When the spill appears

> We trained `pool 1–3, soft τ` again at its eight seeds and applied the edit every four epochs along training, to see when its spill° onto other ops appears. On three of the four seeds that end over the spill criterion it appears early in training, and on the fourth only late, on an op that run learned late. A color probe on the other ops showed nothing that tracks it.

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

`pool 1–3, soft τ` from soft-pool kept the anchor on all eight of its seeds, but only four of its edits stayed within the spill criterion. A cheap early test for spill would let it, or arms tuned from it, run at many seeds, as the loss-rate screen did for losing the anchor. But nothing in the stored trajectories up to epoch 64 predicted the end spill. The stored runs keep only their end checkpoints, so we couldn't tell when the spill appears.

So we trained the condition again at its seeds (704 to 711), with three measurements at every record:[1] the edit itself, a reference edit along another direction as its baseline, and a probe for the colors at the other ops. The reference edit replaces the control as the baseline, since there is no control trajectory to net against. We also scored the edit and the reference edit on the controls, and on the epoch-64 checkpoints the loss-rate screen left.

1 Record…

Observations

- The reference edit beside the control (E1): the reruns match the stored runs. At the end of training, netting the edit against the reference edit sorts the eight seeds much as netting it against the control does. On the controls themselves, the e_1 edit costs some ops a few hundredths more than the reference edit does and others less, so the control adds a seed-to-seed term of its own. An edit at epoch 64 on two low-spill conditions finds no spill° that isn't there at the end, but misses some that is.

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

- Spill along training (E2): spill appears in two ways. On three seeds it appears during the rise, as the pool sharpens and the anchor weight reaches its plateau. On 711 it stays to the end, on 707 it settles near the criterion, and on 708 it falls back under it. These three also swing the most from record to record while it lasts. On two seeds it grows after epoch 100, on `value-hsv`. Around epoch 64 the seeds sort roughly as they do at the end, apart from the late kind; the full order settles at about epoch 180.
- The late kind and learning `value-hsv` (E3): the late kind appears on two of the seeds that learn `value-hsv` last, while they learn it. On the seeds that learn it early, the edit costs `value-hsv` less than a reference edit does. The two quiet conditions learn it as late on some seeds and spill little on it.
- Color on e_1 at the other ops (E4): e_1 holds almost no color information at the other ops, from about epoch 30 on, on every seed, and removing it barely changes how well the color can be read from the rest of the state. The fit from the whole state doesn't follow the spill either.

Scope and method

This is an exploratory study, with no preregistration and no gate. It covers one condition, `pool 1–3, soft τ`, at eight seeds, trained exactly as in soft-pool. That condition pools the pull over slices 1 to 3, and its position and slice pool temperatures sharpen from soft to sharp over the 40-epoch warm-up. Training is deterministic, and the reruns reproduce the stored runs: the end spill° matches to the last digit on every seed (largest difference 0).

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

Eight seeds is few for correlations: a rank correlation of 0.6 over eight seeds comes up by chance about one time in ten. The seeds also split unevenly. One seed (711) spills far more than the rest, and the other three over the criterion are only just over it. We looked at several measurements, windows and slices after the runs came in, so we take a single strong correlation here as a lead only.

The criterion of 0.02 is a convention, carried over from ex-2.2.16, rather than a line the models draw themselves. On runs of conditions that don't spill, the score scatters by about 0.006 (one standard deviation, over 27 stored runs of such conditions), so the criterion sits about three of those above a quiet run. The seeds close to it (705, 708 and 710) could land on either side of a slightly different criterion or score.

The epoch-64 checkpoints are those of `falloff` and `step, soft τ` from the loss-rate screen, 68 in all, and these two conditions spill little. Eleven of the checkpoints belong to runs trained to the end, so their end spill is known, and three of those eleven end at or over the criterion. So the checkpoints can show whether an early edit finds spill that isn't there at the end, and whether it misses spill that is.

