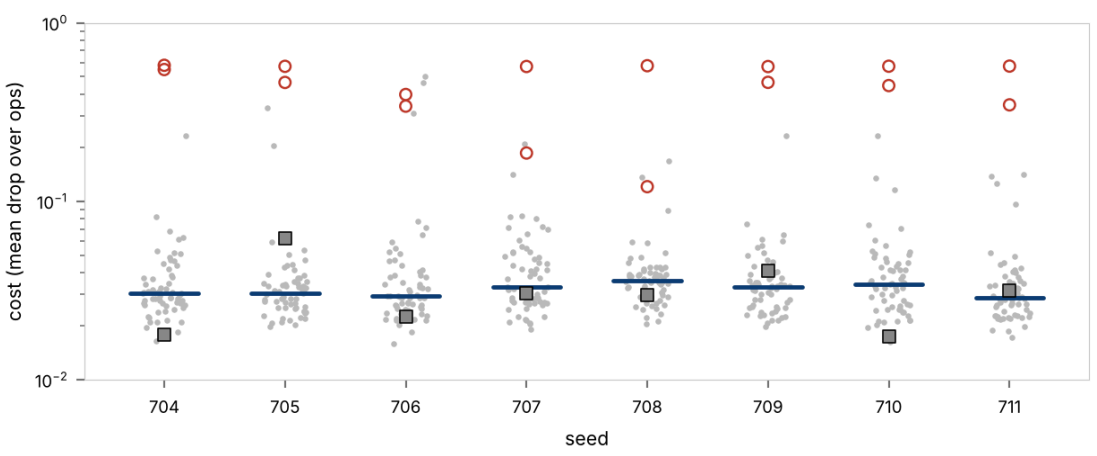
- e_1 in the control: the drop the same e_1 edit causes in the control at the same seed. This is the measurement quoted so far.
- Reference axes, control: the median drop over the reference axes of the control.
- Reference axes, own run: the median drop over the reference axes of the anchored run itself. It needs no control.

E2 also shows removal with nothing subtracted (raw), as rotated-basis recommended.

Where e_1 sits in the control (E1)

How typical an axis is e₁ in a control? If it is always among the quiet ones, the paired baseline and the median of the quiet axes should agree, and nothing quoted so far would move. If its place varies by seed, so does the paired baseline.

What we saw. Its place varies. On two seeds e₁ is the second quietest axis of the control. At seed 705 it is the 5th costliest of 64 (the high square in the figure below), with about twice the cost of the median reference axis. The median itself is steady across seeds.



***The cost of each coordinate axis in the control**, the full edit along it, averaged over the ops. One column per seed: the reference axes as faded dots, their median as a bar, the two syntax axes as open circles, and e₁ as a square. The axis is logarithmic.*

seed	e₁ rank by cost, control	e₁ cost, control	reference median, control	syntax axes cost, control	top energy share, control	top axis, `falloff`
704	63 of 64	0.018	0.030	0.55, 0.58	0.38	same
705	5 of 64	0.062	0.030	0.46, 0.57	0.36	same
706	55 of 64	0.022	0.029	0.40, 0.34	0.36	moved
707	38 of 64	0.031	0.033	0.57, 0.19	0.44	moved
708	48 of 64	0.030	0.036	0.58, 0.12	0.45	moved
709	15 of 64	0.041	0.033	0.46, 0.57	0.37	same
710	63 of 64	0.018	0.034	0.45, 0.57	0.38	same
711	27 of 64	0.032	0.029	0.35, 0.57	0.36	same

***Where e₁ sits in the control**, per seed. The rank counts from the costliest of the 64 coordinate axes. The cost is the drop under the full edit, averaged over the ops. The top energy share is the share of the energy of the stream held by the axis with the most, averaged over slices, and the last column says whether the `falloff` run at that seed has its top axis on the same coordinate.*

So the paired baseline is one draw from the axes of the control, and that draw ranges from one of the quietest axes to one of the busiest few.

The axis with the most energy takes most of the score of every op on every seed, and all of it on 707 and 708.

The second (the lower open circle) costs as much on most seeds, but on 707 and 708 it is a milder axis, and on 706 two reference axes cost more than it does. So ranking by energy only roughly finds the costly axes.

Leaving the syntax axes out of the reference set hardly matters: the median over all 63 axes other than e₁ gives a removal within 0.005 of the median without them, on every seed. A median of a few dozen axes is not swayed by two.

