A TROPICAL GEOMETRY VIEW OF FORGETTING: A PER-UNIT PROJECTOR FOR KNOWLEDGE-PRESERVING FINE-TUNING

Anonymous authors
Paper under double-blind review

ABSTRACT

Fine-tuning a language model on new text degrades what it already does. Replay-free projectors such as Adam-NSCL and GPM forbid one shared subspace of a layer's inputs in every row of the update. The tropical geometry of a ReLU layer shows why this is too coarse. In data space, the units' walls are tropical hypersurfaces whose cells are dual to the upper vertices of a zonotope; in weight space, each old token is a hyperplane, and the tokens cut out a polyhedron, the closure of the weights that keep every token on its side. An exact identity joins the two pictures: the squared change of the layer's output under any weight change splits into in-cell, open→closed and closed→open terms, and the first two live on the tokens each unit fires on (its open tokens). The identity names a gate-aware per-unit projector, and a budget-separation theorem prices exact protection: it costs a unit the rank of its own open tokens, while a shared subspace pays at least the rank of their union in every row. On OPT-1.3b, where 96% of (token, unit) pairs are closed, the projector forgets less than Adam-NSCL at all six matched budgets from 9 to 60 constrained directions per row (24 of 27 seed-pairs, $p = 1.5 \times 10^{-6}$), the gap widening from $1.1\times$ to $4.3\times$; with $1/5.5$ of the directions it halves the forgetting of Adam-NSCL at GPM's energy threshold. On OPT-6.7b, it matches Adam-NSCL's forgetting at matched budget while learning more. As the theory predicts, the open/closed partition is the operative variable: open tokens beat random, sign-blind and anti-gate token sets on 18 of 18 seed-pairs and are equivalent to the first-order criterion, which inherits the partition. In pruning repair, the identity shows that every derivative-based local model of the output error at the dense weights is blind to pairs that open: the minimisers of the gate-weighted objective can leave the polyhedron, the objective's closed-form solution is 1.94 nats worse than no repair on OPT-1.3b, and a convex one-sided penalty bounds the escape.

1 INTRODUCTION

A language model fine-tuned on new text loses some of what it could already do. A leading family of replay-free remedies constrains the update geometrically: Adam-NSCL (Wang et al., 2021) and GPM (Saha et al., 2021) project each layer's update onto the approximate null space of the layer's old inputs; in language models, the same subspace orients or initialises low-rank adapters (Yang et al., 2024; Tang et al., 2026), and AlphaEdit (Fang et al., 2025) projects edits of the MLP output matrix with it. All of them share one object, a single input subspace per layer, forbidden in *every* row of the update.

On a matrix whose outputs pass through a ReLU, that object is too coarse. A unit that is closed on a token contributes exactly zero there, and keeps contributing zero until its pre-activation crosses the threshold; in OPT-1.3b, 96% of (token, unit) pairs are closed, and a shared subspace constrains all of them. ReLU-family language models remain a live design point because this sparsity can be exploited at inference (Mirzadeh et al., 2024; Song et al., 2025).

We start from the geometry (§2). In data space, a ReLU layer's walls are tropical hypersurfaces and its gate patterns are dual to the upper vertices of a zonotope; dually, in weight space, the old tokens cut out a polyhedron, the closure of the weights that keep every token on its side. An exact

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

identity (Theorem 5) splits the squared change in a layer's output under *any* weight change into in-cell, open→closed and closed→open terms, with no remainder and no reference to an objective; the first two live on the tokens each unit fires on. From it we derive (§3) a gate-aware per-unit projector: after each step it holds each unit's pre-activation fixed on up to C of the old tokens it fires on and leaves its closed tokens free. A budget-separation theorem prices this exact protection: it costs a unit the rank of its own open tokens, while a shared subspace pays at least the rank of their union in every row (Theorem 9). The theory makes six predictions (P1–P6, end of §3), which organise the experiments.

The measurements bear them out (§4). On OPT-1.3b, the per-unit projector forgets less than Adam-NSCL at all six matched budgets from 9 to 60 constrained directions per row (24 of 27 seed-pairs, $p = 1.5 \times 10^{-6}$), the gap widening from $1.1\times$ to $4.3\times$; with $1/5.5$ of the directions it halves the forgetting of Adam-NSCL at GPM's energy threshold. On OPT-6.7b, it matches Adam-NSCL's forgetting at matched budget and learns more. At equal per-unit dimension, open tokens beat random, sign-blind and anti-gate token sets on 18 of 18 seed-pairs and are equivalent to the first-order criterion. In pruning repair, every derivative-based local model of the output error at the dense weights is blind to the closed→open term, so the minimisers of the gate-weighted objective can leave the polyhedron (the objective's closed-form solution is 1.94 nats worse than no repair on OPT-1.3b), and a one-sided penalty bounds the escape (§5).

Contributions.

- The two dual pictures of a threshold-activation layer, made rigorous: the tropical form and its dual subdivision, the activation-preserving polyhedron, and an exact three-term decomposition of the output change under any weight change (§2).
- A per-unit projector derived from them, with exactness, budget-separation and first-order-inheritance results, a Ky Fan characterisation of the best per-unit and shared constraints, and six predictions that the experiments test (§§3–4).
- For pruning repair: curvature blindness, polyhedral escape and confinement by a convex one-sided penalty (§5).

2 Two dual pictures of a ReLU layer

A ReLU layer can be read in two spaces (Figure 1). In *data space* its walls tile the inputs into cells, one per gate pattern, dual to the upper vertices of a zonotope. In *weight space* each unit is a point and each old token a hyperplane, and the tokens cut out a polyhedron, the closure of the weights that keep every token on its side. Fine-tuning moves the point: inside it the output changes only on the tokens the unit fires on, and crossing a hyperplane switches one token's gate. The identity below makes the two kinds of damage exact; we use it to decide which motion to forbid, and which to price.

Notation. A layer has m units on inputs $x \in \mathbb{R}^d$. We write $\tilde{x} = (x, 1)$ and $\tilde{w}_i = (w_i, b_i) \in \mathbb{R}^{d+1}$, so $z_i(x) = \langle \tilde{w}_i, \tilde{x} \rangle$ carries the bias as the weight of a ones column. The old tokens $\mathcal{X} = \{x_t\}_{t=1}^T$ (a multiset) give $\tilde{X} \in \mathbb{R}^{T\times(d+1)}$ with rows $\tilde{x}_t$; $\tilde{X}_S$ keeps the rows indexed by S.

Definition 1 (Wall and cells)**.** Let $\mathrm{act}_\theta(u) = u \cdot \mathbb{1}[u > \theta]$ with $\theta \geq 0$. The **wall** of a unit with weights $\tilde{w}$ is $\mathcal{W}_{\tilde{w}} = \{x : z_{\tilde{w}}(x) = \theta\}$. It splits $\mathbb{R}^d$ into the **open cell** $\mathcal{C}_> = \{z_{\tilde{w}} > \theta\}$ and the **closed cell** $\mathcal{C}_\leq = \{z_{\tilde{w}} \leq \theta\}$, on each of which the unit is affine. Unit i's **open set** is $O_i = \{t : z_i(x_t) > \theta\}$, and $G(x) = \{i : z_i(x) > \theta\}$ is the **gate pattern** of x.

Data space. In the max-plus semiring ($a \oplus b = \max(a, b)$, $a \odot b = a + b$) a ReLU unit is the tropical binomial $0 \oplus z$, whose wall is the tropical hypersurface on which its two terms tie (Zhang et al., 2018; Maclagan & Sturmfels, 2015). The sum of the units is their tropical product, with one term per set J of units, and its lifted Newton polytope is a zonotope. Write $\tilde{w}_J^\theta = \sum_{i\in J}(w_i, b_i - \theta)$ and $Z_\theta = \sum_{i=1}^m [0, (w_i, b_i - \theta)]$.

Proposition 2 (Tropical form and the dual subdivision)**.** ***(i)*** $\sum_i \max(0, z_i(x) - \theta) = \max_{J\subseteq[m]} \langle \tilde{w}_J^\theta, \tilde{x} \rangle$*, and J attains the maximum iff $G(x) \subseteq J \subseteq \{i : z_i(x) \geq \theta\}$; so $G(x)$ always attains it, uniquely iff x lies on no wall.* ***(ii)*** *$J \mapsto \tilde{w}_J^\theta$ maps the gate patterns realised off the*

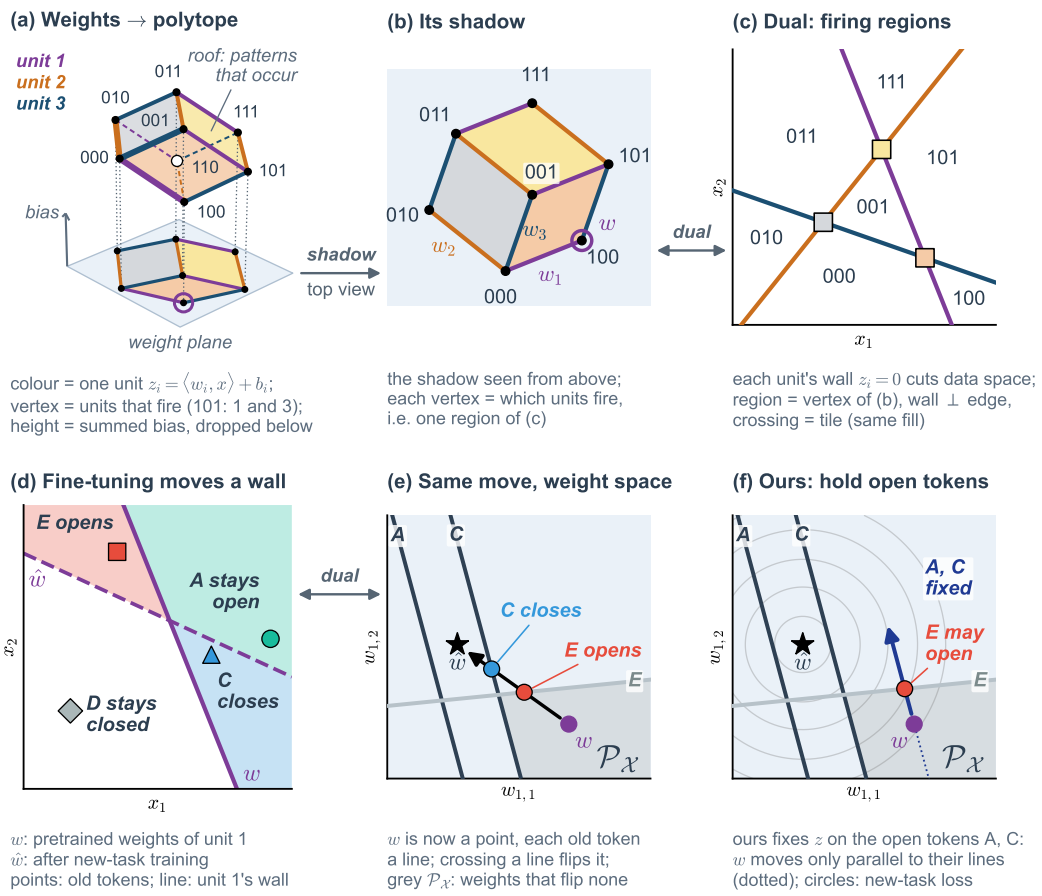


Figure 1: **A layer's weights, its lifted Newton polytope, and its activation pattern**, for a toy layer of three ReLU units on two inputs. Pale blue marks weight space and (c), (d) are data space; arrows marked *dual* join the two. **(a)** Each colour is a unit and each vertex a set of units that fire, in bits (101: units 1 and 3), placed at their summed weights and raised by their summed biases: the Newton polytope of $\sum_i \max(0, z_i)$ lifted by the biases (Zhang et al., 2018) (Appendix A). **(b)** Its shadow, which subdivides the Newton polytope; the circled vertex is unit 1's weights w. **(c)** The units' walls $z_i = 0$ in data space. Regions are labelled by which units fire, each is a vertex of (b), and pattern 110 (hollow in (a)) occurs nowhere. **(d)** Fine-tuning moves unit 1's wall from w to $\hat{w}$, and colours say what that does to each old token: green stays open, blue closes, red opens, grey stays closed (the sets $S_{++}, S_{+-}, S_{-+}, S_{--}$ of Theorem 5). **(e)** The same move in weight space: w is a point and each old token a line, crossing which flips the token (dark: open at w; light: closed; D's line is off-panel). The grey region $\mathcal{P}_\mathcal{X}$ is the closure of the weights that flip no token (Definition 3). **(f)** Our projector (§3) fixes the pre-activation on the open tokens A and C, so w moves only parallel to their lines (dotted) and the closed token E may open; circles are the new-task loss. Biases are fixed in (d)–(f), and A, C lie on one ray, so one direction stays free.

