Ex 2.2.13: does a heavier anchor make the leftover predictable?

> The recipe from ex-2.2.11 leaves some *red* behind on one op of eleven, and how much swings widely from seed to seed. We climb a ladder of anchor weights, on an axis and on a plane, and ask whether a heavier anchor gives us a leftover we can predict. It does not.

Making the anchor 2.8× heavier raises the line margin, the quantity the anchor term optimizes, by about three percent. It does not narrow the seed-to-seed spread of the leftover on either subspace; the tight condition ex-2.2.12 saw was five lucky seeds. What does move the leftover is where *red* lives. On the plane the leftover is smaller at every rung, and the plane at the weight used by the recipe is the lowest in the experiment. Read as the adoption rule intended, that condition qualifies. We keep the recipe at `axis-0.1` all the same: the gain is small and it costs a second coordinate. The plane is now a validated alternative for a concept that needs more room. The anchored-op experiments inherit the leftover as a bounded confound rather than a fixed one.

Findings

- The leftover gets more predictable (H1) — did not hold. On the axis the spread falls a little as the weight rises, to 0.77 of the spread at the reference, against a gate of 0.5. On the plane it widens. The tight plane condition from ex-2.2.12 does not replicate at twenty seeds.
- The leftover does not get smaller (H2) — held on the axis. The plane lowers the mean at every rung, `plane-0.1` by 0.10, with an interval clear of zero. The weight moves the mean in no consistent direction.
- The worst op improves (H3) — held on the axis, where the worst other op falls by 0.062 at the top rung. It missed on the plane, where the worst other op turns back up at the top rung.
- What the weight spends (H4) — held. Every cost statistic is inside its gate at every condition. The line margin rises by about three percent across the ladder, so the treatment does reach the model, and it is close to saturated at the weight used by the recipe.

The adoption rule: `plane-0.1` clears every clause about the leftover and the costs, and fails the $\bar{\alpha}$ clause only on that clause's arithmetic, a fault in the rule as written. Read as intended, it is the one condition that qualifies. The recipe stays at `axis-0.1` by a decision made after the data: the gain is about 0.10 on the leftover's mean, and it costs a second coordinate of the stream.

How to read this draft

The ladder, the four predictions, and the adoption rule were fixed before any run, at commit `c533700`. Everything after that commit is either results filled into their sections or exploratory work, marked as post hoc.

Every anchored run is new. The seeds that produced the observation this experiment follows up are not reused, so the reference is trained again here beside the ladder. The un-anchored control is borrowed from ex-2.2.11. The adoption rule summarizes a leftover by an upper confidence bound rather than by the fixed band ex-2.2.12 used; the adoption rule section gives both verdicts.

Why this experiment

Ex-2.2.11 put *red* on one axis of the residual stream of a small transformer,[1] taught it eleven color operations, and projected the axis out. On ten ops the model lost *red*. On `hue-hsv` it kept about a quarter of the answers that need the red operand, which is over the gate.

Ex-2.2.12 asked why and got a clean negative result. The story about which part of a hue an axis can hold was wrong, no change to the recipe brought the leftover under the gate, and the survival turned out to come from two of the seven red colors.

It also left something unexplained. The leftover swings from seed to seed: the reference keeps anywhere from 0.06 to 0.41 of those answers, depending on which seed trained the model. One condition of the sweep, the plane at twice the anchor weight, kept 0.14 to 0.22 across its five seeds instead. A range over five draws is narrower than a range over twenty at the same spread, so the two ranges are not directly comparable; the standard deviations behind them, 0.036 against 0.125, are what H1 is built on. A standard deviation from five seeds is itself a loose estimate, which is why the gate below asks for less than that ratio.

That is the observation this experiment is built on, and it is not the one the sweep was scored on. A leftover we cannot remove is a confound for every anchored-op experiment that follows.

If it is the same size every time, we can measure it once and subtract it. If it varies three-fold across seeds, every experiment that meets it has to measure it again. So how wide the seed spread is counts as a result in its own right. Read more broadly, the spread of the leftover is one face of how reproducible training is under the recipe on this grammar, and the ladder is a test of whether a heavier anchor makes training more reproducible.

A second reason to expect something from the weight: we run at λ_a = 0.1 because ex-2.1.6 chose a rung safely inside the region where the task is unhurt. The survey in ex-2.1.11 later mapped the weight on the six-op grammar, putting the plateau of the margin at λ_a 0.28–0.94, with the task gate biting only from about 0.38. On that map we have been running near the bottom of the useful range the whole time, and nothing has tested the region in between.

Conditions

One ladder, crossed with the home of *red*, and the un-anchored control beside it.

condition	λ_a	home of *red*	seeds	seen before
`axis-0.1`	0.1	axis	20	ex-2.2.11's recipe, the reference; 20 seeds there
`axis-0.14`	0.14	axis	20	
`axis-0.2`	0.2	axis	20	ex-2.2.12's `lam-0.2`, 5 seeds: loose on the kept share, the tightest line margin and ᾱ in the sweep
`axis-0.28`	0.28	axis	20	
`plane-0.1`	0.1	plane	20	ex-2.2.12's `plane`, 5 seeds
`plane-0.14`	0.14	plane	20	
`plane-0.2`	0.2	plane	20	ex-2.2.12's `plane-lam-0.2`, 5 seeds: the tight kept share
`plane-0.28`	0.28	plane	20	
`control`	0	none; scored on the axis	5	ex-2.2.11's `control`, served from the store at seeds 100–104; the task reference
`control-plane`	0	none; scored on the plane	—	the same checkpoints, a scoring pass; the ᾱ baseline for the plane

The weight. λ_a takes four levels, 0.1, 0.14, 0.2, 0.28, each a factor of √2 above the last. We read a weight on a log scale, so a constant ratio puts the levels evenly apart and makes the shape of any trend across them easy to see. The lowest rung is the recipe from ex-2.2.11, and 0.2 is the one step ex-2.2.12 took. The top rung is where the margin plateau ex-2.1.11 mapped begins, and it stops short of the level where that survey saw its first task failures (about 0.38, on the six-op grammar at a warmer pooling temperature than ours). Where the useful range ends is left open here; the ladder tests the region between the recipe and that point.

What else the weight moves. The anti-subspace term is specified as a ratio to λ_a, peaking at 2.5× λ_a and holding at 0.3× λ_a, so a rung of the ladder raises the repulsion by the same factor as the pull. The ladder is a joint anchor-and-repulsion ladder rather than a pure anchor ladder, and a result along it belongs to the pair. Ex-2.2.12 moved the two separately, one step each (`lam-0.2` and `anti-5`), and neither moved any measurement on its own, so we do not spend runs here separating them; if the ladder moves something, an arm that holds the repulsion fixed at one rung is the follow-up.