On 3 of 8 seeds (706, 707, 708) the `falloff` run puts its most energetic axis (the spike) on a different coordinate from the control at the same seed, at every slice, and that coordinate holds almost nothing in the control. Rotated-basis expected otherwise: it found the spike on the same coordinate in both runs on the seeds it checked, and took a move under the rotated basis as a sign that the anti term steers it.

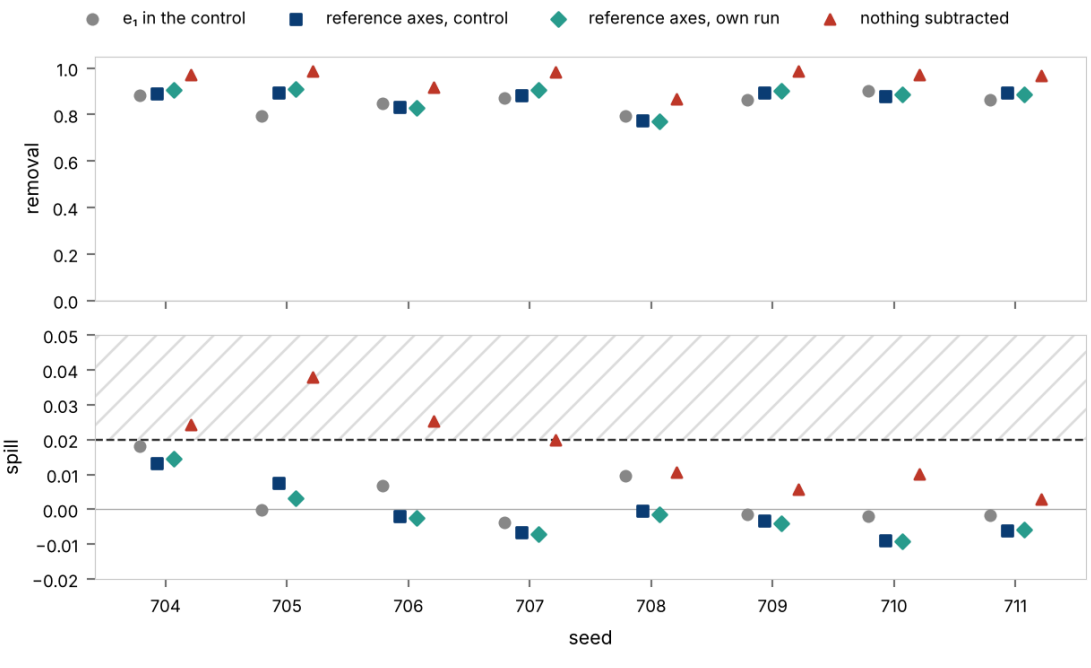
The moves seen here under plain Adam suggest the spike can move without the rotation, so a change of loss alone may be enough to move it.

Net removal and spill against the new baseline (E2)

When netted against the median of the reference axes, do the removal and spill° of `falloff` change from the numbers that split-anti-seeds quoted? The main measurement is removal at the full dose per seed, shown beside spill over the four doses.

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

What we saw. Hardly at all. The mean removal goes from 0.85 against e_1 in the control to 0.87 against the median of the control, and 0.87 against the median of the anchored run. Every run stays within the spill criterion under every baseline.



***Removal and spill under each baseline**, the stored `falloff` runs, per seed. The gray circles (e_1 in the control) are the measurement quoted so far. Removal is at the full dose; spill is the largest over the four doses. The dashed rule is the spill criterion.*

baseline	removal ↑	within criterion ↑	spill, any dose ↓
e_1 in the control	0.85 (0.79 to 0.90)	8 of 8	0.003 (−0.004 to 0.018)
reference axes, control	0.87 (0.77 to 0.90)	8 of 8	−0.001 (−0.009 to 0.013)
reference axes, own run	0.87 (0.77 to 0.91)	8 of 8	−0.002 (−0.009 to 0.014)
nothing subtracted	0.96 (0.87 to 0.99)		0.017 (0.003 to 0.038)

***Removal and spill under each baseline**, the stored `falloff` runs, seed mean and range over 8 runs. Removal is at the full dose; spill is the largest net drop on another op over the four doses, and "within criterion" counts the runs where it is at most 0.02. Arrows mark which direction is better.*

The one seed that moves by more than a few hundredths is 705, where e_1 was among the busiest axes of the control. There the paired baseline (the gray circle in the figure above) gives a removal about a tenth lower than the medians do. Elsewhere the three baselines agree within a few hundredths, small next to the spread of removal across seeds.