*walls injectively into the vertices of Z_θ exposed by functionals $(y, 1)$, and onto them when no unit has $(w_i, b_i) = (0, \theta)$. **(iii)** If x, x' lie on no wall and $G(x') = G(x) \sqcup \{i\}$, the two vertices span an edge of Z_θ parallel to $(w_i, b_i - \theta)$, dual to a nonempty relatively open piece of the wall $\mathcal{W}_i$.*

At $\theta = 0$ this is Figure 1a–c; for $\theta > 0$ the gate pattern is that of $\max(0, z_i - \theta)$, so the proposition describes its cells verbatim.

Weight space. Fix one unit with base weights $\tilde{w}$. Each token defines the hyperplane $H_t = \{\tilde{v} : \langle \tilde{v}, \tilde{x}_t \rangle = \theta\}$ of the unit's weight space ($\tilde{x}_t \neq 0$), and the pairing is symmetric: x_t lies on the wall of $\tilde{v}$ exactly when $\tilde{v}$ lies on H_t.

Definition 3 (Activation-preserving polyhedron)**.** With O the base open set of the unit on $\mathcal{X}$, let

$$\mathcal{P}_{\mathcal{X}} = \big\{\tilde{v} : \langle \tilde{v}, \tilde{x}_t \rangle \geq \theta \ \ \forall t \in O, \quad \langle \tilde{v}, \tilde{x}_t \rangle \leq \theta \ \ \forall t \notin O\big\}, \tag{1}$$

and let the **pattern cell** $\mathcal{K}_{\mathcal{X}} \subseteq \mathcal{P}_{\mathcal{X}}$ be the weights that give every token its base gate bit. The first group of constraints is the **active block**, the second the **inactive block**.

Proposition 4 (Data–weight duality)**.** ***(a)*** *Token t is open at $\tilde{v}$ iff $\tilde{v} \in H_t^{>} = \{\langle \cdot, \tilde{x}_t \rangle > \theta\}$. It changes cell between $\tilde{w}$ and $\hat{w}$ iff exactly one of them lies in $H_t^{>}$, closing if that one is $\tilde{w}$, and along the segment $[\tilde{w}, \hat{w}]$ it changes cell at most once.* ***(b)*** *$\mathcal{K}_{\mathcal{X}} = \mathcal{P}_{\mathcal{X}} \setminus \bigcup_{t \in O} H_t$ is convex, its interior is a chamber of the arrangement $\{H_t\}$, and $\overline{\mathcal{K}_{\mathcal{X}}} = \mathcal{P}_{\mathcal{X}}$.* ***(c)*** *$\mathcal{P}_{\mathcal{X}}$ is full-dimensional, and each of its facets is $\mathcal{P}_{\mathcal{X}} \cap H_t$ for a token t, an active facet if $t \in O$ and an inactive one if not.*

Part (a) is Figure 1d–e: moving a unit's wall across a token in data space and moving its weight point across that token's hyperplane are one event. Proofs of every statement in §§2–3 are in Appendix B.

Theorem 5 (Exact decomposition by cell membership)**.** *For any $\hat{w} \in \mathbb{R}^{d+1}$, any $\theta \geq 0$ and finite $\mathcal{X}$, with $L_{\text{post}} := \sum_{x \in \mathcal{X}} (\text{act}_\theta(z_{\hat{w}}(x)) - \text{act}_\theta(z_{\tilde{w}}(x)))^2$,*

$$L_{\text{post}} = \underbrace{\sum_{x \in S_{++}} \big(z_{\hat{w}}(x) - z_{\tilde{w}}(x)\big)^2}_{\textit{in-cell}} + \underbrace{\sum_{x \in S_{+-}} z_{\tilde{w}}(x)^2}_{\textit{open}\to\textit{closed}} + \underbrace{\sum_{x \in S_{-+}} z_{\hat{w}}(x)^2}_{\textit{closed}\to\textit{open}}, \tag{2}$$

where $S_{++} = \mathcal{X} \cap \mathcal{C}_> \cap \hat{\mathcal{C}}_>$, $S_{+-} = \mathcal{X} \cap \mathcal{C}_> \cap \hat{\mathcal{C}}_\leq$ and $S_{-+} = \mathcal{X} \cap \mathcal{C}_\leq \cap \hat{\mathcal{C}}_>$, which with S_{--} partition $\mathcal{X}$.

Proof. Each x lies in exactly one of the four intersections. On S_{++} both activations equal their pre-activations; on S_{+-} the new one is 0; on S_{-+} the old one is 0; on S_{--} both are 0. Summing the squared differences over the four sets gives Eq. (2). □

Summed over units, the four sets become sets of (token, unit) pairs (t, i). The pattern-changing terms are exact sums over pair sets fixed by the two weight points, where convergence analyses bound them as a remainder (Du et al., 2019). Measured, the two sides agree to a relative gap of at most 5.8×10^{-5} over 27 OPT repair cells and 9.0×10^{-5} over 18 cells on two 7B ReLU-family models (Song et al., 2025; SparseLLM Team, 2023) (Appendix J).

Corollary 6 (Weight-space reading)**.** *$S_{+-} \cup S_{-+}$ are the tokens whose hyperplanes separate $\tilde{w}$ from $\hat{w}$ in the sense of Proposition 4a, and S_{+-} those with $\tilde{w}$ on the open side. L_{post} is quadratic on every set of weights with a fixed gate pattern on $\mathcal{X}$; on $\mathcal{K}_{\mathcal{X}}$ it equals $\|\tilde{X}_O(\hat{w} - \tilde{w})\|^2$, and at $\theta = 0$ it does on all of $\mathcal{P}_{\mathcal{X}}$.*

Moving inside $\mathcal{K}_{\mathcal{X}}$ costs only the in-cell term, and each separating token adds exactly one crossing term. The two applications act on the two blocks: knowledge-preserving fine-tuning holds the active block on the tokens it covers (exactly at $\rho = 0$, to within the ridge residual of Proposition 8a at $\rho > 0$, at the base inputs) and leaves the inactive block free (§3); in pruning repair the curvature objectives leave the inactive block unpriced, and pricing it bounds the escape (§5).

3 The constraint the identity names, and what it predicts

By Theorem 5 a pair closed at base enters only the closed→open term, and contributes zero while it stays closed. At the base inputs the in-cell and open→closed terms of unit i are carried by O_i alone, and both vanish when row i's update satisfies $\tilde{X}_{O_i} u_i = 0$: one equality per open token, and which tokens are open is a property of the unit, not of the layer.

Definition 7 (Gate-aware per-unit projector)**.** Let $\tilde{X}$ hold the base model's inputs to the layer on the T-token *old support*. For unit i let $h_t^{(i)} = \text{act}_\theta(z_i(x_t))$, let the *cap* C bound each unit's budget, let $S_i \subseteq O_i$ index the C largest positive $h_t^{(i)}$ (ties broken by token index), and

$$k_i = |S_i| = \min\big(C, |O_i|\big), \qquad A_i = \tilde{X}_{S_i} \in \mathbb{R}^{k_i \times (d+1)}. \tag{3}$$

After each optimiser step the cumulative displacement $u_i = (dW_i, db_i)$ of row i is replaced by

$$u_i \ \leftarrow \ P_i(\rho_i)\, u_i = u_i - A_i^\top (A_i A_i^\top + \rho_i I)^{-1} A_i u_i, \tag{4}$$

with $P_i(0) := I - A_i^+ A_i$, its limit as $\rho_i \to 0$, and $P_i(\rho) = I$ when $k_i = 0$.

Everything in Eq. (4) is built once from the base model, with ρ_i 10^{-4} times the mean diagonal of $A_i A_i^\top$ padded to $C \times C$ by the identity (Appendix C). At $\rho = 0$ row i loses $\operatorname{rank} A_i \le k_i$ directions (at $\rho > 0$ they are shrunk, not removed), so the reported dims/row, the mean k_i, bounds the directions removed from above. On OPT-1.3b at caps 16 to 128 the residual on the constrained pairs, at the stored base inputs, is 5.4×10^{-9} to 4.0×10^{-8} of the layer's total $|\Delta z|$ mass. All `fc1` layers train at once, so end-to-end numbers also include the drift of each layer's inputs; the statements below concern one layer at its base inputs.

Proposition 8 (Exactness). *Let $u' = P_i(\rho)u$. **(a)** $P_i(0)$ is the orthogonal projector onto $\ker A_i$. For $\rho > 0$ and $k_i \ge 1$, $\|A_i u'\| \le \frac{\rho}{\sigma_+^2 + \rho}\|A_i u\|$, tightly, where σ_+ is the smallest positive singular value of A_i. **(b)** At $\rho = 0$, every pair in S_i stays in S_{++} and contributes zero to all three terms, every pair closed at base contributes only to closed→open, and unit i's in-cell and open→closed terms are sums over $O_i \setminus S_i$, empty when $C \ge |O_i|$. **(c)** If $C \ge |O_i|$ and $\rho = 0$, the flat $\tilde{w}_i + \ker \tilde{X}_{O_i}$ lies in $\bigcap_{t \in O_i} H_t^{>}$: the weight point moves parallel to every active hyperplane, and can leave $\mathcal{K}_{\mathcal{X}}$ only through inactive facets (Figure 1f).*

Theorem 9 (Budget separation). *Let $r_i = \operatorname{rank} \tilde{X}_{O_i}$, and call a subspace U of row updates* protecting *for unit i if unit i's in-cell term vanishes for every update in U. **(a)** U is protecting for unit i iff $U \subseteq \ker \tilde{X}_{O_i}$, and then the open→closed term vanishes on U too; so protection costs at least $r_i \le |O_i|$ directions, and with $C \ge |O_i|$ the projector $P_i(0)$ removes exactly r_i. **(b)** One subspace shared by every row is protecting for all units iff it lies in $\ker \tilde{X}_{O_\cup}$, $O_\cup = \bigcup_i O_i$, so it removes at least $\operatorname{rank} \tilde{X}_{O_\cup} \ge \max_i r_i$ directions from every row. **(c)** With B directions in total, per-unit subspaces protect any set J of units with $\sum_{i \in J} r_i \le B$, while freezing whole rows protects $\#\{i : r_i = 0\} + \min(\lfloor B/(d+1) \rfloor, \#\{i : r_i > 0\})$ units at best.*

In OPT-1.3b with $T = 2\,048$ old tokens and $d + 1 = 2\,049$, the median unit fires on 11 tokens and the mean on 72.1, so $r_i \le 11$ for at least half the units and 19.5% never fire. Caps 64 and 128 give mean k_i of 23.5 and 34.4 and cover 59.6% and 71.7% of the active post-activation energy. Adam-NSCL instead forbids the top k_{sh} eigendirections of $\tilde{X}^\top \tilde{X}/T$ in every row.