The home of *red*. *Red* lives either on the first axis e_1 or on the plane spanned by e_1 and e_2. Where the plane is the home, the anchor term, the anti-subspace term, the alignment measurements, and the removal all take the pair of axes instead of the single axis, as ex-2.2.12 defined them.

The recipe has two anchoring terms and no anti-anchor term. The pull is one minus the alignment of a labelled state with the home; on the axis that alignment is the signed cosine, so the pull is toward $+e_1$ and a state at $-e_1$ is as far from home as it can be. On the plane it is the unsigned length of the projection, so the pull is toward the plane with no preferred direction within it, and $-e_1$ is home. The anti-subspace term is the squared alignment averaged over every live position, labelled or not, and it has no sign in either home. An anti-anchor term, the one-sided hinge that kept every state out of the hemisphere opposite the anchor so a fallback could live there, belonged to M1's fallback and last ran in M2 in ex-2.2.2's fallback arm; the recipe line from ex-2.1.6 onward has never carried it, and on the plane there is no hemisphere for it to name.

Ex-2.2.12 settled against the plane's first rationale, that a single axis is too small a home for a hue. Two things keep it here. The tight condition that prompted this experiment was a plane *and* a heavier weight, while the heavier weight on the axis was among the loosest on the kept share in that sweep even as its line margin and ᾱ were the tightest, so a ladder on the axis alone could not say whether the narrowing comes from the weight, from the subspace, or from the two together. And the plane at the recipe's own weight has been seen at five seeds only; at twenty, `plane-0.1` against `axis-0.1` is a test of the subspace on its own, at a resolution ex-2.2.12 did not have.

The control. The un-anchored control is ex-2.2.11's, served from the store at its seeds rather than retrained. It is the reference for the task gate, which asks for a seed mean within a band of the control's, and the ᾱ baseline for the axis conditions. Neither role touches the observation this experiment follows up, which was read off anchored checkpoints, so the borrow spends nothing the design needs; comparisons against it are unpaired across seed sets, which a comparison of seed means absorbs. Every plane condition is compared against the same checkpoints scored on the plane (`control-plane`): for every state, anchored or not, an unsigned two-dimensional alignment sits higher than a signed one-dimensional one, so the control has to be scored the same way.

The seeds. 20 per condition, all fresh: condition seed *i* trains at model seed 200 + *i*, where ex-2.2.11 and ex-2.2.12 used an offset of 100. Every condition pairs with every other one seed for seed within this experiment, and the reference is retrained rather than borrowed, which makes it a replication of ex-2.2.11's `handover` at seeds it never saw.

The spread question sets the count. A one-sided F-test on the ratio of variances between the top and the bottom of the ladder resolves a halved standard deviation with power 0.90 at twenty seeds a group, 0.81 at fifteen, and 0.63 at ten; a difference in means of the size we care about would be settled by half as many. 160 runs in all.

Everything else is unchanged from ex-2.2.11: table A+, the stochastic corpus, the whole-line labeller, the untied readout, the removal lines chosen by hue, τ = 0.1, the shape of the anti-subspace schedule, and 50 epochs at d64-L4.

Glossary

Kept share

How much of its clean accuracy on an op's removal lines a model keeps after the projection. One means the projection did nothing; zero means every one of those answers changed, which is what the gate takes *red* being gone to mean.

Removal lines

The red lines whose answer needs the hue of the red operand: some permutation of the channels of that operand moves the true answer far. The rule from ex-2.2.11, unchanged.

Seed spread

The standard deviation of a statistic across the seeds of one condition. This is the quantity H1 is about.

Upper bound

The one-sided 95% upper confidence bound on the seed mean of a condition. One condition can have a low mean and a wide spread, and another a higher mean and a narrow spread, and the two can still have the same upper bound. That is the point of using it.

Line margin

How far the anchored states on the labelled lines sit above the rest along the home of *red*. This is the quantity the anchor term optimizes, so it is a check that the treatment landed rather than a result.

ᾱ at op1

The mean alignment with the anchored subspace over every color at the first operand position. How much the colors that are not red have drifted toward the home of *red*.

The leftover gets more predictable (H1)

✗ Miss

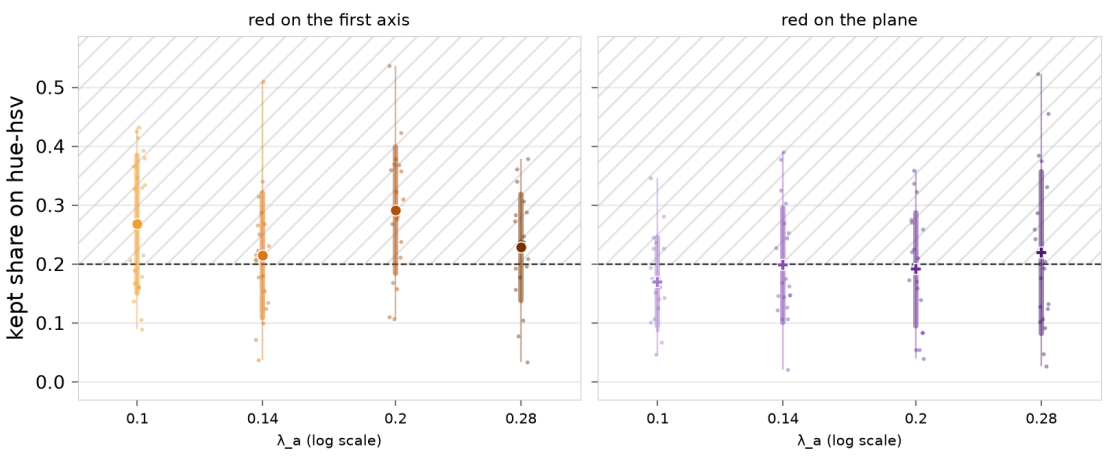
What we expect. Across the ladder, the seed spread of the kept share on the `hue-hsv` removal lines narrows as λ_a rises. The number we score is the ratio of the standard deviation at the top of the ladder to the one at the lowest rung, within a subspace. H1 holds when that ratio falls to 0.5 or below in both subspaces, and the spread falls monotonically enough that a trend contrast across the four levels has a negative slope at 0.05.[2]

It holds in part when one subspace does that and the other does not. On the plane alone, that would say the narrowing needs the plane, and the comparison of `plane-0.1` with the reference then says whether the plane narrows the spread by itself or only once the weight rises. A flat or rising spread in both would mean the tight condition in ex-2.2.12 was five lucky seeds, and that nothing on this ladder buys predictability.