The measurements

- Spill, net of control (at the end, as earlier reports score it): the largest drop on another op under the edit, less the drop on that op under the same edit on the control at the same seed, taken over the four doses. The edit projects e_1 out at every position and slice. A run is within the criterion when its spill is at most 0.02. E1 shows that this score carries a seed-to-seed term from the control, since e_1 isn't a typical direction on an unanchored model either, so it is less reliable than it looked.
- Spill, net of reference (along training): the largest drop on another op under the full edit, less the median drop on that op under eight reference edits. Each reference edit projects out one other basis direction (e_2, e_3, …), chosen at random, and the same eight are used throughout. We use the median because on some seeds one of the eight directions costs every op far more than the rest: Adam gives the residual stream a privileged basis, with the separator tokens on one coordinate that holds about half the stream's energy and `=` on another, and our draw includes the separator coordinate of three seeds (704, 705 and 708). A label-free rule for later experiments is to draw the reference axes from outside the stream's highest-energy coordinates. Each drop is measured in expected exact match° over the whole held-out set, so the noise from the choice of contexts is small (a standard error of about 0.002 per op).

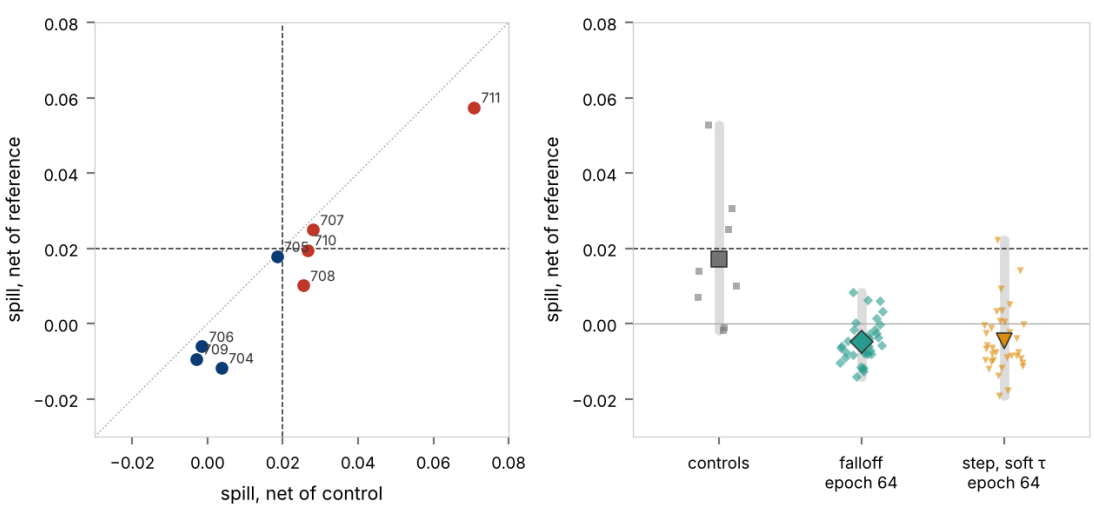
  Expected exact match. The task score, as the chance that a drawn answer agrees with a true one.…
- Smoothed spill and change per record (along training): the spill net of reference, smoothed with a Gaussian over about 8 epochs either side; and the size of its change from one record to the next, smoothed the same way, which measures how much it swings.
- When an op is learned: the first epoch at which its task score (without any edit) reaches half its final value. The ops are learned in jumps, and this marks the middle of the jump.
- The probe: a ridge regression (a linear fit with a small penalty on the weights) from the state at a position to the RGB of the color there. It is fitted on half of the held-out contexts and scored on the other half as R^2 (the share of the variance of the color it predicts). It is fitted at every color position of slices 2 to 4, on the contexts of the other ops and on `difference` contexts. Each fit uses one of three versions of the state: the whole state, the state without its e_1 coordinate, or the e_1 coordinate alone.

The reference edit beside the control (E1)

Can the reference edit stand in for the control as the baseline the spill° is netted against? At the end of training we have both, so the figure (left) and the table compare them on the eight reruns. We also apply the edit and the reference edit to the controls and to the epoch-64 checkpoints, where the edit should have little to act on (right).