Proposition 10 (Budget-limited optimum). *Let $\theta = 0$ and let each row's proposed update u_i satisfy $\mathbb{E}[u_i u_i^\top] = \sigma^2 I$, projected onto U_i of codimension $q \le d + 1$. Unit i's in-cell plus open→closed terms are at most $\|\tilde{X}_{O_i} P_{U_i} u_i\|^2$, whose expectation $D_i(U_i) = \sigma^2 \operatorname{tr}(\Xi_i P_{U_i})$, $\Xi_i = \tilde{X}_{O_i}^\top \tilde{X}_{O_i}$, has minimum $\sigma^2 \sum_{j>q} \lambda_j(\Xi_i)$, attained when $U_i^\perp$ is a top-q eigenspace of the unit's own open-token Gram Ξ_i. One subspace shared by all rows is best when $U^\perp$ is a top-q eigenspace of $\sum_i \Xi_i = \sum_t n_t \tilde{x}_t \tilde{x}_t^\top$, n_t the number of units open on token t, and its total $\sum_i D_i$ is never below that of the best per-unit subspaces of the same total codimension.*

Proposition 11 (The freed term is a near-wall term). *For a closed token t of unit i and an update u, let $\gamma_t = \theta - z_i(x_t) \ge 0$ and $a_t = \langle u, \tilde{x}_t \rangle$. Token t opens iff $a_t > \gamma_t$, so only the near-wall tokens $N(u) = \{t \notin O_i : \gamma_t < \|u\| \, \|\tilde{x}_t\|\}$ can open, and at $\theta = 0$ the closed→open term is $\sum_{t \notin O_i} (a_t - \gamma_t)_+^2 \le \|\tilde{X}_{N(u)} u\|^2$.*

Proposition 12 (First-order criteria inherit the partition). *Let the inputs be fixed and a loss $L = g(h)$ be differentiable in the post-activations $h = (h_{t,i})$, $h_{t,i} = \mathrm{act}_\theta(z_{t,i})$ (any downstream network), with $\mathrm{act}'_\theta(\theta) := 0$ as in automatic differentiation. Then $\partial L / \partial z_{t,i} = \mathbb{1}[z_{t,i} > \theta] \, \partial g / \partial h_{t,i}$ vanishes on every closed pair; the row gradient $\nabla_{\tilde{w}_i} L = \sum_{t \in O_i} (\partial L / \partial z_{t,i}) \, \tilde{x}_t$ lies in the row space of $\tilde{X}_{O_i}$; and the Fisher of row i annihilates $\ker \tilde{X}_{O_i}$ (with O_i taken over the Fisher data), while every term of EWC's diagonal Fisher is indexed by an open pair.*

Which matrices carry the relaxation. On an ungated matrix every pair acts as open, so the exact per-unit constraint of every row is the shared null space $\ker \tilde{X}$, which Adam-NSCL approximates; the two exact constraints coincide there and differ only on a matrix whose outputs each pass through a threshold, in OPT `fc1`, where we compare the methods. In a ReLU-gated GLU the same constraint on the gate and up rows keeps every open pair's output (Lemma 21).

Predictions. The results fix what §4 measures. **P1** (Theorem 9, Proposition 10): exact protection costs a unit r_i directions of its own row but a shared subspace $\operatorname{rank} \tilde{X}_{O_\cup}$ in every row, and in the damage bound D_i the best per-unit constraint of each codimension is never worse than the best

shared one. The gate projector attains $D_i = 0$ once its tokens span the row space of $\tilde{X}_{O_i}$; below that the theory does not rank it against Adam-NSCL, and we predict that at matched dims/row it forgets less once the cap covers most of each unit's open-token span (§4.1). **P2** (Propositions 8 and 12): at the base inputs a covered open pair is held exactly and a covered closed pair only against opening, so we predict that at fixed k_i sets chosen within O_i protect comparably, sets chosen off it lose, and a first-order criterion, which stays inside O_i whenever at least k_i open pairs have $\partial g/\partial h \neq 0$, should match the gate; among the published families EWC is the one whose criterion reads it (§§4.2, 4.3). **P3** (Proposition 11, $\theta = 0$): at the base inputs only near-wall closed tokens $N(u)$ can open, and their term is at most $\|\tilde{X}_{N(u)}u\|^2$, so we predict that a constrained update frees far less closed→open mass than an unconstrained one. Measured end to end on the probe it is 1 785 at $C = 64$ and 1 622 for the shared subspace of the same size, against 9 239 (seeds 0–2, Appendix C.1); the theory does not order the two constrained methods. **P4** (Proposition 8b): exactness is reached unit by unit at $C \geq |O_i|$, so where open sets are larger the per-unit advantage should need a larger cap (§4.1). **P5** (Theorem 9c): a frozen row spends $d + 1$ directions on one unit, so protecting whole units should be weaker at equal budget (§4.2). **P6** (Lemma 13, Theorem 14): every derivative-based local model of the repair error at the dense weights sees open pairs only, so its minimisers should violate inactive constraints of $\mathcal{P}_{\mathcal{X}}$, and pricing them should bound the escape (§5).

4 EXPERIMENTS

Setup. Pre-LN OPT-125m, OPT-1.3b and OPT-6.7b (Zhang et al., 2022) (ReLU, $\theta = 0$; on OPT-6.7b the `fc1` of the last four blocks). Each run is one fine-tuning transition from the pretrained model, the setting of Biderman et al. (2024), CorDA and LoRA-Null (Yang et al., 2024; Tang et al., 2026). The knowledge to preserve is the pretraining distribution, represented by MiniPile (Kaddour, 2023), a subset of the Pile, parts of which OPT was pretrained on; it is scored by MiniPile loss, as in Ibrahim et al. (2024), and by five zero-shot tasks. A 2 048-token support builds the projectors, and a disjoint 1 024-token probe measures forgetting. The new task is WikiText-103 (Merity et al., 2017) or Python code from CodeParrot (CodeParrot, 2022). A run trains `fc1` on 65 536 tokens: AdamW (Loshchilov & Hutter, 2019) at lr 10^{-4}, cosine schedule, 128 steps, batch 4. *Forgetting* is the rise in old-probe NLL, *learning* the change in held-out new-task NLL (more negative is better). Runs are paired by seed; we report paired t-tests, sign-flip permutation tests and random-effects meta-analyses (DerSimonian & Laird, 1986) (Appendix C).

4.1 MATCHED BUDGET, CAPACITY AND SCALE (P1, P4)

Theorem 9 prices protection in constrained directions per row, so that is the axis we hold equal: each cap is paired with Adam-NSCL at k_{sh} equal to the rounded mean $\bar{k}$ of its k_i (Figure 2; every row in Table 3). On OPT-1.3b, the per-unit projector forgets less at all six budgets from 9 to 60 dims/row, on 24 of 27 seed-pairs (sign-flip permutation $p = 1.5 \times 10^{-6}$; random-effects difference -0.018 nats, 95% CI $[-0.028, -0.009]$), and Adam-NSCL learns at most 0.003 nats more. The gap widens from $1.1\times$ at 9 dims/row to $4.3\times$ at 60: the shared subspace plateaus at 0.046–0.050 nats between 15 and 47 dims/row (seeds 0–2), while the per-unit projector keeps falling, to 0.010. The reading P1 suggests is that each added shared direction is spent in every row, including on the 96% of pairs that are closed, while each added per-unit direction lands on a token its own unit fires on. OPT-125m's units fire on more tokens (a mean of 111 of 2 048, against 72), and as P4 predicts, the per-unit advantage needs a larger cap there: at 5.7 dims/row, where the cap covers under a tenth of a unit's open tokens, the shared subspace forgets less, and the two meet at 17. From 17 upward the per-unit projector forgets less on 14 of 18 seed-pairs (exact permutation $p = 5 \times 10^{-4}$) and learns more, by 0.023 nats at 17 dims/row and 0.008 at 42.

Capacity. Against Adam-NSCL at a uniform $k_{\mathrm{sh}} = 259$ per row (the mean dimension GPM's energy criterion selects at 0.95), our cap 256 at 47.3 dims/row forgets $2.3\times$ less ($3/3$, $p = 0.009$) and learns 0.052 nats more, and cap 128 learns more and forgets less on WikiText-103 (5 of 6 seeds) and on Python (3 of 3), where every constrained method ends below the base old-probe NLL, so forgetting less means improving the old probe more. Cap 256 removes 95.9% of unconstrained forgetting on OPT-1.3b and 97.7% on OPT-125m (Appendix D.1).

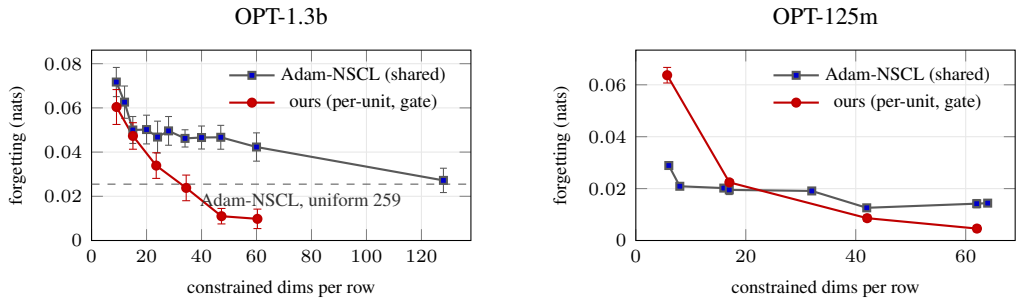


Figure 2: **Forgetting at matched constrained dimensions.** Means over seeds 0–2 $\pm$ one standard error, WikiText-103, lr 10^{-4}; dashed: Adam-NSCL at a uniform 259 dims/row, the mean dimension GPM's energy criterion selects at 0.95.

Table 1: **Identical per-unit dimension, different tokens.** OPT-1.3b, WikiText-103, $\bar{k} = 23.5$, $n = 6$ ($n = 3$ for random-within-open). Gate minus the other selector. Superiority: one-sided paired t, 95% CI; equivalence: TOST (Schuirmann, 1987) at $\delta = 0.0068$ nats, 2.5% of unconstrained forgetting, 90% CI.

other selector	Δ forget	CI	verdict
random tokens (rank-matched)	-0.0317	$[-0.0436, -0.0197]$	gate better, $p = 0.0005$
sign-blind $\lvert z - \theta \rvert$	-0.0575	$[-0.0756, -0.0393]$	gate better, $p = 0.0002$
anti-gate (most closed)	-0.0660	$[-0.0799, -0.0520]$	gate better, $p < 10^{-4}$
first-order $\lvert \partial L/\partial z \rvert$	$+0.0024$	$[-0.0011, +0.0059]$	equivalent
random within the open set	$+0.0015$	$[-0.0021, +0.0051]$	equivalent

Scale. On OPT-6.7b, cap 128 constrains 68.7 dims/row and forgets the same as Adam-NSCL at $k_{\text{sh}} = 69$ (0.0192 against 0.0184, equivalent under TOST), and it learns 0.011 nats more (3/3 seeds, paired $p = 1.5 \times 10^{-4}$). Both remove 85% of the unconstrained forgetting of 0.126.