One reading of a narrowing has to be ruled out before it counts. The kept share is a proportion over a fixed set of lines, so its spread is bounded below by sampling noise that depends on where the mean sits: a condition whose mean is near zero or near one cannot vary much. H1 is therefore read together with H2. A narrowing that arrives with a mean the ladder also moved is a narrowing we get for free, and the report says so rather than claiming the recipe bought it.

The kept share is a fair thing to score here. The anchor term acts during training on how well a state aligns with the home of *red*; the kept share is taken afterwards, on held-out lines, through a projection the term never sees. Nothing in the treatment is set up to reduce its spread, so a narrowing would be telling us something.

The quantities the two terms do act on — the line margin and ᾱ at op1 — are reported under H4 as manipulation checks rather than with gates on them.



Kept share on the `hue-hsv` removal lines along the ladder. *Left, red on the first axis; right, red on the plane. Each column is one condition: 20 faint seed dots, a thin bar spanning the seed range, a thick bar spanning one standard deviation either side of the mean, and the seed mean as the large mark. H1 is about the height of the thick bar rather than where the mark sits. The dashed line is the 20% gate, hatched above.*

condition	seed mean ↓	SD ↓	range	n	seen before (mean, SD, n)
`axis-0.1`	0.268	0.116	0.09–0.43	20	ex-2.2.11 0.239, 0.105, 20
`axis-0.14`	0.215	0.106	0.04–0.51	20	—
`axis-0.2`	0.292	0.107	0.11–0.54	20	ex-2.2.12 0.276, 0.118, 5
`axis-0.28`	0.229	0.090	0.03–0.38	20	—
`plane-0.1`	0.170	0.075	0.05–0.35	20	ex-2.2.12 0.256, 0.125, 5
`plane-0.14`	0.198	0.097	0.02–0.39	20	—
`plane-0.2`	0.191	0.096	0.04–0.36	20	ex-2.2.12 0.192, 0.036, 5
`plane-0.28`	0.219	0.137	0.03–0.52	20	—

Kept share on the `hue-hsv` removal lines per condition, at 20 seeds. The last column is the same measurement as the experiment that saw that rung took it, at its own seed count, as a replication rather than as data. axis: SD ratio top/bottom 0.77 against a gate of 0.5, Levene W = 1.59 (p = 0.198), trend slope -0.0330 per log unit of λ_a (one-sided p = 0.032); plane: SD ratio top/bottom 1.83 against a gate of 0.5, Levene W = 1.97 (p = 0.126), trend slope +0.0435 per log unit of λ_a (one-sided p = 0.988).

What we saw. H1 did not hold. On the axis the spread falls a little with the weight: the standard deviation at the top rung is 0.77 of the reference's, against a gate of 0.5, and the trend contrast has a negative slope at one-sided p = 0.032. The trend clause passes and the ratio clause does not, so the narrowing is there and it is small. On the plane the spread widens: the ratio is 1.83 and the slope is positive.

The observation this experiment followed up did not replicate. Ex-2.2.12's `plane-0.2` had a standard deviation of 0.036 over five seeds; the same condition at twenty fresh seeds has 0.096, in line with every other rung. Its mean did replicate (0.192 then, 0.191 now), and so did the reference's (0.239 at ex-2.2.11's seeds, 0.268 here). A standard deviation from five draws has the wide interval the prereg noted, and this is what the low end of that interval looks like when it comes up.

The sampling-noise reading H1 asked us to rule out does not arise: no mean moved close enough to zero or one to bound its spread. The means themselves are H2.

> **Miss**
>
> The spread narrows a little on the axis and widens on the plane; neither reaches the gate, and the tight condition from ex-2.2.12 does not replicate.

The leftover does not get smaller (H2)

~ Partial

What we expect. The seed-mean kept share on `hue-hsv` is flat across the ladder: no level differs from the reference by more than 0.05, which is about what twenty paired seeds can resolve at the spread of the reference. H2 is a statement about what we can see at this resolution rather than a claim that the mean is unmoved, and a reviewer reading it as an equivalence test with a wide band is reading it right.

The evidence for it: the one step ex-2.2.12 took moved the mean up on the axis and down on the plane, and its tight plane condition sits almost exactly on the mean of the five reference seeds it pairs with. A level that does move the mean down by more than 0.05 would be a better result than we expect, and the adoption rule is written to take it.

Together, H1 and H2 give the shape of the claim: climbing the ladder leaves a leftover that is no smaller and that comes out the same size every time.

The means themselves are the large marks of the H1 figure, read along the λ_a axis; the table below pairs them seed for seed with the reference.

condition	seed mean ↓	paired Δ from `axis-0.1`	95% interval	|Δ| ≤ 0.05
`axis-0.1`	0.268	–	–	–
`axis-0.14`	0.215	-0.053	-0.117 to +0.011	no
`axis-0.2`	0.292	+0.024	-0.049 to +0.096	**yes**
`axis-0.28`	0.229	-0.039	-0.093 to +0.015	**yes**
`plane-0.1`	0.170	-0.098	-0.170 to -0.026	no
`plane-0.14`	0.198	-0.070	-0.134 to -0.005	no
`plane-0.2`	0.191	-0.077	-0.153 to -0.000	no
`plane-0.28`	0.219	-0.048	-0.132 to +0.035	**yes**

The seed-mean kept share on `hue-hsv` and its paired difference from the reference. Every condition trains at the same 20 seeds, so each difference is paired seed for seed; the interval is Student's t at 19 degrees of freedom. H2's band is 0.05.

What we saw. H2 held on the axis and not on the plane. On the axis, two of the three rungs sit inside the band and `axis-0.14` just outside it, with no trend in the weight: the mean goes down, up, and down again along the ladder. On the plane every rung is below the reference, three of the four by more than the band, and `plane-0.1` by 0.098 with a 95% interval of 0.026 to 0.170 below it. That is the comparison ex-2.2.12 could not resolve at five seeds, and at twenty it says the subspace lowers the leftover on its own. Adding weight on top of the plane moves the mean back up a little rather than further down.

So the shape of the claim in the prereg, no smaller and the same size every time, came out the other way round: the plane makes the leftover somewhat smaller, and nothing on the ladder makes it more predictable.

> **Partial**
>
> Flat within the band on the axis; lower than the reference at every rung on the plane.