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

We first tried a random direction orthogonal to e_1 as the reference edit, as planned. On a control it cost every op far more than the e_1 edit does, so a random direction is no baseline for e_1. A single other basis direction behaves much more like e_1 on a control, though, so we used eight of those as the reference edits. The gap comes from the few high-energy coordinates: a random direction touches all of them, and a random direction kept off them costs about what an ordinary axis does.



The reference edit beside the control. *Left: one point per rerun at the end of training. Vertical: the spill at the full dose, netted against the median of the eight reference edits. Horizontal: the spill as earlier reports score it, netted against the control at the same seed. Red where the run ends over the criterion, blue where it ends within it. The dashed lines are the criterion, and the dotted line is equality. Right: the reference-netted spill of the e_1 edit on other models, one small mark each, with the mean as a large mark and the range as a gray bar: the controls of ex-2.2.23 at seeds 704 to 711, and the checkpoints of the loss-rate screen at epoch 64.*

seed	stored ↓	rerun ↓	op	net of reference ↓	op	control, net of reference
704	0.004	0.004	value-hsv	−0.012	mix	−0.001
705	0.018	0.018	sat-hsv	0.018	value-hsv	0.053
706	−0.001	−0.001	lighten	−0.006	mix	0.014
707	0.028	0.028	darken	0.025	mix	0.007
708	0.025	0.025	value-hsv	0.010	mix	0.031
709	−0.003	−0.003	mix	−0.009	mix	0.010
710	0.027	0.027	value-hsv	0.020	value-hsv	−0.002
711	0.071	0.071	darken	0.057	darken	0.025

Spill at the end of training, by seed. *Stored and rerun: the stored run of soft-pool and the rerun, each netted against the control at the same seed (the largest net drop on another op at any dose), then the op that drop falls on. Net of reference: the rerun netted against the median reference edit at the full dose, then its op. Last column: the e_1 edit on the control, netted against the median reference edit on that control. The criterion is 0.02.*

The two scores rank the eight seeds alike (rank correlation 0.90), though they often name different ops. The reference-netted score runs a little lower, so the criterion would need restating for it. The seeds over the criterion stay at or over it, apart from 708, which falls under it.