4.2 What the criterion reads (P2, P5)

To isolate the criterion, we hold each unit's k_i fixed (cap 64, $\bar{k} = 23.5$) and change only which tokens fill it (Table 1). The unit's own open tokens beat rank-matched random tokens, a sign-blind ranking by $|z - \theta|$ and the most closed tokens, on all 18 of 18 seed-pairs (permutation $p = 3.8 \times 10^{-6}$). The random-token contrast holds again on OPT-125m (-0.0114, $p = 0.005$), on Python (-0.0224, $p = 0.0009$) and on OPT-6.7b at cap 128 (-0.0098, 3/3, $p = 0.007$).

The two equivalences are the ones P2 names. Random tokens drawn *within* each unit's open set protect as well as its most active ones, although the two sets differ on 30% of pairs and cover 35.2% against 59.6% of the active energy. The criterion is therefore the partition itself, not a ranking inside it. The first-order criterion agrees because it inherits the partition (Proposition 12): 99.6% of the pairs $|\partial L/\partial z|$ selects are open. The gate reads the same set off one forward pass, with no gradient. As P5 predicts, protecting whole units is far weaker: freezing the `fc1` rows of the units with the largest old-support activity, at the same 23.5 dims/row, forgets 0.210 against the per-unit projector's 0.025 (6/6, $p < 10^{-4}$) and learns 0.039 nats less. Freezing enough rows to hold the same 59.6% of the active energy costs 69 dims/row and still forgets 0.172 (Appendix E.1).

4.3 Against the published method families

Figure 3 places the per-unit projector among the replay-free families on the forgetting–learning plane, each swept over its own hyperparameter on `fc1`. EWC (Kirkpatrick et al., 2017), whose diagonal Fisher reads open pairs only (Proposition 12), is dominated from $\lambda = 4$ to 4×10^4 and reaches the per-unit frontier at $\lambda = 4 \times 10^5$ (0.0175, between our caps 128 and 256; no paired difference significant). MIGU (Du et al., 2024), which masks rows by output magnitude on the new batch, forgets 2$\times$ more than our cap 128 at its default mask ratio 0.7 and learns 0.051 nats less

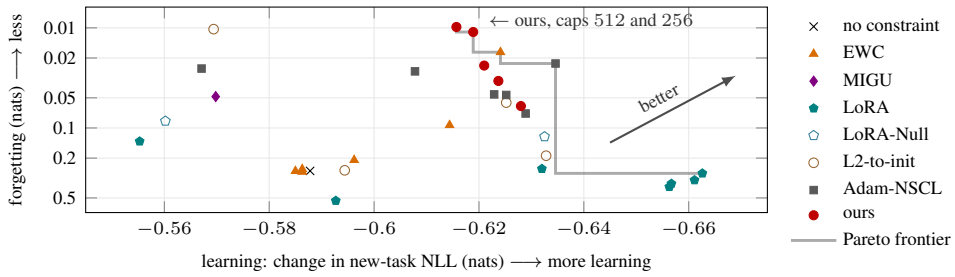


Figure 3: **Method families on the forgetting–learning plane.** OPT-1.3b, WikiText-103, means over seeds 0–2. Axes reversed: up and right is better. Each marker is one setting of its family's own hyperparameter. Grey steps: the Pareto frontier of the plotted points, whose low-forgetting end is our caps 512 and 256. Settings, and five points off the plotted range, each learning less than unconstrained fine-tuning, are in Appendix F.1.

($p = 0.0005$). LoRA (Hu et al., 2022), to which O-LoRA (Wang et al., 2023) reduces at the first task, never comes within $5\times$ of our forgetting; the four configurations that learn significantly more than ours, by 0.035–0.042 nats, forget 12–16$\times$ more. LoRA-Null (Tang et al., 2026), which initialises the adapter in the null space of the old inputs, cuts LoRA's forgetting by 2.3–3$\times$ and still forgets 3.6–5.1$\times$ more than our cap 128 (dominated at rank 32). L2-to-init, EWC with a unit Fisher, forgets at $\lambda = 0.3$ as little as our cap 256 (TOST-equivalent) and learns 0.049 nats less ($3/3$, $p = 0.02$). Adam-NSCL with its dimension set by GPM's energy criterion (mean 259 per row) learns 0.014 nats more than our cap 128 at equal forgetting; the per-unit projector matches that forgetting with $1/7.5$ of the dimensions and halves it with $1/5.5$ (-0.0117, $3/3$), while learning 0.016 nats less (Appendix F). Post-hoc interpolation toward the base, WiSE-FT (Wortsman et al., 2022), composes with both projectors: interpolated, every constrained run learns more at zero forgetting than the interpolated unconstrained run (0.027–0.043 nats, $3/3$ seeds for each method; Appendix F.2).

Downstream accuracy. The protection measured in old-probe NLL carries over to zero-shot accuracy on LAMBADA, HellaSwag, ARC-easy, Winogrande and SciQ (Paperno et al., 2016; Zellers et al., 2019; Clark et al., 2018; Sakaguchi et al., 2020; Welbl et al., 2017) (Table 9; LM Evaluation Harness, Gao et al., 2024). Unconstrained fine-tuning on WikiText-103 costs OPT-1.3b 2.94 points of mean accuracy, and our cap 128 costs 1.02 ($3/3$ seeds, $p = 0.013$); at matched budget the two projectors agree to within 0.1 points.

5 Second application: repair after pruning

One-shot pruning zeroes a fraction of each weight matrix under a mask, such as Wanda's saliency mask (Sun et al., 2024), and SparseGPT (Frantar & Alistarh, 2023) then *repairs* the surviving weights so that the layer reproduces its old output. We repair only the wall matrices (`fc1`, `gate_proj`); attention and the output matrix stay mask-only (Appendix I). We call $\mathrm{Esc}(\hat{w}) := \sum_{x \in \mathcal{X} \cap \mathcal{C}_{\leq}} \max(z_{\hat{w}}(x) - \theta, 0)^2$, the squared amount by which a repaired row's closed calibration tokens overshoot the threshold, its *escape* from $\mathcal{P}_{\mathcal{X}}$: it is zero exactly when every inactive constraint holds, and at $\theta = 0$ it equals the closed→open term of Eq. (2). Here a row is $w \in \mathbb{R}^d$ with its bias fixed, $\mathcal{X}$ is the calibration multiset, $\mathcal{P}_{\mathcal{X}}$ is the slice of Definition 3 at that bias, M is the mask with kept coordinates K, $\mathcal{Q}_M$ is the set of masked $\hat{w}$ that satisfy every inactive constraint, and $\zeta = (\langle w_{K^c}, x_{K^c} \rangle)_{x \in \mathcal{X} \cap \mathcal{C}_>}$ is the mask's residual.

Lemma 13 (Curvature blindness). *Assume no calibration token lies on the wall, $z_w(x) \neq \theta$ for all $x \in \mathcal{X}$. Then $L_{\mathrm{post}} = L_{\mathrm{gated}}$ on a neighbourhood of w, where the gate-weighted objective $L_{\mathrm{gated}}(\hat{w}) = \sum_{x \in \mathcal{X} \cap \mathcal{C}_>} (z_{\hat{w}}(x) - z_w(x))^2$ is the in-cell term evaluated as if no token changed cell. Hence every local model of L_{post} at w built from its derivatives, of any order (the second-order Taylor model, the Gauss–Newton matrix, the Fisher of a Gaussian output model), is the corresponding model of L_{gated}; both crossing terms vanish identically on the neighbourhood.*

Theorem 14 (Polyhedral escape). *Write $X_K = [A; \tilde{A}]$ for the kept columns of the calibration inputs, split into active and inactive rows.* ***(i)*** *L_{gated} depends on the active tokens alone, through A*

*and the residual ζ. **(ii)** If $\ker A \not\subseteq \ker \tilde{A}$, in particular if X_K has full column rank and there are fewer active tokens than kept columns, the minimiser set of L_{gated} is an affine subspace on which L_{gated} is constant and the escape is unbounded above. **(iii)** For n i.i.d. calibration tokens with finite second moments, a fixed mask and a non-singular active second moment on the kept coordinates, the minimiser is almost surely unique for large n and converges to a point $w^\star$ the active tokens determine, and its per-token escape converges to $e^\star = \mathbb{E}\big[\mathbb{1}[z_w(x) \le \theta]\,(z_{w^\star}(x) - \theta)_+^2\big]$, which is positive if and only if $\Pr[z_w(x) \le \theta < z_{w^\star}(x)] > 0$.*

Theorem 15 (Confinement by a one-sided penalty)**.** *For $\lambda > 0$, $L_{\mathrm{gated_h}}(\hat{w}, \lambda) = L_{\mathrm{gated}}(\hat{w}) + \lambda\,\mathrm{Esc}(\hat{w})$ over masked $\hat{w}$ is convex and attains its minimum; its penalty is zero exactly on $\mathcal{Q}_M \supseteq \mathcal{P}_{\mathcal{X}} \cap \{\hat{w} = \hat{w} \odot M\}$; and every $\hat{w}$ with $L_{\mathrm{gated_h}}(\hat{w}, \lambda) \le L_{\mathrm{gated_h}}(w \odot M, \lambda)$, every minimiser in particular, has escape at most $\|\zeta\|^2/\lambda + E_0$, where $\|\zeta\|^2 = L_{\mathrm{gated}}(w \odot M)$ and $E_0 = \mathrm{Esc}(w \odot M)$ is the escape of the mask alone. The escape is the same at every minimiser, is non-increasing in λ, and tends to 0 as $\lambda \to \infty$ if and only if $\mathcal{Q}_M \neq \varnothing$.*

Proofs are in Appendices G and H (Remark 24 covers the implemented penalty at $\theta > 0$). The bias witness $b \le \theta$ certifies $\mathcal{Q}_M \neq \varnothing$ on 81–96% of the wall units measured (Remark 23). Solved in closed form in `fp64`, L_{gated} is 1.94 nats worse than no repair (OPT-1.3b, sparsity 0.7). Its crossings run closed→open (1.15% of pairs against 0.18% open→closed), as its unpriced inactive block predicts; the pre-activation objective L_{pre} runs the other way (0.52% against 0.34%; Appendix G.3). Pricing both crossings orders the objectives: $L_{\mathrm{post}} < L_{\mathrm{pre}}$ on 46 of 47 probe cells, $15/15$ on WikiText-2 (Merity et al., 2017) and $20/20$ on two 7B ReLU-family models, and a price $\lambda > 1$ beats $\lambda = 1$ on ReLU, and on GELU with the exact objective (Pythia, Biderman et al., 2023; Appendix L).

6 Related work

Projection-based continual learning. GPM (Saha et al., 2021), Adam-NSCL (Wang et al., 2021) and AdNS (Kong et al., 2022) confine each layer's update to the complement of a shared subspace of its old inputs, OGD (Farajtabar et al., 2020) does so in parameter space, later work relaxes the subspace (Lin et al., 2022; Saha & Roy, 2023; Liang & Li, 2024), and in language models it orients or initialises adapters (Yang et al., 2024; Tang et al., 2026); sparse activations enlarge the null space (Abbasi et al., 2022). We change which pairs are constrained.

Other continual-learning methods. O-LoRA (Wang et al., 2023) keeps task adapters orthogonal, LoRA learns less and forgets less (Biderman et al., 2024), and MIGU (Du et al., 2024), HAT, AGS-CL (Serrà et al., 2018; Jung et al., 2020) and neuron-level editors (Jiang et al., 2024; Pan et al., 2025) decide per unit whether it may move; EWC (Kirkpatrick et al., 2017), LDIFS (Mukhoti et al., 2024) and replay (Ibrahim et al., 2024) complete the set.