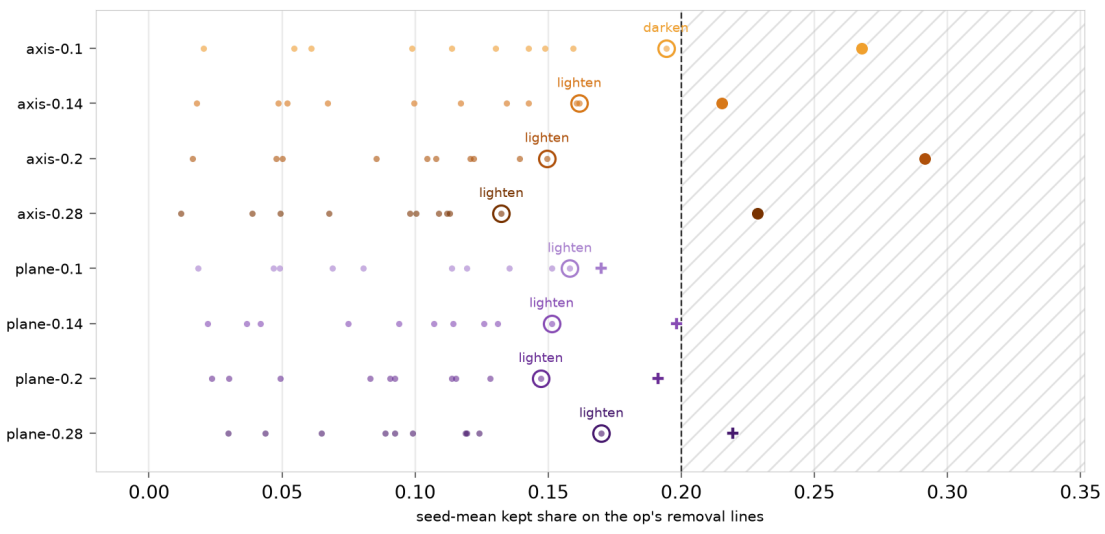
The worst op improves (H3)

~ Partial

What we expect. On the ten ops other than `hue-hsv`, the highest per-op seed-mean kept share falls as λ_a rises. The number we score is that highest value at the top of the ladder against the same quantity at the reference, both at twenty seeds; H3 holds when it falls by at least 0.03.

Any claim that a recipe removes *red* cleanly is limited by whichever op does worst, and that is not the same op in every condition, so a mean over the ten would hide it.

The margin is the 0.03 in the statement above, and it is there because a kept share is a proportion over some 330–400 removal lines per op, so on one checkpoint it carries a sampling error of about 0.02 at the values the worst op sits at, and the seed spread of the other ops in ex-2.2.12 was about 0.03. A drop smaller than that is one we could not tell from no drop. In ex-2.2.12 the worst other op was `darken` at the reference, `darken` again one step up the axis, and `lighten` on the plane at twice the weight — the only condition in that sweep with no op over the gate at all — and the spacing between those three values is close to the margin, which is why H3 asks for a margin rather than for a difference. H3 fails if the worst op is flat or rises, which would mean the ladder buys nothing on the ops that already remove *red*.



Per-op seed-mean kept share, per condition. *One row per condition; each small mark is one op's kept share on its own removal lines, averaged over the 20 seeds. The hollow ring is the worst of the ten ops other than `hue-hsv`, named beside it; the large mark is `hue-hsv` itself. The dashed line is the 20% gate, hatched to the right.*

condition	worst other op	its seed mean ↓	Δ from `axis-0.1`	ten-op mean ↓
`axis-0.1`	`darken`	**0.194**	—	0.112
`axis-0.14`	`lighten`	**0.162**	**-0.033**	0.100
`axis-0.2`	`lighten`	**0.150**	**-0.045**	0.094
`axis-0.28`	`lighten`	**0.132**	**-0.062**	0.083
`plane-0.1`	`lighten`	**0.158**	**-0.036**	0.094
`plane-0.14`	`lighten`	**0.151**	**-0.043**	0.090
`plane-0.2`	`lighten`	**0.147**	**-0.047**	0.087
`plane-0.28`	`lighten`	**0.170**	-0.024	0.095

The worst of the ten ops other than `hue-hsv` per condition, and the mean over those ten. H3 asks for a fall of at least 0.03 from the reference at the top of the ladder; bold in the last column marks a fall that large. The worst op is not the same op in every condition.

What we saw. H3 held on the axis and missed on the plane. The frozen wording does not split the top of the ladder by subspace, so we read it with H1's convention and call it partial. On the axis the worst other op falls at every rung, and by 0.062 at `axis-0.28`, twice the margin. On the plane it falls by the margin or more at the two middle rungs and turns back up at `plane-0.28`, to 0.024 under the reference, short of the margin. The worst op is `darken` at the reference and `lighten` everywhere else, as the spacing in ex-2.2.12 suggested, and every condition's worst other op is under the gate. The ten ops that already remove *red* have some room to spare, and a heavier anchor on the axis uses a little of it.

> **Partial**
>
> The worst other op falls by twice the margin at the top of the axis ladder, and turns back up at the top of the plane's.

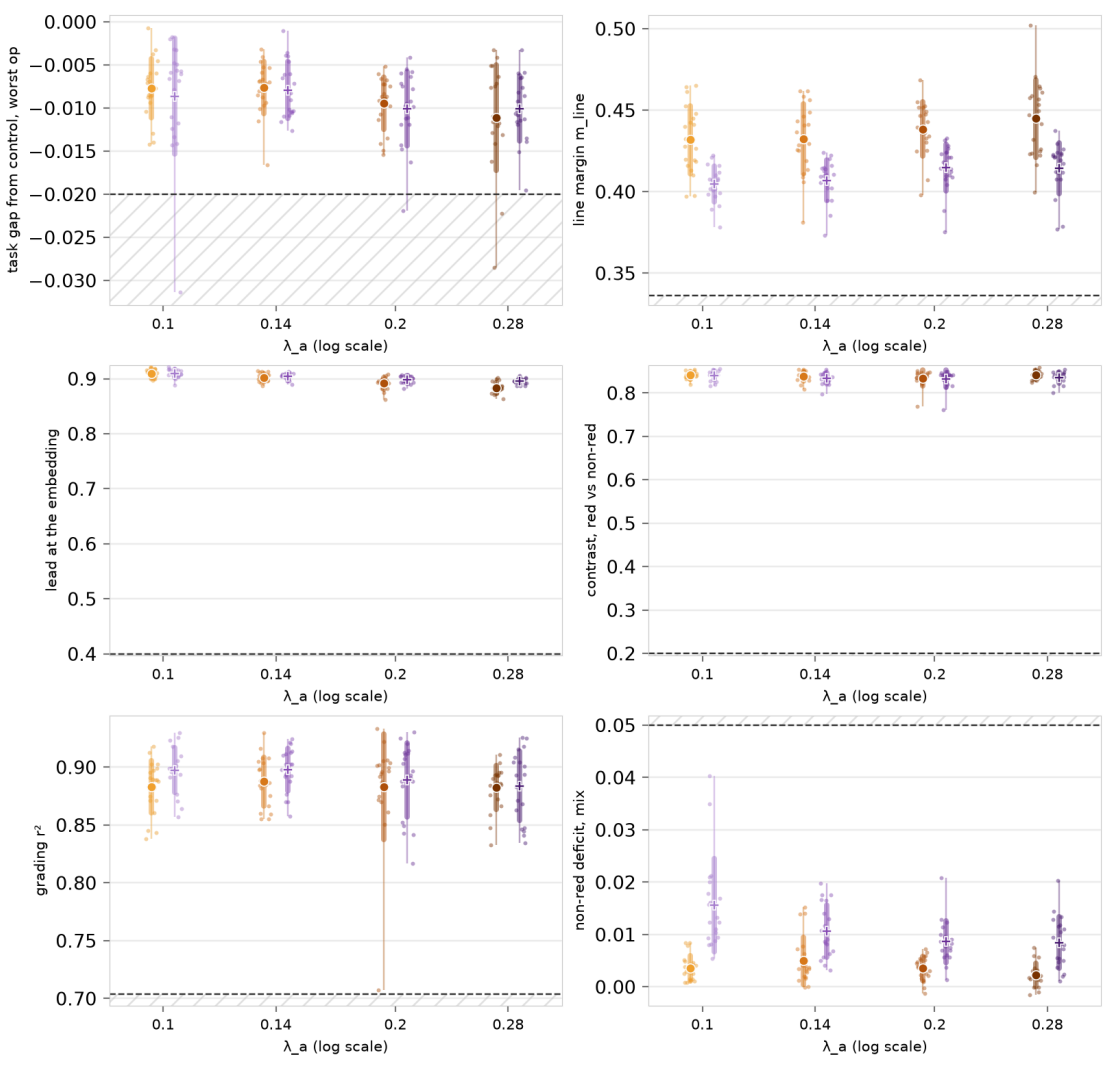
What the weight spends (H4)