seed	removal, e_1 in the control ↑	removal, reference axes, control ↑	removal, reference axes, own run ↑	removal, nothing subtracted ↑
704	0.88	0.89	0.91	0.97
705	0.79	0.90	0.91	0.99
706	0.85	0.83	0.83	0.92
707	0.87	0.88	0.90	0.98
708	0.79	0.77	0.77	0.87
709	0.86	0.89	0.90	0.99
710	0.90	0.88	0.89	0.97
711	0.86	0.89	0.89	0.97

***Removal at the full dose by seed**, under each baseline, the stored `falloff` runs. These are the values in the top panel of the figure above.*

The two medians agree on every seed, within 0.02 of removal. So in these runs the quiet axes of an anchored run cost about what the quiet axes of its control cost, and the anchor hardly touches them.

Spill falls slightly under both medians. On most seeds, e_1 in the control costs a little less on the other ops than the median does, so the paired baseline subtracts less.

The raw spill, with nothing subtracted, is higher (up to 0.038), so some baseline is still needed to separate the anchored op from the general cost of an edit.

The median of eight reference axes drawn at random, as spill-timing takes it, varies with the draw. Over many draws at each seed, the middle nine tenths of the removal spans at most 0.06, about half the seed range of removal, so eight axes are enough to tell a removable run from a weak one.

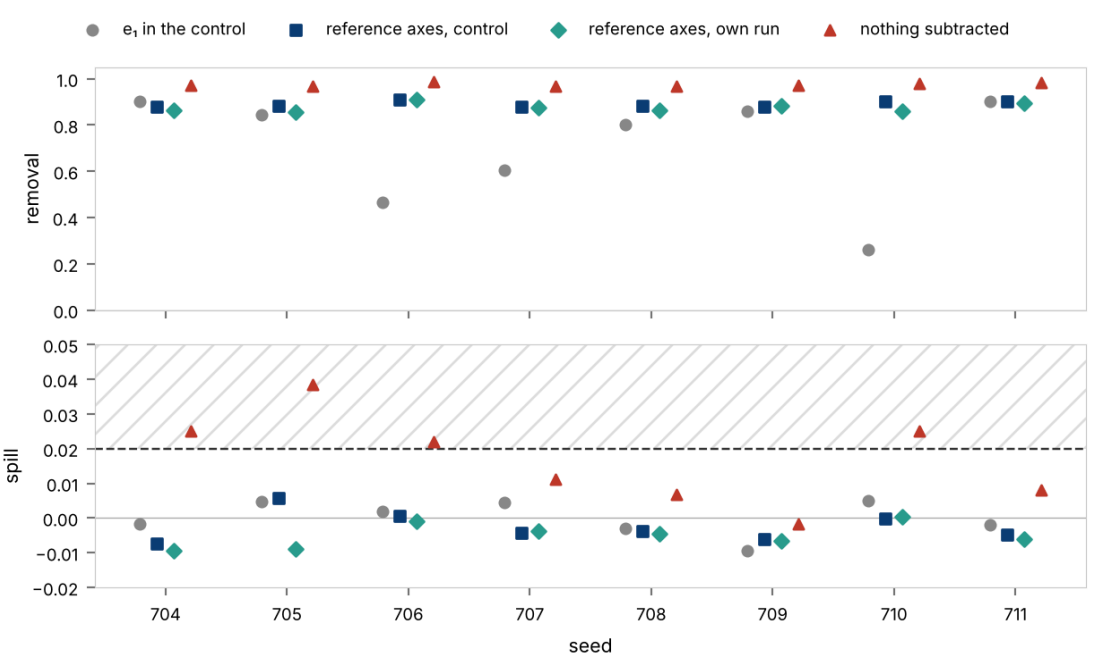
Over the same draws, spill stays at or below 0.018 on nineteen draws in twenty. The median over all the reference axes needs no draw.

The rotated-basis runs (E3)

In the rotated-basis runs, Adam put the cheap directions along the rotated axes. With the median taken along those axes, does the baseline put the short seeds back beside the rest?

What we saw. Yes. Against e_1 in the control, removal falls below half on seeds 706 and 710, and to about six tenths on 707. Against either median it sits near nine tenths on every seed, with every run within the spill° criterion.