Activation patterns and repair. Layer-wise repair after pruning or quantisation (Frantar & Alistarh, 2023; Frantar et al., 2023; Li et al., 2021; Kim et al., 2025) fits pre-activations or curvature-weighted outputs. For piecewise-linear activations, the Hessian's diagonal blocks equal the Gauss–Newton blocks (Botev et al., 2017); Lemma 13 extends this to the objectives near w. The inequalities of $\mathcal{P}_{\mathcal{X}}$ appear in provable repair (Tao et al., 2023), Net-Trim (Aghasi et al., 2017) and Pilanci & Ergen (2020), and our penalty is the inactive-side term of the activation-boundary loss of Heo et al. (2019) at zero margin.

Tropical geometry of ReLU networks. ReLU networks are tropical rational maps (Zhang et al., 2018; Maragos et al., 2021) whose number of linear regions Montúfar et al. (2014) bound; tropical compression approximates their Newton polytopes from the weights (Smyrnis & Maragos, 2020; Misiakos et al., 2022). We use the data-space subdivision and its weight-space dual.

7 Conclusion

An exact identity joins a ReLU layer's two dual pictures. In fine-tuning, it names the per-unit constraint; in pruning repair, it names the block that curvature cannot see.

AI USE STATEMENT

The authors designed the study and every experiment. We used a large language model for the following. *Implementation and checking:* it refined and debugged the code for the method, the baselines, the experiments and the analyses *Retrieval and discovery:* it searched for and summarised related work. *Writing:* it helped polish the text, refine the proof. The authors take full responsibility for all content.

REPRODUCIBILITY STATEMENT

The supplementary material contains the implementation of the gate-aware per-unit projector (Definition 7), of the shared-subspace baseline, of the decomposition of Eq. (2) and of the repair objectives of §5, a CPU self-test of Eq. (2), Definition 7, Proposition 8 and Theorem 15 on random layers, and numerical checks of the statements of §§2, 3 and 5. The complete training and evaluation code, the per-run histories and the scripts that regenerate every table and figure will be released with the paper. Every run logs its fully resolved arguments and checks Eq. (2) against its own activations. Models, corpora, training settings and statistics are given in Appendices C and I.

REFERENCES

Ali Abbasi, Parsa Nooralinejad, Vladimir Braverman, Hamed Pirsiavash, and Soheil Kolouri. Sparsity and heterogeneous dropout for continual learning in the null space of neural activations. In *Conference on Lifelong Learning Agents (CoLLAs)*, volume 199 of *PMLR*, pp. 617–628, 2022.

Alireza Aghasi, Afshin Abdi, Nam Nguyen, and Justin Romberg. Net-trim: Convex pruning of deep neural networks with performance guarantee. In *Advances in Neural Information Processing Systems*, volume 30, 2017.

Dan Biderman, Jacob Portes, Jose Javier Gonzalez Ortiz, Mansheej Paul, Philip Greengard, Connor Jennings, Daniel King, Sam Havens, Vitaliy Chiley, Jonathan Frankle, Cody Blakeney, and John P. Cunningham. LoRA learns less and forgets less. *Transactions on Machine Learning Research*, 2024.

Stella Biderman, Hailey Schoelkopf, Quentin Anthony, Herbie Bradley, Kyle O'Brien, Eric Hallahan, Mohammad Aflah Khan, Shivanshu Purohit, USVSN Sai Prashanth, Edward Raff, Aviya Skowron, Lintang Sutawika, and Oskar van der Wal. Pythia: A suite for analyzing large language models across training and scaling. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*. PMLR, 2023.

Aleksandar Botev, Hippolyt Ritter, and David Barber. Practical Gauss-Newton optimisation for deep learning. In *International Conference on Machine Learning (ICML)*, 2017. arXiv:1706.03662.

Peter Clark, Isaac Cowhey, Oren Etzioni, Tushar Khot, Ashish Sabharwal, Carissa Schoenick, and Oyvind Tafjord. Think you have solved question answering? try ARC, the AI2 reasoning challenge, 2018. arXiv:1803.05457.

CodeParrot. CodeParrot dataset cleaned, validation split. `https://huggingface.co/datasets/codeparrot/codeparrot-clean-valid`, 2022. Python files from public GitHub repositories, deduplicated and filtered.

Rebecca DerSimonian and Nan Laird. Meta-analysis in clinical trials. *Controlled Clinical Trials*, 7 (3):177–188, 1986.

Simon S. Du, Xiyu Zhai, Barnabás Póczos, and Aarti Singh. Gradient descent provably optimizes over-parameterized neural networks. In *International Conference on Learning Representations (ICLR)*, 2019. arXiv:1810.02054.

Wenyu Du, Shuang Cheng, Tongxu Luo, Zihan Qiu, Zeyu Huang, Ka Chun Cheung, Reynold Cheng, and Jie Fu. Unlocking continual learning abilities in language models. In *Findings of the Association for Computational Linguistics: EMNLP*, 2024.

Junfeng Fang, Houcheng Jiang, Kun Wang, Yunshan Ma, Shi Jie, Xiang Wang, Xiangnan He, and Tat-seng Chua. AlphaEdit: Null-space constrained knowledge editing for language models. In *International Conference on Learning Representations (ICLR)*, 2025. arXiv:2410.02355.

Mehrdad Farajtabar, Navid Azizan, Alex Mott, and Ang Li. Orthogonal gradient descent for continual learning. In *International Conference on Artificial Intelligence and Statistics (AISTATS)*, 2020. arXiv:1910.07104.

Marguerite Frank and Philip Wolfe. An algorithm for quadratic programming. *Naval Research Logistics Quarterly*, 3(1–2):95–110, 1956.

Elias Frantar and Dan Alistarh. SparseGPT: Massive language models can be accurately pruned in one-shot. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*. PMLR, 2023.

Elias Frantar, Saleh Ashkboos, Torsten Hoefler, and Dan Alistarh. GPTQ: Accurate post-training quantization for generative pre-trained transformers. In *International Conference on Learning Representations*, 2023.

Leo Gao, Stella Biderman, Sid Black, Laurence Golding, Travis Hoppe, Charles Foster, Jason Phang, Horace He, Anish Thite, Noa Nabeshima, Shawn Presser, and Connor Leahy. The Pile: An 800GB dataset of diverse text for language modeling. *arXiv preprint arXiv:2101.00027*, 2020.

Leo Gao, Jonathan Tow, Baber Abbasi, Stella Biderman, Sid Black, Anthony DiPofi, Charles Foster, Laurence Golding, Jeffrey Hsu, Alain Le Noac'h, Haonan Li, Kyle McDonell, Niklas Muennighoff, Chris Ociepa, Jason Phang, Laria Reynolds, Hailey Schoelkopf, Aviya Skowron, Lintang Sutawika, Eric Tang, Anish Thite, Ben Wang, Kevin Wang, and Andy Zou. The language model evaluation harness, 2024. Zenodo record of v0.4.3; we use v0.4.12.

Dan Hendrycks and Kevin Gimpel. Gaussian error linear units (GELUs). *arXiv preprint arXiv:1606.08415*, 2016.

Byeongho Heo, Minsik Lee, Sangdoo Yun, and Jin Young Choi. Knowledge transfer via distillation of activation boundaries formed by hidden neurons. In *AAAI Conference on Artificial Intelligence*, 2019. arXiv:1811.03233.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations (ICLR)*, 2022.

Adam Ibrahim, Benjamin Therien, Kshitij Gupta, Mats L. Richter, Quentin Anthony, Timothée Lesort, Eugene Belilovsky, and Irina Rish. Simple and scalable strategies to continually pre-train large language models. *Transactions on Machine Learning Research (TMLR)*, 2024.

Houcheng Jiang, Junfeng Fang, Tianyu Zhang, An Zhang, Ruipeng Wang, Tao Liang, and Xiang Wang. Neuron-level sequential editing for large language models, 2024. arXiv:2410.04045.

Sangwon Jung, Hongjoon Ahn, Sungmin Cha, and Taesup Moon. Continual learning with node-importance based adaptive group sparse regularization. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2020. arXiv:2003.13726.

Jean Kaddour. The MiniPile challenge for data-efficient language models. *arXiv preprint arXiv:2304.08442*, 2023.

Jinuk Kim, Marwa El Halabi, Wonpyo Park, Clemens Js Schaefer, Deokjae Lee, Yeonhong Park, Jae W. Lee, and Hyun Oh Song. GuidedQuant: Large language model quantization via exploiting end loss guidance. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 30011–30037. PMLR, 2025.

James Kirkpatrick, Razvan Pascanu, Neil Rabinowitz, Joel Veness, Guillaume Desjardins, Andrei A. Rusu, Kieran Milan, John Quan, Tiago Ramalho, Agnieszka Grabska-Barwinska, Demis Hassabis, Claudia Clopath, Dharshan Kumaran, and Raia Hadsell. Overcoming catastrophic forgetting in neural networks. *Proceedings of the National Academy of Sciences*, 114(13):3521–3526, 2017.

Yajing Kong, Liu Liu, Zhen Wang, and Dacheng Tao. Balancing stability and plasticity through advanced null space in continual learning. In *European Conference on Computer Vision (ECCV)*, 2022. arXiv:2207.12061.

Mark Kurtz, Justin Kopinsky, Rati Gelashvili, Alexander Matveev, John Carr, Michael Goin, William Leiserson, Sage Moore, Nir Shavit, and Dan Alistarh. Inducing and exploiting activation sparsity for fast inference on deep neural networks. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*. PMLR, 2020.

Yuhang Li, Ruihao Gong, Xu Tan, Yang Yang, Peng Hu, Qi Zhang, Fengwei Yu, Wei Wang, and Shi Gu. BRECQ: Pushing the limit of post-training quantization by block reconstruction. In *International Conference on Learning Representations*, 2021.

Yan-Shuo Liang and Wu-Jun Li. InfLoRA: Interference-free low-rank adaptation for continual learning. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024.

Sen Lin, Li Yang, Deliang Fan, and Junshan Zhang. TRGP: Trust region gradient projection for continual learning. In *International Conference on Learning Representations (ICLR)*, 2022. arXiv:2202.02931.

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *International Conference on Learning Representations (ICLR)*, 2019. arXiv:1711.05101.

Diane Maclagan and Bernd Sturmfels. *Introduction to Tropical Geometry*, volume 161 of *Graduate Studies in Mathematics*. American Mathematical Society, 2015.

Petros Maragos, Vasileios Charisopoulos, and Emmanouil Theodosis. Tropical geometry and machine learning. *Proceedings of the IEEE*, 109(5), 2021.

James Martens and Roger Grosse. Optimizing neural networks with Kronecker-factored approximate curvature. In *International Conference on Machine Learning (ICML)*, volume 37 of *PMLR*, pp. 2408–2417, 2015.

Stephen Merity, Caiming Xiong, James Bradbury, and Richard Socher. Pointer sentinel mixture models. In *International Conference on Learning Representations*, 2017.

Iman Mirzadeh, Keivan Alizadeh, Sachin Mehta, Carlo C Del Mundo, Oncel Tuzel, Golnoosh Samei, Mohammad Rastegari, and Mehrdad Farajtabar. ReLU strikes back: Exploiting activation sparsity in large language models. In *International Conference on Learning Representations*, 2024.

Panagiotis Misiakos, Georgios Smyrnis, George Retsinas, and Petros Maragos. Neural network approximation based on Hausdorff distance of tropical zonotopes. In *International Conference on Learning Representations (ICLR)*, 2022.

Guido Montúfar, Razvan Pascanu, Kyunghyun Cho, and Yoshua Bengio. On the number of linear regions of deep neural networks. In *Advances in Neural Information Processing Systems*, volume 27, 2014.

Jishnu Mukhoti, Yarin Gal, Philip H. S. Torr, and Puneet K. Dokania. Fine-tuning can cripple your foundation model; preserving features may be the solution. *Transactions on Machine Learning Research (TMLR)*, 2024.