✓ Pass

What we expect. Every cost statistic stays inside its gate from ex-2.2.11 on the seed mean through the whole ladder: the task, the line margin, the lead at the embedding, the contrast between red and non-red, the grading r^2, and the deficit on the non-red lines. The ladder stops under the level where the six-op survey saw the task give way, so we expect every cost to hold; the one we name as most likely to leave its gate first is the non-red deficit on the plane conditions, which ex-2.2.12 measured four times higher on the plane than on the axis.

If a statistic leaves its gate at or below $\lambda_a = 0.28$, the useful range of the recipe on this grammar is narrower than the map in ex-2.1.11 suggested, which is worth knowing on its own. If none does, the range is at least this wide, and where it ends stays open.

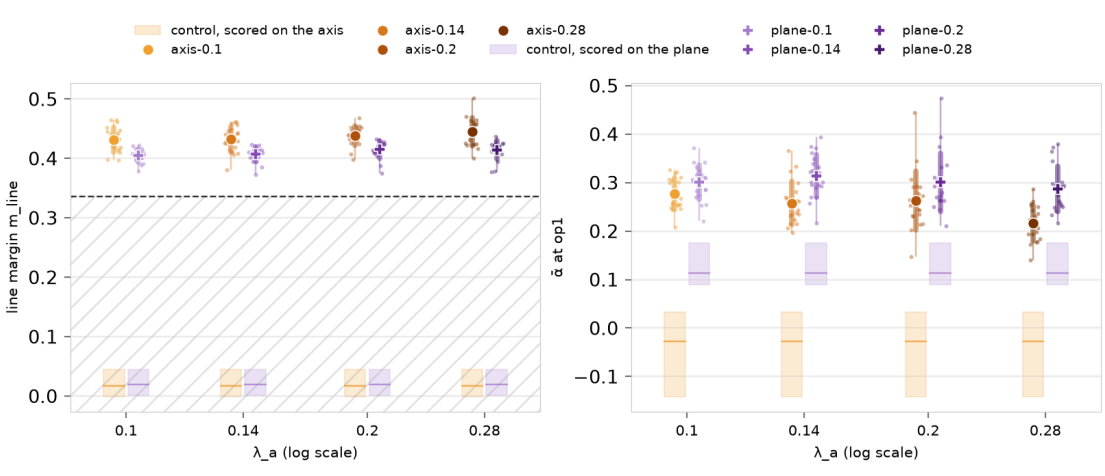
We report the line margin here as the manipulation check. It is the quantity the anchor term optimizes, and a heavier pull should raise it along the ladder — one direction, with no argument available for the other. If it stays flat, the weight is not reaching the model, and the other three sections have no treatment to interpret. $\bar{\alpha}$ at op1 goes in the same figure without a direction attached to it: it is what the anti-subspace term acts on, and that term is climbing the ladder alongside the pull.



The cost side along the ladder. *One panel per statistic H4 names, each against its gate from ex-2.2.11 (dashed, failing side hatched): the held-out expected exact match less the control's on whichever op the condition is worst on; the line margin on the `mix` lines; the lead at the embedding; the contrast between the red and non-red groups; the grading r^2; and the non-red deficit under the projection. At each weight the left column is the axis and the right the plane; each is 20 seed dots with a standard-deviation bar and the seed mean.*

condition	the task	the line margin	the lead at the embedding	the contrast between red and non-red	the grading r^2	the deficit on the non-red lines
`axis-0.1`	-0.008 (`hsvmix`)	0.432	0.910	0.841	0.883	0.004
`axis-0.14`	**-0.008 (`value-hsv`)**	0.432	0.902	0.838	0.888	0.005
`axis-0.2`	**-0.009 (`value-hsv`)**	0.438	0.892	0.834	0.883	0.004
`axis-0.28`	**-0.011 (`value-hsv`)**	0.445	0.883	0.841	0.882	0.002
`plane-0.1`	**-0.009 (`value-hsv`)**	0.405	0.909	0.841	0.897	0.016
`plane-0.14`	**-0.008 (`value-hsv`)**	0.407	0.904	0.834	0.898	0.011
`plane-0.2`	**-0.010 (`hsvmix`)**	0.415	0.899	0.833	0.889	0.009
`plane-0.28`	**-0.010 (`hsvmix`)**	0.415	0.896	0.836	0.884	0.008

Every cost statistic per condition, on the seed mean, with a value inside its ex-2.2.11 gate in bold. The gates: task ≥ −0.02, margin ≥ 0.336, lead ≥ 0.4, contrast ≥ 0.2, $r^2 \geq 0.704$, deficit ≤ 0.05. The task column names the op the condition is worst on.



The two quantities the treatment acts on. *Left, the line margin, which the anchor term optimizes: the manipulation check, with a heavier pull expected to raise it. Right, $\bar{\alpha}$ at op1, which the anti-subspace term acts on, with no direction attached to it. The tinted box under each column is the un-anchored control's seed range, scored on that column's subspace: on the axis for the axis conditions and on the plane (`control-plane`) for the plane ones. An unsigned two-dimensional alignment sits higher than a signed one-dimensional one for every state, so the two subspaces compare against different boxes.*

What we saw. H4 held: every statistic is inside its gate at every condition, and none comes near one. The non-red deficit on the plane is two to eight times the axis's, in the direction predicted, and still a factor of three under its gate; it falls with the weight rather than rising. So the useful range of the recipe on this grammar reaches at least $\lambda_a = 0.28$, and where it ends is still open.