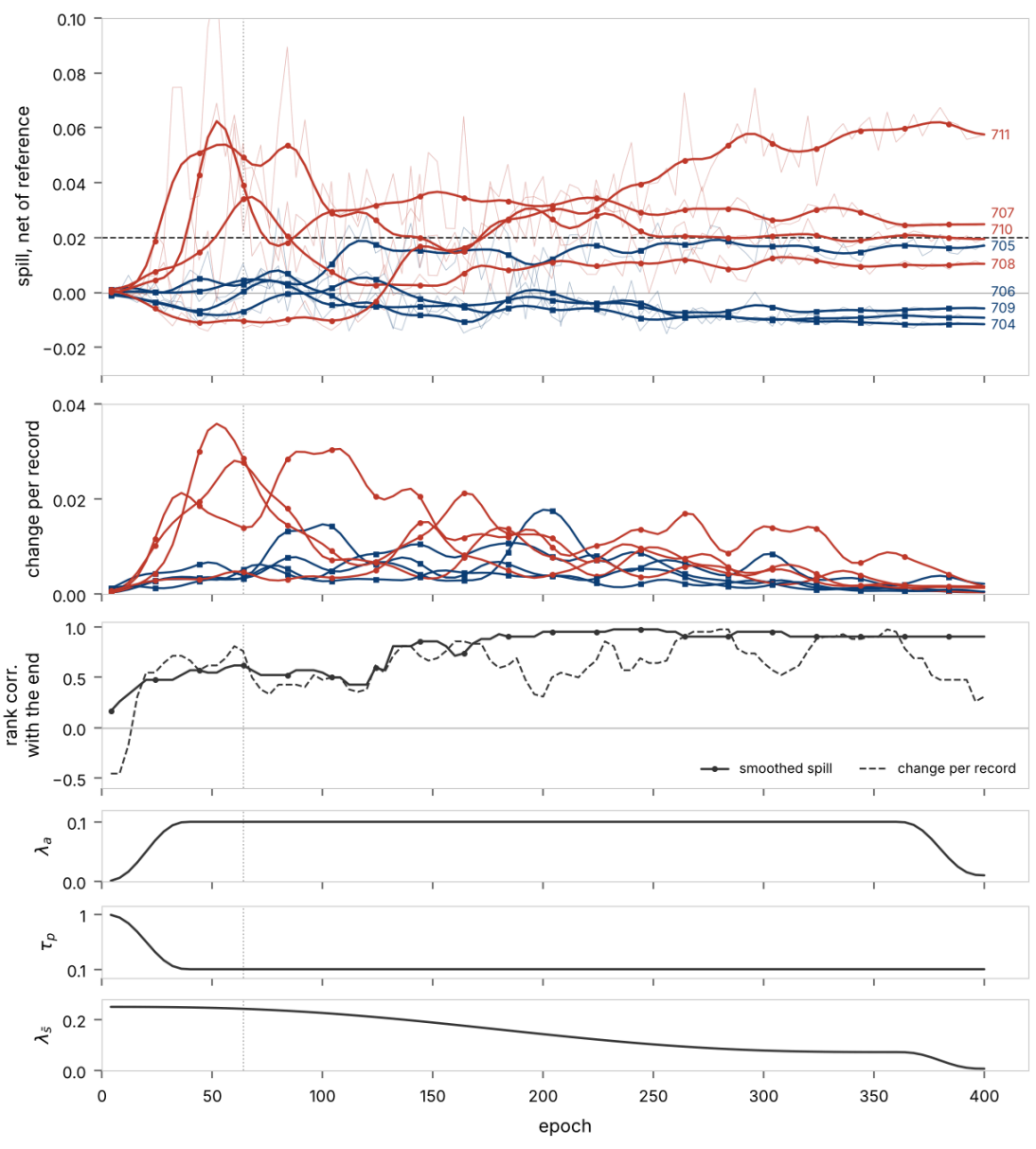
On the controls, the e_1 edit costs some ops up to a few hundredths more than the reference edit does, and others as much less, differently on each seed (the table shows only the largest). So e_1 on an unanchored model isn't quite a typical direction, and netting against the control at the same seed adds that seed-to-seed term to the spill score. That may be part of why quiet runs scored as high as 0.018 in earlier reports.

On the epoch-64 checkpoints of the two low-spill conditions, the edit costs the other ops a little less than the reference edit, as it should when the anti term keeps them off e_1. Only 1 of the 68 goes over the criterion, and on the eleven with a known end spill the early score stays below 0.01. So an early edit doesn't seem to invent spill. But it also misses the spill that three of the eleven end with (at or over the criterion). That is the late kind of spill from E2, seen here in two other conditions.

Spill along training (E2)

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

When does the spill° appear, and from when does it sort the seeds as the end does? The figure shows the reference-netted spill smoothed over neighboring records, how much it swings from one record to the next, and how well each of the two sorts the seeds as the end does. The table gives the spill averaged over a few windows.



Spill along training, smoothed. *Top: the largest drop on another op under the full edit, less the median drop under the eight reference edits, smoothed over about 8 epochs either side (a Gaussian), one line per seed, with the raw record (every four epochs) as a hairline: red for the seeds whose run ends over the criterion, blue for those within it. Second: the size of the change from one record to the next, smoothed the same way. Third: the rank correlation of each of the two with the end spill over the eight seeds, at every record. The dotted line marks where the loss-rate screen stopped. The strips underneath are the anchor weight λ_a, the position pool temperature τ_p (log scale), and the anti weight at slice 3, $\lambda_{\bar{s}}$.*