Haowen Pan, Xiaozhi Wang, Yixin Cao, Zenglin Shi, Xun Yang, Juanzi Li, and Meng Wang. Precise localization of memories: A fine-grained neuron-level knowledge editing technique for LLMs. In *International Conference on Learning Representations (ICLR)*, 2025. arXiv:2503.01090.

Denis Paperno, Germán Kruszewski, Angeliki Lazaridou, Quan Ngoc Pham, Raffaella Bernardi, Sandro Pezzelle, Marco Baroni, Gemma Boleda, and Raquel Fernández. The LAMBADA dataset: Word prediction requiring a broad discourse context. In *Annual Meeting of the Association for Computational Linguistics (ACL)*, 2016.

Mert Pilanci and Tolga Ergen. Neural networks are convex regularizers: Exact polynomial-time convex optimization formulations for two-layer networks. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pp. 7695–7705. PMLR, 2020.

Gobinda Saha and Kaushik Roy. Continual learning with scaled gradient projection. In *AAAI Conference on Artificial Intelligence*, 2023. arXiv:2302.01386.

Gobinda Saha, Isha Garg, and Kaushik Roy. Gradient projection memory for continual learning. In *International Conference on Learning Representations (ICLR)*, 2021.

Keisuke Sakaguchi, Ronan Le Bras, Chandra Bhagavatula, and Yejin Choi. WinoGrande: An adversarial Winograd schema challenge at scale. In *AAAI Conference on Artificial Intelligence*, 2020.

Donald J. Schuirmann. A comparison of the two one-sided tests procedure and the power approach for assessing the equivalence of average bioavailability. *Journal of Pharmacokinetics and Biopharmaceutics*, 15(6):657–680, 1987.

Joan Serrà, Dídac Surís, Marius Miron, and Alexandros Karatzoglou. Overcoming catastrophic forgetting with hard attention to the task. In *International Conference on Machine Learning (ICML)*, volume 80, pp. 4548–4557, 2018. arXiv:1801.01423.

Noam Shazeer. GLU variants improve transformer. *arXiv preprint arXiv:2002.05202*, 2020.

Georgios Smyrnis and Petros Maragos. Multiclass neural network minimization via tropical Newton polytope approximation. In *International Conference on Machine Learning (ICML)*, volume 119 of *PMLR*, 2020.

Chenyang Song, Xu Han, Zhengyan Zhang, Shengding Hu, Xiyu Shi, Kuai Li, Chen Chen, Zhiyuan Liu, Guangli Li, Tao Yang, and Maosong Sun. ProSparse: Introducing and enhancing intrinsic activation sparsity within large language models. In *Proceedings of the 31st International Conference on Computational Linguistics (COLING)*, 2025. arXiv:2402.13516.

SparseLLM Team. ReluLLaMA-7B. `https://huggingface.co/SparseLLM/ReluLLaMA-7B`, 2023. A ReLU-activated fine-tune of Llama-2-7B.

Mingjie Sun, Zhuang Liu, Anna Bair, and J. Zico Kolter. A simple and effective pruning approach for large language models. In *International Conference on Learning Representations*, 2024.

Pengwei Tang, Xiaolin Hu, Yong Liu, Lizhong Ding, Dongjie Zhang, Xing Wu, and Debing Zhang. Put the space of LoRA initialization to the extreme to preserve pre-trained knowledge. In *AAAI Conference on Artificial Intelligence*, 2026. arXiv:2503.02659.

Zhe Tao, Stephanie Nawas, Jacqueline Mitchell, and Aditya V. Thakur. Architecture-preserving provable repair of deep neural networks. *Proceedings of the ACM on Programming Languages*, 7(PLDI):124, 2023. arXiv:2304.03496.

Shipeng Wang, Xiaorong Li, Jian Sun, and Zongben Xu. Training networks in null space of feature covariance for continual learning. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021. arXiv:2103.07113.

Xiao Wang, Tianze Chen, Qiming Ge, Han Xia, Rong Bao, Rui Zheng, Qi Zhang, Tao Gui, and Xuanjing Huang. Orthogonal subspace learning for language model continual learning. In *Findings of the Association for Computational Linguistics: EMNLP*, 2023.

Johannes Welbl, Nelson F. Liu, and Matt Gardner. Crowdsourcing multiple choice science questions. In *Workshop on Noisy User-generated Text (W-NUT)*, 2017.

Mitchell Wortsman, Gabriel Ilharco, Jong Wook Kim, Mike Li, Simon Kornblith, Rebecca Roelofs, Raphael Gontijo-Lopes, Hannaneh Hajishirzi, Ali Farhadi, Hongseok Namkoong, and Ludwig Schmidt. Robust fine-tuning of zero-shot models. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022.

Yibo Yang, Xiaojie Li, Zhongzhu Zhou, Shuaiwen Leon Song, Jianlong Wu, Liqiang Nie, and Bernard Ghanem. CorDA: Context-oriented decomposition adaptation of large language models for task-aware parameter-efficient fine-tuning. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2024. arXiv:2406.05223.

Rowan Zellers, Ari Holtzman, Yonatan Bisk, Ali Farhadi, and Yejin Choi. HellaSwag: Can a machine really finish your sentence? In *Annual Meeting of the Association for Computational Linguistics (ACL)*, 2019.

Liwen Zhang, Gregory Naitzat, and Lek-Heng Lim. Tropical geometry of deep neural networks. In *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pp. 5824–5832. PMLR, 2018.

Susan Zhang, Stephen Roller, Naman Goyal, Mikel Artetxe, Moya Chen, Shuohui Chen, Christopher Dewan, Mona Diab, Xian Li, Xi Victoria Lin, Todor Mihaylov, Myle Ott, Sam Shleifer, Kurt Shuster, Daniel Simig, Punit Singh Koura, Anjali Sridhar, Tianlu Wang, and Luke Zettlemoyer. OPT: Open pre-trained transformer language models. *arXiv preprint arXiv:2205.01068*, 2022.

A Preliminaries: a ReLU unit as a tropical binomial

In the max-plus notation of Section 2, the summed output of a ReLU layer is the tropical product of its units' binomials $0 \oplus z_i$, and multiplying it out gives one monomial per subset J of units, the case $\theta = 0$ of Proposition 2(i):

$$\sum_i \max(0, z_i) \;=\; \bigodot_i (0 \oplus z_i) \;=\; \max_J \Big(\Big\langle \sum_{i \in J} w_i,\, x \Big\rangle + \sum_{i \in J} b_i \Big). \tag{5}$$

Following Zhang et al. (2018), lift each exponent $\sum_{i \in J} w_i$ by its coefficient $\sum_{i \in J} b_i$. The convex hull of the lifted points is the zonotope $Z_0 = \sum_i [0, (w_i, b_i)]$, and projecting its upper faces gives a subdivision of the weight plane whose vertices are the gate patterns realised off the walls and whose cells are dual to the points where walls meet (Figure 1a–c). With real weights the exponents are real, so the polynomial is strictly a tropical signomial; the duality used here is that of regular subdivisions and holds unchanged. For $\theta > 0$ the activation act_θ is discontinuous and is not tropical; Proposition 2 describes its cells through the gate pattern it shares with $\max(0, z - \theta)$. Figure 4 draws Eq. (5) for a layer with one input, where every object of Figure 1a–c is one dimension lower, and Figure 5 reads Theorem 5 one token at a time on the toy layer of Figure 1.

B Proofs for Sections 2 and 3

Every numbered statement of §§2–3 except Theorem 5, which the body proves, is stated here in full and proved. A full form keeps the part letters of its body statement (an unlettered one is split into parts); parts that appear only here carry a primed letter, as in (a′), or a letter after the body's last one.

All statements concern one layer at fixed inputs. Tokens may repeat, so $\mathcal{X} = \{x_t\}_{t=1}^{T}$ is a finite multiset and $\tilde{X}$ may have equal rows; nothing below assumes distinct tokens, full rank or general position unless stated. Accordingly, the sets O, O_i, S_i, S_{++}, S_{+-}, S_{-+}, S_{--} and $N(u)$ are sets of token indices $t \in [T]$, a sum over $x \in \mathcal{X} \cap A$ runs over the indices t with $x_t \in A$, and $\tilde{X}_S \in \mathbb{R}^{|S| \times (d+1)}$ has one row per index in S; for $S = \varnothing$ it is the empty matrix, of rank 0 and with kernel $\mathbb{R}^{d+1}$. Throughout, d, m and T are positive integers, $\theta \geq 0$ is a real threshold, $[m] = \{1, \dots, m\}$, and $\langle \cdot, \cdot \rangle$ and $\| \cdot \|$ are the Euclidean inner product and norm, with $\| \cdot \|_2$ the induced operator norm of a matrix. Every lifted token $\tilde{x}_t = (x_t, 1)$ has last coordinate 1, so $\tilde{x}_t \neq 0$; we say where this is used. For one unit with base weights $\tilde{w}$ and new weights $\hat{w} = \tilde{w} + u$, $u \in \mathbb{R}^{d+1}$, we write $z_t = \langle \tilde{w}, \tilde{x}_t \rangle$ and $\hat{z}_t = \langle \hat{w}, \tilde{x}_t \rangle = z_t + \langle u, \tilde{x}_t \rangle$; *at the base inputs* means that the $\tilde{x}_t$ are the base model's inputs to the layer and do not change with u. Confining a row's update to a linear subspace $U \subseteq \mathbb{R}^{d+1}$ *removes* $\operatorname{codim} U = d + 1 - \dim U$ directions, and P_U denotes the orthogonal projector onto U. For a symmetric positive semidefinite $\Xi \in \mathbb{R}^{n \times n}$ with eigenvalues $\lambda_1(\Xi) \geq \cdots \geq \lambda_n(\Xi)$, listed with multiplicity, a *top-q eigenspace*, $0 \leq q \leq n$, is the span of q orthonormal eigenvectors whose eigenvalues are $\lambda_1(\Xi), \dots, \lambda_q(\Xi)$.