The manipulation check is the finding of this section. The line margin rises with the weight on both subspaces, but by very little: from 0.432 to 0.445 on the axis and from 0.405 to 0.415 on the plane, about three percent for an anchor 2.8 times heavier. The weight reaches the model, and the quantity it optimizes is close to saturated at the recipe's weight already. That is the likely reason H1 and H2 had so little to respond to: between 0.1 and 0.28 the recipe sits on a plateau of the thing it trains. $\bar{\alpha}$ at op1 falls with the weight on the axis, from 0.277 to 0.216, which is the anti-subspace term climbing beside the pull, and barely moves on the plane.

> **Pass**
>
> Every cost statistic is inside its gate at every rung, and the line margin rises with the weight, by about three percent.

The adoption rule

The rule, frozen before the run:

> A condition qualifies when the one-sided 95% upper confidence bound on its seed-mean kept share sits under the 0.2 gate on `hue-hsv` and on each of the ten other ops, the task, the line margin, the lead at the embedding, the contrast between red and non-red, the grading r^2, and the deficit on the non-red lines are inside ex-2.2.11's gates on the seed mean, and its $\bar{\alpha}$ at op1 stands in no higher ratio to the un-anchored baseline for its own subspace than the reference stands to the axis one. Among those that qualify, the one with the lowest upper bound on `hue-hsv` is adopted, and the smaller λ_a breaks a tie. The rule sees the measurements it names; if a qualifying condition looks harmful on something it does not name, the report says what and keeps the reference, and that override is marked as a decision made after the data. As in ex-2.2.12, this experiment proposes: the adopted recipe is confirmed at fresh seeds by the anchored-op experiment that inherits it, and no number quoted here for it is free of the selection. If none qualifies, the recipe stays at `axis-0.1` and the report carries the ladder's best characterization of the leftover — its level, its spread, and which lines it is made of — for the anchored-op preregistration to treat as a bounded confound.

What changed from the rule in ex-2.2.12, and why. Ex-2.2.12 asked instead for a seed mean under the gate by at least a fixed band of 0.06, the spread its reference showed at twenty seeds. That band is a statement about one condition's resolution, and applying it to a condition three times tighter charges it for a spread it does not have. The upper bound above asks the same question — is this condition's leftover under the gate, or does the seed spread leave that unsettled — of each condition at its own precision. It is the stricter rule on a wide condition and the looser one on a tight one, and it is fixed here before the runs exist. Both verdicts are reported.

condition	upper bound, `hue-hsv` ↓	upper bound, worst other op ↓	cost gates	$\bar{\alpha}$ at op1	$\bar{\alpha}$ ratio ↓	qualifies	under ex-2.2.12's band (0.14)	$\bar{\alpha}$ excess ↓	qualifies, as intended
`axis-0.1`	0.313	0.218 (`darken`)	all pass	0.277	**-10.27**	no	no	**0.30**	no
`axis-0.14`	0.256	**0.184 (`lighten`)**	all pass	0.258	-9.56	no	no	**0.28**	no
`axis-0.2`	0.333	**0.170 (`lighten`)**	all pass	0.263	-9.76	no	no	**0.29**	no
`axis-0.28`	0.263	**0.151 (`lighten`)**	all pass	0.216	-8.02	no	no	**0.24**	no
`plane-0.1`	**0.199**	**0.192 (`lighten`)**	all pass	0.300	2.66	no	no	**0.19**	**yes**
`plane-0.14`	0.236	**0.167 (`lighten`)**	all pass	0.314	2.77	no	no	**0.20**	no
`plane-0.2`	0.228	**0.171 (`lighten`)**	all pass	0.300	2.65	no	no	**0.19**	no
`plane-0.28`	0.272	0.203 (`lighten`)	all pass	0.287	2.54	no	no	**0.17**	no

The adoption rule clause by clause. The upper bound is the one-sided 95% bound on the seed mean, and the gate it is read against is 0.2. The $\bar{\alpha}$ ratio is $\bar{\alpha}$ at op1 over the un-anchored baseline for the condition's own subspace, and it passes when it is no higher than the reference's -10.27. Those two baselines are -0.027 on the axis, where the alignment is signed and the control's five seeds straddle zero, and +0.113 on the plane, where it is unsigned. The last column applies ex-2.2.12's fixed band to the seed mean instead of the upper bound, with the same cost and $\bar{\alpha}$ clauses. The last two columns read the $\bar{\alpha}$ clause as intended, as an excess over the condition's own baseline, passing when it is no higher than the reference's 0.30; that reading was adopted at review, after the data.

To the letter, no condition qualifies under either rule. Under the fixed band from ex-2.2.12 nothing comes close: the lowest seed mean is `plane-0.1` at 0.170, against a band that asks for 0.14.

Under the rule frozen here, `plane-0.1` clears every clause about the leftover and the costs. Its upper bound on `hue-hsv` is 0.199 and the upper bound on its worst other op is 0.192. Both are under the 0.2 gate, and every cost statistic passes. The one clause it fails is the $\bar{\alpha}$ clause, and it fails on the arithmetic of that clause. The clause takes the $\bar{\alpha}$ of each condition as a ratio to the un-anchored baseline of its own subspace, then compares that with the ratio of the reference to the axis baseline. The axis baseline is a signed cosine, and it averages -0.027 over the five control seeds, so the ratio for the reference is -10.3. The plane baseline is an unsigned length, 0.113, so every plane ratio is a small positive number, and no positive number is below a negative one. In raw terms the $\bar{\alpha}$ of the plane conditions runs about 0.29 to 0.30, against 0.22 to 0.28 on the axis. The difference is modest, and the ratio turns it into an impossible one.

The clause was written badly, so we read it as it was meant. Its purpose was to catch a condition that pulls the other colors toward the home of *red* more than the reference does, and the ratio was an attempt in the prereg to put the two subspaces on one scale. The comparison ex-2.2.12 made across subspaces, and the one the clause was reaching for, is the excess over the baseline of each subspace. On that reading `plane-0.1` sits 0.19 above its own baseline where the reference sits 0.30 above the axis one, so it passes. The last column of the table applies that reading to every condition, and `plane-0.1` is the one condition that qualifies. Reading a frozen rule by its intent is a decision made after the data, so the table keeps the literal column beside it and a reader can weigh the two.