seed	epochs 24–40	epochs 48–80	epochs 100–160	epochs 200–240	epochs 360–400	end, net of control ↓	op at 64	op at 400
704	−0.005	0.001	−0.006	−0.002	−0.012	0.004	mix	mix
705	0.001	0.004	0.015	0.015	0.017	0.018	hue-hsv	value-hsv
706	0.000	0.006	−0.004	−0.004	−0.006	−0.001	darken	mix
707	0.001	0.046	0.033	0.033	0.025	0.028	hue-hsv	mix
708	0.012	0.028	0.003	0.011	0.010	0.025	darken	mix
709	−0.004	−0.006	0.001	−0.006	−0.009	−0.003	darken	mix
710	−0.009	−0.011	0.003	0.026	0.020	0.027	hue-hsv	value-hsv
711	0.041	0.050	0.022	0.034	0.060	0.071	darken	darken
rank correlation with the end	0.52	0.57	0.81	0.98	0.90			
correlation with the end	0.84	0.73	0.67	0.85	0.98			
correlation, without 711	0.30	0.51	0.64	0.92	0.93			

Spill along training, by window. *The reference-netted spill averaged over the records in each window, the spill at the end as earlier reports score it (net of the control), and the op the reference-netted spill falls on at epochs 64 and 400. The last three lines are the correlation of each window with the end spill over the eight seeds: by rank (Spearman), by value (Pearson), and by value over the seven seeds other than 711, which spills far more than the rest.*

Spill appears in two ways. The early kind rises on seeds 708 and 711 between epochs 24 and 32, as the pool temperature reaches its sharp value and the anchor weight its plateau, and on 707 at about epoch 44, just after. It falls mostly on `darken` and `hue-hsv`, and it stays up past where the loss-rate screen stopped. After that the three seeds differ. On 711 it dips around epoch 160 and climbs again to the end, on `darken`. On 707 it moves to `mix` and settles near the criterion, and on 708 it falls back to about half.

The late kind grows on seeds 705 and 710, from about epoch 120 and 150 respectively, on `value-hsv`, while the anti weight on slice 3 is partway through its slow anneal. It ends just under the criterion on 705 and just over it on 710. E3 looks at why these two seeds.

The early kind is also the jumpy kind. During the rise (epochs 20 to 180) the spill of 707, 708 and 711 changes by 0.016 per record on average, against 0.006 on the other five seeds, including 705 and 710, whose late spill grows smoothly. After epoch 200 every seed settles (a standard deviation of about 0.003 after epoch 300). So the swings flag the same runs as the level of the early spill does, and nothing more.

The seeds within the criterion settle a little below zero (704, 706 and 709, at about −0.01): there the edit costs every other op a little less than a typical other axis does. That is how a very clean anchor looks, with the other ops leaving e_1 alone while making some use of most other coordinates, though a score a little above zero is still within the criterion.