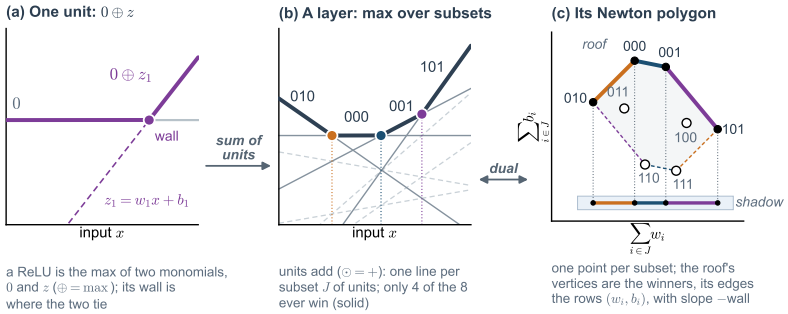


Figure 4: **A ReLU layer as a tropical polynomial, for one input.** The toy layer has three units on a scalar input x, with $z_1 = x - 1.2$, $z_2 = -0.8x - 0.8$ and $z_3 = 0.6x - 0.12$; each unit keeps its colour from Figure 1 (unit 1 purple, unit 2 amber, unit 3 navy). **(a) One unit.** The grey line is the monomial 0 and the dashed purple line the monomial z_1. Their maximum, the thick line, is the ReLU $0 \oplus z_1 = \max(0, z_1)$. The two monomials tie at $x = 1.2$, the unit's wall: to its left the unit outputs 0, to its right it outputs z_1. **(b) The whole layer.** Adding the three outputs is tropical multiplication, $\bigodot_i (0 \oplus z_i)$, and multiplying out gives one line $\sum_{i \in J} z_i$ for each of the 8 subsets J of units (Eq. (5)). The thick envelope is their maximum, which is the layer's summed output. Each piece of the envelope is labelled by the subset that attains it, written as a bit string in which bit i is 1 when unit i is in J, that is, open. Reading left to right: at $x = -1.7$ only unit 2 fires, so the winning line is z_2 and the output is 0.56 (pattern 010); between $x = -1.0$ and $x = 0.2$ no unit fires (000); between 0.2 and 1.2 only unit 3 fires (001); beyond 1.2 units 1 and 3 fire, and at $x = 1.8$ the output is $z_1 + z_3 = 1.56$ (101). The coloured dots are the three walls, where the envelope bends. The four dashed lines never reach the envelope. Pattern 100, for example, would need unit 1 open and unit 3 closed, that is $x > 1.2$ and $x < 0.2$ at once. **(c) The same layer in exponent space.** Each subset J becomes the point $(\sum_{i \in J} w_i, \sum_{i \in J} b_i)$: 000 is the origin, $001 = (0.6, -0.12)$, $101 = (1.6, -1.32)$, and so on. The upper hull of the eight points, the roof, has as its vertices exactly the four patterns that win in (b). Each roof edge is one unit's row. The navy edge from 000 to 001 is unit 3's row $(0.6, -0.12)$, and its slope -0.2 is minus unit 3's wall at $x = 0.2$. In the same way the amber edge has slope 1, minus unit 2's wall at -1.0, and the purple edge has slope -1.2, minus unit 1's wall at 1.2. The hollow points lie below the roof and never win: $100 = (1.0, -1.2)$, for example, sits 0.6 below the roof point above it. The dashed edges bound the polygon from below and carry no winner. Dropping the height (dotted) projects the roof onto the weight axis and splits it into three pieces, one per unit, which is the one-input counterpart of the subdivision in Figure 1b. The arrow marked *dual* records that each roof vertex is a region of (b) and each roof edge a wall.

For a set $V \subseteq \mathbb{R}^n$, $\operatorname{conv} V$ and $\operatorname{aff} V$ are its convex and affine hulls and $\dim V = \dim \operatorname{aff} V$; the relative interior of a convex set is its interior within its affine hull. For a linear functional ℓ on $\mathbb{R}^n$ and a compact convex set $P \subseteq \mathbb{R}^n$, the *face of P exposed by ℓ* is $F_\ell(P) = \arg\max_{p \in P} \ell(p)$; a *vertex* is an exposed face that is a single point, and an *edge* is an exposed face that is a segment of positive length. For a polyhedron P, possibly unbounded, a *face* is a set $P \cap \{y : c^\top y = \beta\}$ for an inequality $c^\top y \le \beta$ valid on P, and a *facet* is a face of dimension $\dim P - 1$. A *chamber* of a finite family of hyperplanes is a connected component of the complement of their union.

Lemma 16. *Let $V \subseteq \mathbb{R}^n$ be finite and nonempty, $P = \operatorname{conv} V$, ℓ a linear functional on $\mathbb{R}^n$ and $M = \max_{v \in V} \ell(v)$. Then $\max_P \ell = M$ and $F_\ell(P) = \operatorname{conv}\{v \in V : \ell(v) = M\}$.*

Proof. Let $p \in P$. Since $P = \operatorname{conv} V$, there are weights $\lambda_v \ge 0$, $v \in V$, with $\sum_v \lambda_v = 1$ and $p = \sum_v \lambda_v v$. By linearity of ℓ,

$$M - \ell(p) = \sum_{v \in V} \lambda_v \big(M - \ell(v)\big),$$

a sum of non-negative terms. Hence $\ell(p) \le M$, with equality if and only if $\lambda_v = 0$ for every v with $\ell(v) < M$. The points $v \in V$ with $\ell(v) = M$ lie in P, so $\max_P \ell = M$. If $\ell(p) = M$, the representation of p uses only points of $V_M = \{v \in V : \ell(v) = M\}$, so $p \in \operatorname{conv} V_M$; conversely, every point of $\operatorname{conv} V_M$ lies in P and, by linearity, has value M. Therefore $F_\ell(P) = \operatorname{conv} V_M$. $\square$

Lemma 17. *Let $a_1, \dots, a_m \in \mathbb{R}^n$ and, for $J \subseteq [m]$, $a_J = \sum_{i \in J} a_i$, so that $a_\varnothing = 0$. Then the Minkowski sum $\sum_{i=1}^m [0, a_i]$ equals $\operatorname{conv}\{a_J : J \subseteq [m]\}$.*

Proof. Write $Z = \sum_{i=1}^m [0, a_i]$. A Minkowski sum of convex sets is convex, and Z contains every a_J (take the i-th summand at a_i for $i \in J$ and at 0 otherwise), so $\operatorname{conv}\{a_J\} \subseteq Z$. Conversely, let $p \in Z$, so that $p = \sum_i \lambda_i a_i$ with every $\lambda_i \in [0, 1]$. Order the indices as $(1), \dots, (m)$ so that $\lambda_{(1)} \geq \cdots \geq \lambda_{(m)}$, put $\lambda_{(0)} = 1$ and $\lambda_{(m+1)} = 0$, and let $J_k = \{(1), \dots, (k)\}$ for $0 \leq k \leq m$. We claim that

$$p = \sum_{k=0}^{m} \big(\lambda_{(k)} - \lambda_{(k+1)}\big)\, a_{J_k}.$$

Indeed, $a_{(j)}$ occurs in a_{J_k} exactly when $k \geq j$, so its coefficient on the right is $\sum_{k=j}^m (\lambda_{(k)} - \lambda_{(k+1)}) = \lambda_{(j)} - \lambda_{(m+1)} = \lambda_{(j)}$. The weights $\lambda_{(k)} - \lambda_{(k+1)}$ are non-negative, because $1 \geq \lambda_{(1)}$, the $\lambda_{(k)}$ are sorted and $\lambda_{(m)} \geq 0$, and they sum to $\lambda_{(0)} - \lambda_{(m+1)} = 1$. Hence $p \in \operatorname{conv}\{a_J\}$. □

Lemma 18. *Let K be a positive integer, let $g_1, \dots, g_K : \mathbb{R}^n \to \mathbb{R}$ be affine and non-constant, let $P = \{y \in \mathbb{R}^n : g_t(y) \leq 0,\ t = 1, \dots, K\}$, and assume that P has nonempty interior.*

***(i)** The boundary of P is $\partial P = \bigcup_t (P \cap \{g_t = 0\})$, and the interior points of P are the points with $g_t < 0$ for every t.*

***(ii)** Every facet of P equals $P \cap \{g_t = 0\}$ for some t.*

***(iii)** Every $y \in \partial P$ lies on a facet, and every facet containing y equals $P \cap \{g_t = 0\}$ for some t with $g_t(y) = 0$.*

Proof. Write $g_t(y) = c_t^\top y - \beta_t$ with $c_t \in \mathbb{R}^n \setminus \{0\}$, which is what non-constant means, and $\beta_t \in \mathbb{R}$. The set P is convex and, as a finite intersection of closed half-spaces, closed.

(i) Let $y \in P$. If $g_t(y) < 0$ for every t, then by continuity of the finitely many g_t a neighbourhood of y lies in P, so y is an interior point. If $g_t(y) = 0$ for some t, then for every $\varepsilon > 0$ the point $y + \varepsilon c_t$ satisfies $g_t(y + \varepsilon c_t) = \varepsilon \|c_t\|^2 > 0$ and lies outside P; hence y is a limit of points outside P, that is, a boundary point. Since P is closed, $\partial P \subseteq P$, and the two cases prove both claims.

(ii) Let $F = P \cap \{c^\top y = \beta\}$ be a facet, where $c^\top y \leq \beta$ is valid on P. Since $\dim F = n - 1 < n = \dim P$, the face F is neither P nor empty, which forces $c \neq 0$. Let y_0 be a point of the relative interior of F, which is nonempty because F is a nonempty convex set. The point y_0 maximises $c^\top y$ over P, and $c \neq 0$, so y_0 is not an interior point of P (otherwise $y_0 + \varepsilon c \in P$ for small $\varepsilon > 0$ would have a larger value); by (i), $g_t(y_0) = 0$ for some t. Fix such a t and let $y \in F$. Because y_0 lies in the relative interior of F and $y_0 + \varepsilon(y_0 - y) \in \operatorname{aff} F$, the point $y_0 + \varepsilon(y_0 - y)$ lies in $F \subseteq P$ for all small $\varepsilon > 0$. Since g_t is affine and $g_t(y_0) = 0$,

$$0 \geq g_t\big(y_0 + \varepsilon(y_0 - y)\big) = (1 + \varepsilon) g_t(y_0) - \varepsilon g_t(y) = -\varepsilon g_t(y),$$

so $g_t(y) \geq 0$, and $y \in P$ gives $g_t(y) \leq 0$; hence $g_t(y) = 0$. Thus $F \subseteq G := P \cap \{g_t = 0\}$. The set G is convex and lies in the hyperplane $\{g_t = 0\}$, so $\dim G \leq n - 1 = \dim F$; together with $F \subseteq G$ this gives $\operatorname{aff} F = \operatorname{aff} G$. It follows that y_0, which has a neighbourhood in $\operatorname{aff} F$ contained in $F \subseteq G$, lies in the relative interior of G. Now let $y \in G$. For all small $\varepsilon > 0$ the point $y_0 + \varepsilon(y_0 - y)$ lies in $G \subseteq P$, and validity of $c^\top y \leq \beta$ together with $c^\top y_0 = \beta$ gives

$$\beta \geq c^\top \big(y_0 + \varepsilon(y_0 - y)\big) = (1 + \varepsilon)\beta - \varepsilon\, c^\top y,$$

so $c^\top y \geq \beta$. With validity this gives $c^\top y = \beta$, that is, $y \in F$. Therefore $F = G$.

(iii) Let $I \subseteq [K]$ be minimal under inclusion with $P = \{y : g_t(y) \leq 0 \text{ for all } t \in I\}$; it exists because $[K]$ is finite. Let $y \in \partial P$. Part (i), applied to the system $\{g_t\}_{t \in I}$, which describes the same set P, gives an $s \in I$ with $g_s(y) = 0$. We show that $F_s = P \cap \{g_s = 0\}$ is a facet. By minimality of I there is a point $p \in \mathbb{R}^n$ with $g_s(p) > 0$ and $g_t(p) \leq 0$ for every $t \in I \setminus \{s\}$, since otherwise the constraint s could be dropped. Let q be an interior point of P; by (i), $g_t(q) < 0$ for every t. The function $\lambda \mapsto g_s((1 - \lambda)q + \lambda p)$ is affine on $[0, 1]$, negative at 0 and positive at 1, so it vanishes at some $\lambda^* \in (0, 1)$. At $y_0 = (1 - \lambda^*)q + \lambda^* p$, every $t \in I \setminus \{s\}$ satisfies $g_t(y_0) = (1 - \lambda^*) g_t(q) + \lambda^* g_t(p) < 0$. By continuity there is a neighbourhood N of y_0 on

which these finitely many inequalities stay strict, and since the system I describes P, this gives $N \cap \{g_s = 0\} \subseteq F_s$. The set $N \cap \{g_s = 0\}$ is a nonempty relatively open subset of the hyperplane $\{g_s = 0\}$, so $\dim F_s \geq n-1$, and F_s lies in that hyperplane, so $\dim F_s = n-1$. Since $g_s \leq 0$ is valid on P, the set F_s is a face, hence a facet, and it contains y. Finally, by (ii) a facet containing y equals $P \cap \{g_t = 0\}$ for some $t \in [K]$, and y lying in it forces $g_t(y) = 0$. □

Proposition 2: the tropical form and the dual subdivision. Here $a_i = (w_i, b_i - \theta) \in \mathbb{R}^{d+1}$ for $i \in [m]$, so that $\tilde{w}^\theta_J = \sum_{i \in J} a_i$ and $Z_\theta = \sum_{i=1}^m [0, a_i]$. For $y \in \mathbb{R}^d$ the linear functional $\ell_y = \langle \cdot, (y, 1) \rangle$ on $\mathbb{R}^{d+1}$ is the functional *given by* $(y, 1)$. A set $J \subseteq [m]$ *maximises at* x if it attains the maximum in (i) below, and a gate pattern is *realised off the walls* if it equals $G(x)$ for a point x that lies on no wall.