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



***Removal and spill under each baseline**, the rotated-basis `falloff` runs, per seed. The gray circles (e_1 in the control) are the measurement quoted so far. Removal is at the full dose; spill is the largest over the four doses. The dashed rule is the spill criterion.*

baseline	removal ↑	within criterion ↑	spill, any dose ↓
e_1 in the control	0.71 (0.26 to 0.90)	8 of 8	0.000 (−0.009 to 0.005)
reference axes, control	0.89 (0.88 to 0.91)	8 of 8	−0.003 (−0.007 to 0.006)
reference axes, own run	0.88 (0.86 to 0.91)	8 of 8	−0.005 (−0.009 to 0.000)
nothing subtracted	0.98 (0.97 to 0.99)		0.017 (−0.002 to 0.038)

***Removal and spill under each baseline**, the rotated-basis `falloff` runs, seed mean and range over 8 runs. Removal is at the full dose; spill is the largest net drop on another op over the four doses, and "within criterion" counts the runs where it is at most 0.02. Arrows mark which direction is better.*

seed	removal, e_1 in the control ↑	removal, reference axes, control ↑	removal, reference axes, own run ↑	removal, nothing subtracted ↑
704	0.90	0.88	0.86	0.97
705	0.84	0.88	0.86	0.97
706	0.47	0.91	0.91	0.99
707	0.60	0.88	0.88	0.97
708	0.80	0.88	0.86	0.97
709	0.86	0.88	0.88	0.97
710	0.26	0.90	0.86	0.98
711	0.90	0.90	0.89	0.98

***Removal at the full dose by seed**, under each baseline, the rotated-basis `falloff` runs. These are the values in the top panel of the figure above.*

Again the two medians agree, within 0.04 of removal. The median of eight drawn axes spans at most 0.09 in removal over the middle nine tenths of draws, a little wider than under plain Adam.

The reference axes have to be the cheap ones for the basis the model trained in. Under the rotated basis a coordinate axis is an ordinary direction and costs more, as rotated-basis found, so a median over coordinate axes would overstate the baseline there.

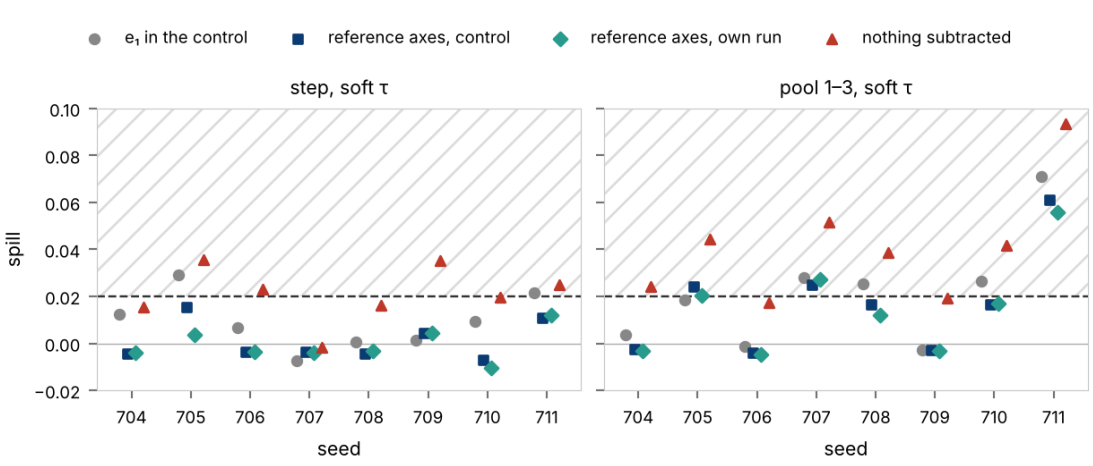
With an optimizer that has no privileged basis at all, every direction touches the spike, and no set of cheap reference directions exists in the same sense.

Runs near the spill criterion (E4)

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

The stored `falloff` runs are all well within the spill° criterion, so a change of baseline could not move a run across it. Two arms of soft-pool, `step, soft τ` and `pool 1–3, soft τ`, come close to it at the same seeds: against e_1 in the control, 2 and 4 of their 8 runs spill over it. How do those counts change against the medians? We compare the spill of each run, the largest over the four doses, under each baseline.

What we saw. For `step, soft τ` the count drops to none: both runs that were over the criterion fall under it against either median. For `pool 1–3, soft τ` the count barely changes (4 against e_1 in the control, 3 against the median of the control, 3 against the median of the run), but which runs are over it changes. Seeds 708 and 710 fall under the criterion, and seed 705 rises to it: just over against the median of the control, and over by less than a thousandth against the median of the run (the bold 0.020 in the table). The figure shows each run under each baseline and, for scale, with no baseline subtracted.