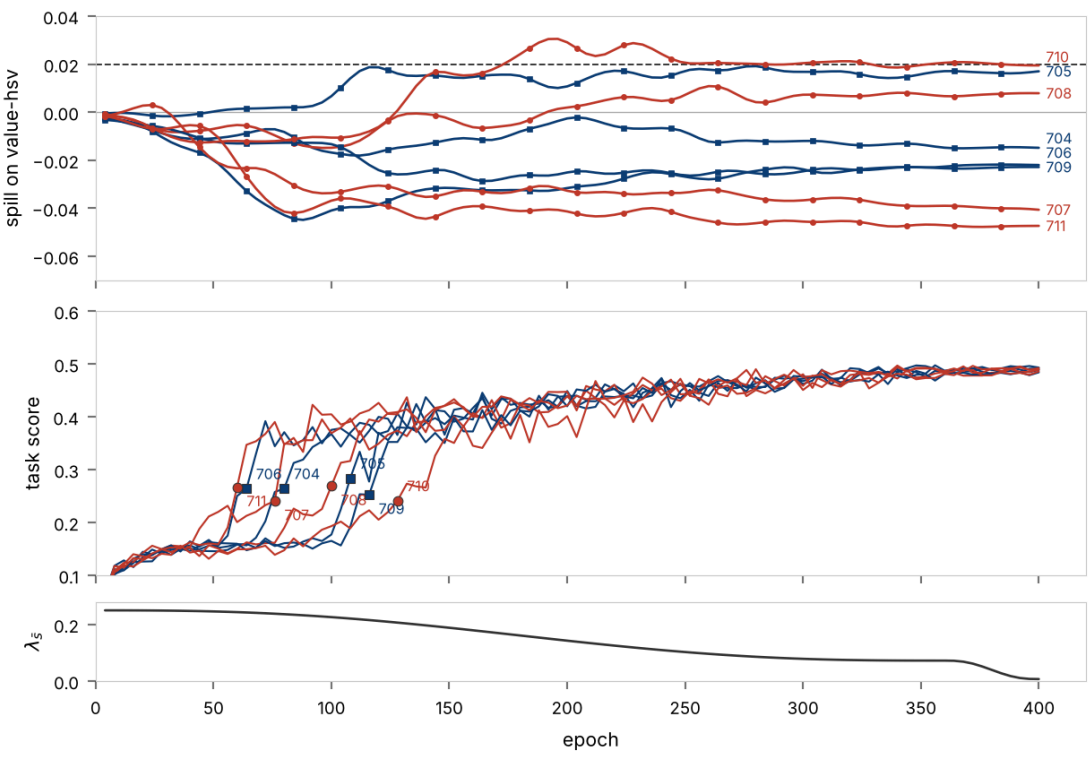
Smoothing doesn't bring the order forward much. The smoothed spill sorts the seeds as the end does from about epoch 180 on (a rank correlation over 0.9), and before about epoch 130 its rank correlation hovers around 0.5 to 0.6. Averaged over epochs 48 to 80, 707, 708 and 711 are clearly up, 710 is lowest of all, and 705 is among the runs within the criterion. The window over epochs 100 to 160 does better by rank (0.81) because the late rise of 705 and 710 has begun by then, but its correlation by value is 0.67. The correlations by value lean on 711 early on: without it, the window over epochs 24 to 40 correlates 0.30 with the end.

So a screen at epoch 64 would catch the early kind of spill and miss the late kind. On these seeds it would catch three of the four runs over the criterion, though two of those three end near it.

The late kind and learning `value-hsv` (E3)

Is the late kind tied to when the run learns `value-hsv`? Each op is learned in a jump, at a different time on each seed. The figure shows the spill° on `value-hsv` beside its task score.

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



Spill on `value-hsv` beside when it is learned. *Top: the drop on `value-hsv` under the full edit, less the median drop under the reference edits, smoothed as in the figure above, one line per seed. Middle: the task score on `value-hsv` (expected exact match without any edit), with a mark where it first reaches half its final value. Bottom: the anti weight at slice 3, $\lambda_{\bar{s}}$.*

The seeds learn `value-hsv` between about epoch 60 and 128. The two seeds with the late kind are among the last to learn it (705 at epoch 108, 710 at 128, the latest), and their spill on it grows while they learn it and then stays. 708, which learns it at 100, ends with a little spill on it too (0.008). On the seeds that learn it early, the edit costs `value-hsv` less than a reference edit does, by 0.015 to 0.05 at the end, so there `value-hsv` keeps further off e_1 than off a typical axis. Over the eight seeds, the later a seed learns `value-hsv`, the more the edit costs it at the end (rank correlation 0.71). 709 is the exception: it learns `value-hsv` late (epoch 116) and still keeps it off e_1.