We keep the recipe at `axis-0.1` all the same. That is a second decision made after the data, on grounds the rule does not name. The plane lowers the mean of the leftover by about 0.10. In return the concept takes two coordinates of the stream instead of one, the non-red deficit is a few times the one on the axis, and every anchored-op experiment that follows inherits a two-dimensional home for a hue that one dimension holds well enough on ten ops of eleven. That is a small gain at a standing cost. What the ladder has settled is that the plane works, and by how much, so it is a validated alternative for a concept that turns out to need more room than an axis. If it is preregistered, the $\bar{\alpha}$ clause should be stated as an excess.

Exploratory analyses

Four measurements are planned as descriptions rather than tests: they carry no gate, and no finding rests on them. Anything we think of after seeing the data goes here too, marked as post hoc.

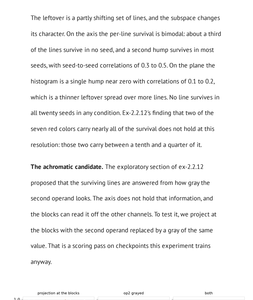
The spread of everything else. H1 scores the spread of one statistic, The same seed-spread table for every cost statistic, the line margin, and the task_per condition, says whether a heavier anchor makes training as a whole more reproducible on this grammar or only settles the leftover. It carries no gate because we have no prediction for the direction of most of them: a heavier pull could hold the margin to a tighter value across seeds, or could amplify whatever differs between initializations. Ex-2.2.12 gives a hint that the two spreads can move apart: one step up the axis, the line margin and Δ at op1 were the tightest in the sweep while the kept share stayed as wide as the reference, and on the plane at the same step it was the other way round.

condition	kept share, the hue-hsv	the task margin	the line margin	the leak at the embedding	the contrast between red and non-red	the grading r^2	the deficit on the non-red lines
axis-0.1	0.1364 +1.00	0.0054 +1.00	0.0207 +1.00	0.0078 +1.00	0.0088 +1.00	0.0225 +1.00	0.0025 +1.00
axis-0.14	0.1056 +0.91	0.0050 +0.87	0.0214 +1.05	0.0070 +0.88	0.0115 +1.31	0.0212 +0.94	0.0045 +1.84
axis-0.2	0.1069 +0.92	0.0050 +0.87	0.0164 +0.79	0.0106 +1.36	0.0176 +2.00	0.0458 +2.03	0.0023 +0.91
axis-0.28	0.0901 +0.77	0.0061 +1.26	0.0258 +1.25	0.0097 +1.24	0.0094 +1.07	0.0189 +0.84	0.0022 +0.91
plane-0.1	0.0745 +0.64	0.0067 +1.56	0.0209 +0.92	0.0071 +0.92	0.0106 +1.21	0.0194 +0.86	0.0009 +0.59
plane-0.14	0.0972 +0.84	0.0053 +0.91	0.0125 +0.60	0.0058 +0.75	0.0128 +1.46	0.0184 +0.82	0.0050 +3.05
plane-0.2	0.0961 +0.83	0.0043 +0.25	0.0141 +0.68	0.0068 +0.87	0.0099 +1.27	0.0320 +1.42	0.0040 +1.62
plane-0.28	0.1567 +1.57	0.0058 +1.09	0.0215 +0.75	0.0053 +0.69	0.0126 +1.44	0.0298 +1.32	0.0049 +1.98

Seed standard deviation of every statistic, per condition, with its ratio to the reference's beside it. H1 scores the first column only; the rest are here to say whether a heavier anchor settles training as a whole or only the leftover. No gate is read off this table.

A heavier anchor does not settle training as a whole. Relative to the reference, the line margin's spread tightens on the plane and not on the axis, the contrast and the grading r^2 widen at $\lambda_a = 0.2$ on both subspaces, and the task's spread roughly doubles at `axis-0.28` and `plane-0.1`. The hint from ex-2.2.12, that the spread of the kept share and the spread of the margin move independently, holds up here.

Which lines are left. Ex-2.2.12 found that nearly all of the `hue-hsv` survival comes from two of the seven red colors, but it scored conditions rather than individual lines. Scoring per line across the ladder tells us whether a narrow seed spread means the same lines survive every time, which would let us characterize the leftover, or a shifting set that happens to be the same size.



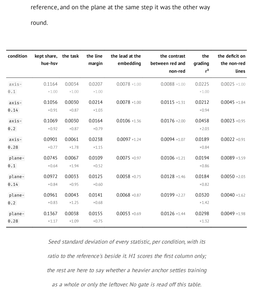
How often each `hue-hsv` removal line survives, per condition. One histogram per condition over the removal lines: a line at 0 is answered differently in every seed after the projection, and a line at 1 keeps its answer in all 20. A leftover made of the same lines each time piles up at the two ends; a leftover of a shifting set fills the middle. Post hoc, no gate.

condition	lines scored	never survives	survives in every seed	mean seed-to-seed r	share from the two reddest op2 colors
axis-0.1	553	34%	0%	0.50	11%
axis-0.14	553	35%	0%	0.51	17%
axis-0.2	553	31%	0%	0.50	11%
axis-0.28	553	41%	0%	0.44	11%
plane-0.1	553	20%	0%	0.53	11%
plane-0.14	553	16%	0%	0.21	21%
plane-0.2	553	22%	0%	0.21	10%
plane-0.28	553	10%	0%	0.11	25%

The `hue-hsv` leftover line by line. A line counts as surviving in a seed when it keeps more than half of its clean expected exact match. The correlation is the mean over every pair of seeds of the correlation between their per-line kept vectors: near one means the same lines survive every time. The last column is the share of all surviving weight carried by lines whose second operand is one of the two reddest palette colors, which is where ex-2.2.12 found the survival. Post hoc.

The leftover is a partly shifting set of lines, and the subspace changes its character. On the axis the per-line survival is bimodal: about a third of the lines survive in no seed, and a second hump survives in most seeds, with seed-to-seed correlations of 0.5 to 0.5. On the plane the histogram is a single hump near zero with correlations of 0.1 to 0.2, which is a thinner leftover spread over more lines. No line survives in all twenty seeds in any condition. Ex-2.2.12's finding that two of the seven red colors carry nearly all of the survival does not hold at this resolution: those two carry between a tenth and a quarter of it.

The achromatic candidate. The exploratory section of ex-2.2.12 proposed that the surviving lines are answered from how gray the second operand looks. The axis does not hold that information, and the blocks can read it off the other channels. To test it, we project at the blocks with the second operand replaced by a gray of the same value. That is a scoring pass on checkpoints this experiment trains anyway.