***Spill under each baseline, soft-pool runs**, per seed, for the two arms of soft-pool that come close to the criterion. Spill is the largest net drop on another op over the four doses; the triangles subtract no baseline. The dashed rule is the spill criterion.*

arm	baseline	704	705	706	707	708	709	710	711	over criterion ↓
`step, soft τ`	e_1 in the control	0.013	**0.029**	0.007	−0.007	0.001	0.002	0.009	**0.022**	2 of 8
`step, soft τ`	reference axes, control	−0.004	0.015	−0.004	−0.004	−0.004	0.004	−0.007	0.011	0 of 8
`step, soft τ`	reference axes, own run	−0.004	0.003	−0.003	−0.004	−0.003	0.004	−0.011	0.012	0 of 8
`pool 1–3, soft τ`	e_1 in the control	0.004	0.018	−0.001	**0.028**	**0.025**	−0.003	**0.027**	**0.071**	4 of 8
`pool 1–3, soft τ`	reference axes, control	−0.003	**0.024**	−0.004	**0.025**	0.017	−0.003	0.017	**0.061**	3 of 8
`pool 1–3, soft τ`	reference axes, own run	−0.003	**0.020**	−0.005	**0.027**	0.012	−0.003	0.017	**0.056**	3 of 8

***Spill by seed under each baseline**, the two soft-pool arms: the largest net drop on another op over the four doses. Values over the criterion of 0.02 are in bold. These are the values in the figure above.*

Which way a run moves depends on the op that spills. On every run over the criterion against e_1 in the control, the e_1 edit costs the spilling op less in the control than the median axis does. So the median subtracts more, and most of the runs that were just over the criterion fall under it. On seed 705 of `pool 1–3, soft τ` it goes the other way. There e_1 is among the busiest axes of the control (E1), and it costs `value-hsv` more than the median does, so the paired baseline hides the spill on that op. That fits the spill-timing experiment, which tracked this run against a reference edit and saw its spill grow late on `value-hsv`, ending near the criterion. The run that spills well over it (seed 711 of `pool 1–3, soft τ`) stays over under every baseline. So does seed 707, though it was as close to the criterion as 708 and 710. Against the median its largest spill moves to another op, `mix`, which e_1 in the control costs more than the median axis does.

Removal moves as little as it did for `falloff`, within a hundredth or two on average.

So the soft-pool spill counts depend on the baseline at the edge: a run within about a hundredth of the criterion can fall on either side. The counts against the median of the run are the ones to compare with later experiments, since the D2.2 design now scores that way. That design takes its median with the syntax axes in, which puts seed 705 just under the criterion, so `pool 1–3, soft τ` has 2 runs over it on that rule (seeds 707 and 711) and `step, soft τ` has none.

Discussion

For the stored runs of `falloff` the paired baseline was sound, in that it rarely moved the result. On most seeds e_1 happened to be a quiet axis of the control. Where it was not, the error was a tenth of removal, on one seed. The numbers that split-anti-seeds quoted stand. The paired baseline moved a spill° count only for runs at the edge of the criterion, as in the soft-pool runs of E4.

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

The median over the quiet axes removes that dependence on the draw, and it takes a few minutes more of scoring per run. Where the draw goes badly, as in the rotated-basis runs, it is the difference between a run that seems to have lost the anchor and one that removes nine tenths of the op.

The agreement of the two medians suggests a baseline needs no control at all: the quiet axes of the anchored run serve. That would matter for a recipe meant for real models, where a matched control is costly or impossible, and it is what spill-timing already does along training.

But the agreement rests on these sixteen pairs, and a recipe that disturbed the quiet axes of the anchored run would break it. So a control is still the way to check that a run-internal baseline is sound.

For later experiments the D2.2 design now takes a simpler scoring rule from this report. Leaving the syntax axes in barely moves the median, so the baseline is the median over every coordinate axis but e_1, with no ranking by energy, taken in the anchored run. An experiment then needs controls only now and then, to confirm that the two medians still agree, rather than one per seed. The one difference to watch is that the median of the run puts spill a little lower than the median of the control (by up to 0.012 here, on one seed). All of this assumes an optimizer like Adam, which keeps most axes quiet. Under one with no privileged basis, any reference direction touches the spike, so the baseline would be larger and an edit of any one direction dearer.

---

1. Adam and the basis: Adam scales the update of each coordinate separately, so the coordinate axes are special to it. Rotated-basis ran it in a rotated basis instead. ↩

2. Energy: the sum of the squared projections of the states onto a direction. The share of a direction is its energy over the total. ↩