When an op is learned isn't fixed by the seed, though (table). The control also learns `value-hsv` at different times on different seeds, between epochs 48 and 144, but which seeds are late changes with the condition, and in this condition the order is unrelated to the control's. So the anchor and anti terms shift when the op is learned. The two quiet conditions learn it as late or later on some seeds, `step, soft τ` at 180 on 709 and `falloff` at 164 on 706, and spill little on it. Learning late doesn't put an op on e_1 by itself; in those conditions something keeps e_1 closed to it.

condition	704	705	706	707	708	709	710	711	rank corr. with the control	spill on `value-hsv`
control	96	72	136	52	48	72	144	80		
`falloff`	100	136	164	48	84	100	112	72	0.58	none
`step, soft τ`	100	160	64	92	92	180	168	88	0.01	704
`pool 1–3, soft τ`	80	108	64	76	100	116	128	60	0.04	708, 710

When `value-hsv` is learned, by condition and seed. *The first epoch at which the task score on `value-hsv` reaches half its final value. The control and the two quiet conditions are read from their stored trajectories, which score each op on a subset of the held-out contexts; `pool 1–3, soft τ` is read from the reruns. Rank corr. with the control: the rank correlation of the row with the control row over the eight seeds. Spill on `value-hsv`: the seeds whose largest drop on `value-hsv` at the end, net of the control at the same seed, is over 0.01. Net of the control, 705 drops out: the edit costs its control `value-hsv` too (E1).*

The early kind fits the same pattern loosely. Every seed learns `darken` by epoch 28, during the warm-up, and the early spill of 708 and 711 falls on it. On 707 the early spill falls on `hue-hsv`, which 707 learns at epoch 48, earlier than any other seed.

So, as a lead: in this condition the spill lands on an op that is being learned at a time when e_1 is open to it, during the rise or later, and the quiet conditions keep e_1 closed to an op they learn as late. Within this condition, it doesn't say which way it runs. Learning `value-hsv` late may leave it free to land on e_1, or whatever draws it to e_1 may also slow its learning.

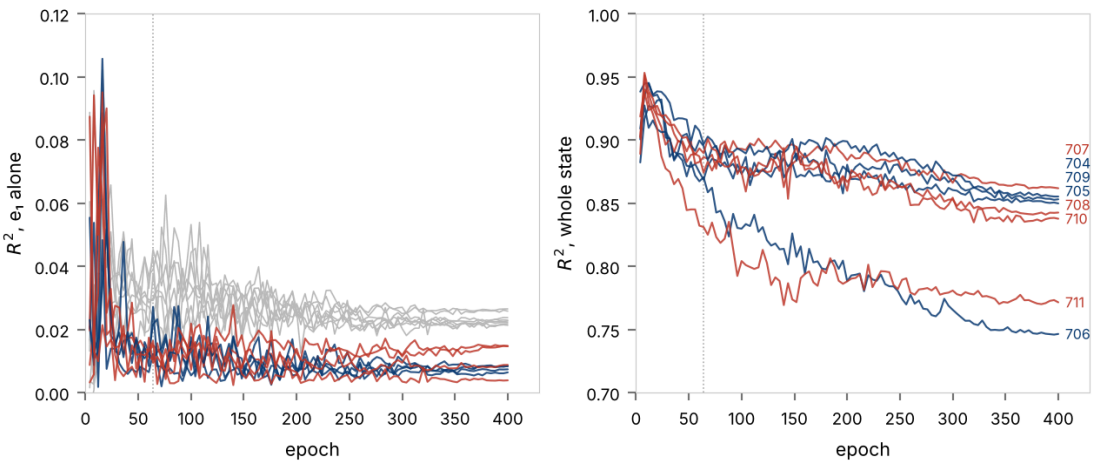
Color on e_1 at the other ops (E4)

Does e1 hold color information at the example answers of the other ops, which the edit would take away, and does that track the spill°?

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

This is a geometric measurement that doesn't apply the edit.

The figure shows the probe at the example answers, at slice 3.