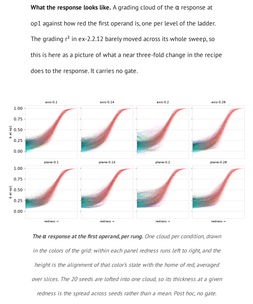
The achromatic edit on the `hue-hsv` removal lines. *Each panel is a kept share against the unedited line's own answer: the projection at the blocks, the gray substitution on its own, and the two together. If the surviving answers are read off how achromatic the second operand looks, graying it leaves them in place; if they need that operand's own hue, it does not. Post hoc, no gate.*

condition	projection at the blocks	op1 grayed	both	both − blocks	both − grayed
axis-0.1	0.271	0.696	0.680	+0.409	-0.016
axis-0.14	0.214	0.706	0.689	+0.475	-0.017
axis-0.2	0.328	0.703	0.694	+0.365	-0.008
axis-0.28	0.329	0.705	0.695	+0.364	-0.010
plane-0.1	0.294	0.700	0.677	+0.383	-0.022
plane-0.14	0.208	0.698	0.680	+0.472	-0.017
plane-0.2	0.326	0.704	0.693	+0.367	-0.011
plane-0.28	0.255	0.700	0.690	+0.435	-0.012

Seed-mean kept share under each edit, over the removal-op1 removal lines, all read against the unedited line's own answer. Graying the second operand moves about three answers in ten on its own, so the two difference columns say different things: the first is what graying adds to the projection, and the second is what the projection still costs once the operand is already gray. Post hoc.

Graying the second operand is a large edit on its own, moving about thirty percent of the answers with no projection at all. Against that, projecting at the blocks costs almost nothing once the operand is gray, where the same projection with the operand intact costs most of the answers. So on these lines the whole effect of the removal runs through the second operand's hue, and once that hue is gone there is nothing left for the projection to take. The achromatic candidate as ex-2.2.12 phrased it, an answer read from how gray the operand looks, is not what this shows; the hue of the second operand is doing the work, through channels the anchored subspace does not hold.

What the response looks like. A grading cloud of the α response at op1 against how red the first operand is, one per level of the ladder. The grading r^2 in ex-2.2.12 barely moved across its whole sweep, so this is here as a picture of what a near three-fold change in the recipe does to the response. It carries no gate.



The α response at the first operand, per rung. *One cloud per condition, drawn in the colors of the grid: within each panel, redness runs left to right, and the height is the alignment of that color's state with the home of red, averaged over slices. The 20 seeds are lofted into one cloud, so its thickness at a given redness is the spread across seeds rather than a mean. Post hoc, no gate.*

The curve is a sigmoid in redness at every rung, and the clouds differ in texture. At $\lambda_a = 0.1$ the cloud is smooth on both subspaces. At 0.2 it is noisier, with more low-alignment strays through the middle of the redness range. At 0.28 it is stepped, and the middle of the curve separates into bands. The rung used by the recipe, `axis-0.1`, has the cleanest response of the eight. This is the plateau in the line margin seen from the other side: the mean curve barely moves as the weight rises, and what the weight changes is the texture across seeds.

Discussion

The ladder was built on the premise that we had been running near the bottom of the useful range, so a heavier anchor would give the downstream measurements something to respond to. What it shows instead is a plateau. Between λ_a 0.1 and 0.28 the line margin moves by about three percent, every cost stays well inside its gate, and the leftover on `hue-hsv` neither shrinks nor settles. On this grammar the weight is not the lever we thought it was, so the recipe can stay where it is, with some confidence that the choice does not matter much.

The subspace *is* a lever, but a small one. At the weight used by the recipe, the plane lowers the leftover by about 0.10. It costs a non-red deficit a few times the one on the axis, still far under the gate. It qualifies under the adoption rule read as intended, and the anchored-op experiments inherit `axis-0.1` anyway, for the reasons the adoption section gives. The plane will most likely stay on the shelf. What this ladder adds is that it is a validated shelf: if a concept turns up that a single axis cannot hold, we know the plane works and we know what it costs.

For those experiments the leftover is a bounded confound rather than a fixed one. On the axis the kept share is near a quarter, with a standard deviation of about 0.12 across seeds, and it comes from no fixed set of lines: about a third of the lines never survive, a second group survives in most seeds, and none survives in every seed. Any experiment that meets it will have to measure it at its own seeds, which at twenty seeds resolves a shift of about 0.05. Two questions stay open: where the useful range of the weight ends, and whether τ trades against it the way the survey in ex-2.1.11 found. Neither is needed before the anchored-op work goes ahead.

Method

Fresh seeds

Ex-2.2.11 and ex-2.2.12 trained at model seeds 100–119. The observation this experiment follows up was read off those checkpoints, and a rule written in advance does not make data we have already seen unseen, so no anchored checkpoint here is served from the store: every anchored condition, the reference included, trains at seeds 200–219. The four conditions that earlier experiments saw appear in the H1 table beside their earlier values, as a replication rather than as data.

We do not hold seeds back to quote the adopted condition from. The search is eight conditions, so the selection bias on the winner is small, and ex-2.2.12 set the pattern this experiment follows: the ladder proposes, and the anchored-op experiment that inherits the recipe confirms it at seeds of its own. Deciding the rule on half the seeds would halve the precision of the upper bound it rests on, for a correction the next experiment pays anyway.

The plane

As ex-2.2.12 defined it. The home of *red* is the first two axes of the stream together. Alignment is the length of the projection of a state onto that pair, and the anchor term pulls that length toward one on the labelled lines. The anti-subspace term is the square of that length over every live position, and the removal projects the whole plane out.

So the concept holds two coordinates of sixty-four rather than one, and its share of the variance counts both.

Budget

160 runs at d64-L4, each as long as a run in ex-2.2.11, at about three minutes a run on an L4. Scoring adds the eleven ops under the operators from ex-2.2.11, the per-line pass, and the achromatic edit. That is about three and a half times the sweep in ex-2.2.12, which cost six dollars on Modal.

What this experiment does not vary

Depth, the pooling temperature τ, and the anti-subspace ratios stay at the values from ex-2.2.11, which means the anti-subspace weight itself climbs with the ladder. Ex-2.2.12 moved each of them one step and resolved nothing on any measurement, so a third factor here would cost runs without a prediction behind it. The weight above 0.28 is also left alone: the ladder tests the region ex-2.1.11's map calls useful, and finding its edge on this grammar is a different experiment.

The survey in ex-2.1.11 found that the weight and τ trade against each other, so a $\lambda_a \times \tau$ factorial would be the natural follow-up if this ladder finds a level worth having.

---

1. The *residual stream* is the running vector of activations that each layer of a transformer reads from and writes back to. ↩

2. The *trend contrast* regresses each run's absolute distance from the median of its own condition on log λ_a. Working from the median rather than the mean keeps a couple of extreme seeds from deciding it, and asking for a slope rather than for any difference between the levels means a spread that rose and then fell does not count as a pass. ↩