***The color probe at the example answers**, slice 3: held-out R^2 of a ridge fit from the state to the RGB of the answer color, averaged over the three example answers. One line per seed, red for the seeds whose run ends over the criterion and blue for those within it. Left: from the e1 coordinate alone, on the contexts of the other ops (solid) and on `difference` contexts (ghosted). Right: from the whole state, on the contexts of the other ops. The dotted line marks where the loss-rate screen stopped.*

Removing e1 barely changes how well the color can be read from the state: the largest difference between the whole state and the state without e1 is 0.016 in R^2, at any position, slice or record, on the other ops and on `difference` alike. The color is spread over many coordinates, so one coordinate more or less makes little difference to a linear readout.

On its own, e1 holds little color at the example answers of the other ops. After the first few epochs its R^2 is about 0.01, with the seeds that spill and those that don't intermixed (rank correlation with the end spill −0.43 at epoch 64). On `difference` contexts it holds more, about 0.02, which is still little. What the anchor puts on e1 there is the op, `difference`, rather than a color, so there is no reason for e1 to carry much of the RGB.

The whole-state fit on the other ops falls slowly over training, and on two seeds it falls further: from the start on 711, which spills most, and after about epoch 60 on 706, which doesn't spill. So the fit doesn't follow the spill (rank correlation −0.43 at epoch 64, −0.07 at the end).

So the probe doesn't see the spill. The edit takes away e1, but what the other ops lose may be how later blocks read e1, rather than color information stored there.

Discussion

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

The edit spills° in two ways in this condition. The early kind appears during the rise, when the pool is still sharpening and the anchor weight reaches its plateau. That fits the reading of soft-pool that the soft pool lifts the other ops toward e_1 at that time. The late kind appears on `value-hsv` on two of the seeds that learn it last, while they learn it, as the anti term is partway through its slow anneal (E3). On 705 it starts while the anti weight is still near 0.2, so the anneal may matter less than when the op is learned. Both kinds land on an op that is being learned at the time. Neither reading is tested here; each would need an arm that changes the schedule at that time.

A screen that edits a checkpoint around epoch 64 would see only the early kind (E2). On these seeds the late kind was the smaller, but it put one run over the criterion and another just under it, so such a screen would undercount. A screen late enough to see both, around epoch 200, would cost half a full run.

Pooled over seeds, epoch 64 does tell the three conditions apart (table). `pool 1–3, soft τ` averages 0.014 at epoch 64, and half its runs end over the criterion. `falloff` and `step, soft τ` average about −0.005, and few of their runs end over it. So a condition whose runs often spill early shows it in the mean at epoch 64, even when a single run is noisy. But the mean can't see the late kind, and the two quiet conditions look alike at epoch 64 though they differ a little at the end. With three conditions this is a hint only: a pooled early score could rank arms by their early kind of spill, with a later check still needed for the late kind. The swings during the rise could serve the same way, since they mark the same runs.

Is that enough to stop a condition early? As a rule for dropping a condition whose runs spill early, it may be. Scoring a record costs about 1% of training, and the runs of a condition that passes simply carry on, so the screen costs almost nothing, and each condition it stops saves the remaining five-sixths of its runs. It can't stop a condition for the late kind, so a condition that passes still needs its full runs before it is adopted. It pays when many arms are screened and a fair share of them spill early. With few arms, or arms that spill late, it saves little.

condition	runs at 64	mean spill at 64	over at 64	runs at 400	over at 400
`pool 1–3, soft τ`	8	0.014	3	8	4
`falloff`	34	−0.005	0	12	1
`step, soft τ`	34	−0.005	1	11	2

The three conditions at epoch 64 and at the end. *Runs at 64: the reruns of this report, and the loss-rate screen's checkpoints. Mean spill at 64: the reference-netted spill at epoch 64 (a single record), averaged over the runs, and how many of them are over the criterion there. Runs at 400: those trained to the end, with their spill netted against the control as earlier reports score it, and how many end over the criterion (0.02).*

Netting against a reference edit works about as well as netting against a control at the end of training, and it needs no control run. The controls showed that e_1 isn't a neutral direction on an unanchored model either, so some of the noise in the earlier spill scores may come from the control rather than from the anchored run.

---

1. Record: a point along training where the measurements are taken, every four epochs. ↩