Proposition (Full form of Proposition 2)**.** ***(i)*** *For every $x \in \mathbb{R}^d$, $\sum_{i=1}^m \max(0, z_i(x) - \theta) = \max_{J \subseteq [m]} \langle \tilde{w}^\theta_J, \tilde{x} \rangle$; at $\theta = 0$ this is $\sum_i \mathrm{act}_0(z_i(x)) = \bigodot_i (0 \oplus z_i(x))$. A set $J \subseteq [m]$ maximises at x if and only if $G(x) \subseteq J \subseteq \{i : z_i(x) \geq \theta\}$. Hence the set $G(x)$ always maximises at x, and it is the only maximiser if and only if x lies on no wall.*

(ii) *The map $J \mapsto \tilde{w}^\theta_J$ sends the gate patterns realised off the walls injectively into the set of vertices of Z_θ exposed by a functional ℓ_y, $y \in \mathbb{R}^d$, and onto that set when no unit is constant at the threshold, that is, when $(w_i, b_i) \neq (0, \theta)$ for every i.*

(iii) *Let $x, x' \in \mathbb{R}^d$ lie on no wall, and let $G(x') = G(x) \sqcup \{i\}$ for a unit $i \notin G(x)$. Then $[\tilde{w}^\theta_{G(x)}, \tilde{w}^\theta_{G(x')}]$ is an edge of Z_θ with direction $a_i = (w_i, b_i - \theta)$. The set*

$$\Omega = \big\{ y \in \mathcal{W}_i : \ z_j(y) > \theta \text{ for } j \in G(x), \ \ z_j(y) < \theta \text{ for } j \notin G(x) \cup \{i\} \big\}$$

of points of $\mathcal{W}_i$ that lie on no other wall and at which the other units have pattern $G(x)$ is a nonempty relatively open subset of the hyperplane $\mathcal{W}_i$, whose normal w_i is the projection of the edge's direction to the first d coordinates; and for every $y \in \Omega$ the face of Z_θ exposed by ℓ_y is this edge.

Proof. Fix $x \in \mathbb{R}^d$ and put $c_i = z_i(x) - \theta$ for $i \in [m]$. Since $\langle a_i, \tilde{x} \rangle = \langle w_i, x \rangle + b_i - \theta = c_i$, every $J \subseteq [m]$ has $\langle \tilde{w}^\theta_J, \tilde{x} \rangle = \sum_{i \in J} c_i$, the empty sum being 0.

(i), the identity. For every J, $\sum_{i \in J} c_i \leq \sum_{i \in J} \max(0, c_i) \leq \sum_{i \in [m]} \max(0, c_i)$, because $c_i \leq \max(0, c_i)$ and $\max(0, c_i) \geq 0$. The set $\{i : c_i > 0\}$ attains the right-hand side, so $\max_J \langle \tilde{w}^\theta_J, \tilde{x} \rangle = \sum_i \max(0, c_i)$, which is the identity. At $\theta = 0$, $\max(0, z_i(x)) = \mathrm{act}_0(z_i(x)) = 0 \oplus z_i(x)$, and the ordinary sum of the m units is their tropical product $\bigodot_i (0 \oplus z_i(x))$. Expanding this product by distributivity of $\odot$ over $\oplus$ gives $\bigoplus_J \bigodot_{i \in J} z_i(x) = \max_J \sum_{i \in J} z_i(x)$, which is the maximum above at $\theta = 0$.

(i), the maximisers. For every J,

$$\sum_{i \in [m]} \max(0, c_i) - \sum_{i \in J} c_i = \sum_{i \notin J} \max(0, c_i) + \sum_{i \in J} \max(0, -c_i),$$

because $\max(0, c_i) - c_i = \max(0, -c_i)$. Every summand on the right is non-negative, so J maximises at x if and only if every summand vanishes, that is, if and only if $c_i \leq 0$ for every $i \notin J$ and $c_i \geq 0$ for every $i \in J$. The first condition says $\{i : c_i > 0\} \subseteq J$, that is, $G(x) \subseteq J$, and the second says $J \subseteq \{i : c_i \geq 0\} = \{i : z_i(x) \geq \theta\}$. In particular $J = G(x)$ maximises. If x lies on no wall, no c_i is 0, so $G(x) = \{i : z_i(x) \geq \theta\}$ and only one set lies between them. If x lies on the wall of a unit j, then $c_j = 0$, and $G(x)$ and $G(x) \cup \{j\}$ are two different maximisers.

(ii) By Lemma 17, $Z_\theta = \mathrm{conv}\{\tilde{w}^\theta_J : J \subseteq [m]\}$. For $y \in \mathbb{R}^d$, the value of ℓ_y at $\tilde{w}^\theta_J$ is the term of J in the maximum at y, so by Lemma 16 the face $F_{\ell_y}(Z_\theta)$ is the convex hull of the points $\tilde{w}^\theta_J$ of the sets J that maximise at y.

Into. Let x lie on no wall. By (i) the set $G(x)$ is the only maximiser at x, so $F_{\ell_x}(Z_\theta) = \{\tilde{w}^\theta_{G(x)}\}$, a vertex exposed by ℓ_x.

Injective. Let x and x' lie on no wall, with $\tilde{w}^\theta_{G(x')} = \tilde{w}^\theta_{G(x)}$. The term of $G(x')$ at x is then $\langle \tilde{w}^\theta_{G(x)}, \tilde{x} \rangle$, the maximum, so $G(x')$ maximises at x; since the maximiser at x is unique by (i), $G(x') = G(x)$.

Onto. Assume that $a_j \neq 0$ for every j, which is the condition $(w_j, b_j) \neq (0, \theta)$. Let v be a vertex of Z_θ exposed by ℓ_y for some $y \in \mathbb{R}^d$, so that $F_{\ell_y}(Z_\theta) = \{v\}$ and every J that maximises at y has $\tilde{w}^\theta_J = v$. Suppose, for a contradiction, that y lies on the wall of some unit j. Then $j \notin G(y)$ and $z_j(y) \geq \theta$, so by (i) both $G(y)$ and $G(y) \cup \{j\}$ maximise at y; their points differ by $a_j \neq 0$, yet both equal v. Hence y lies on no wall, $G(y)$ is a pattern realised off the walls, and $v = \tilde{w}^\theta_{G(y)}$.

(iii) Let $J = G(x)$, so that $G(x') = J \sqcup \{i\}$. For $j \neq i$, $j \in G(x)$ if and only if $j \in G(x')$, and since neither x nor x' lies on $\mathcal{W}_j$, the numbers $z_j(x) - \theta$ and $z_j(x') - \theta$ are nonzero and of the same sign. For the unit i, $i \notin G(x)$ and $x \notin \mathcal{W}_i$ give $z_i(x) < \theta$, and $i \in G(x')$ gives $z_i(x') > \theta$. In particular z_i is not constant, so $w_i \neq 0$ and $\mathcal{W}_i = \{y : \langle w_i, y\rangle = \theta - b_i\}$ is a hyperplane with normal w_i.

The set Ω is nonempty. For $s \in [0, 1]$ let $x_s = (1 - s)x + sx'$. Since each z_j is affine, $z_j(x_s) - \theta = (1 - s)(z_j(x) - \theta) + s(z_j(x') - \theta)$. For $j \neq i$ this is a convex combination of two nonzero numbers of the same sign, hence nonzero with that sign. For $j = i$ it is an affine function of s that is negative at $s = 0$ and positive at $s = 1$, so it vanishes at exactly one $s^* \in (0, 1)$. The point x_{s^*} therefore lies on $\mathcal{W}_i$, has $z_j(x_{s^*}) > \theta$ for $j \in J$ and $z_j(x_{s^*}) < \theta$ for $j \notin J \cup \{i\}$, and so belongs to Ω.

The set Ω is relatively open. It is the intersection of $\mathcal{W}_i$ with the finitely many sets $\{z_j > \theta\}$, $j \in J$, and $\{z_j < \theta\}$, $j \notin J \cup \{i\}$, each of which is open in $\mathbb{R}^d$ because z_j is continuous; hence Ω is open in $\mathcal{W}_i$.

The exposed face. Let $y \in \Omega$. Then $z_i(y) = \theta$, and for $j \neq i$, $z_j(y) \neq \theta$ with $z_j(y) > \theta$ if and only if $j \in J$. Hence $G(y) = J$ and $\{j : z_j(y) \geq \theta\} = J \cup \{i\}$, and by (i) the maximisers at y are exactly J and $J \cup \{i\}$. By (ii) the face exposed by ℓ_y is $\operatorname{conv}\{\tilde{w}^\theta_J, \tilde{w}^\theta_J + a_i\} = [\tilde{w}^\theta_{G(x)}, \tilde{w}^\theta_{G(x')}]$, a segment of length $\|a_i\| \geq \|w_i\| > 0$. It is therefore an edge of Z_θ, its direction is a_i, and the first d coordinates of a_i form w_i, the normal of $\mathcal{W}_i$. Since Ω is nonempty, this proves every claim of (iii). □

Proposition 4: the data–weight duality. Here one unit is fixed, with base weights $\tilde{w} \in \mathbb{R}^{d+1}$, pre-activations $z_t = \langle \tilde{w}, \tilde{x}_t\rangle$ and base open set $O = \{t : z_t > \theta\}$. Token t is *open at* $\tilde{v} \in \mathbb{R}^{d+1}$ if $\langle \tilde{v}, \tilde{x}_t\rangle > \theta$ and closed otherwise, and it *changes cell* between two weight points if it is open at exactly one of them. The sets H_t, $\mathcal{P}_\mathcal{X}$ and $\mathcal{K}_\mathcal{X}$ are those of Definition 3, so that $\mathcal{K}_\mathcal{X} = \{\tilde{v} : \langle \tilde{v}, \tilde{x}_t\rangle > \theta \ (t \in O),\ \langle \tilde{v}, \tilde{x}_t\rangle \leq \theta \ (t \notin O)\}$, and e_{d+1} is the last standard basis vector of $\mathbb{R}^{d+1}$. Each H_t is a hyperplane because $\tilde{x}_t \neq 0$.

Proposition (Full form of Proposition 4)**.** ***(a)*** *Token t is open at $\tilde{v} \in \mathbb{R}^{d+1}$ if and only if $\tilde{v}$ lies in the open half-space $H_t^> = \{\tilde{v} : \langle \tilde{v}, \tilde{x}_t\rangle > \theta\}$. For $\hat{w} \in \mathbb{R}^{d+1}$, token t changes cell between $\tilde{w}$ and $\hat{w}$ if and only if exactly one of them lies in $H_t^>$, and it closes ($t \in S_{+-}$) if and only if that one is $\tilde{w}$. Along the segment $[\tilde{w}, \hat{w}]$ it changes cell at most once: the gate bit $s \mapsto \mathbb{1}[\langle (1-s)\tilde{w} + s\hat{w}, \tilde{x}_t\rangle > \theta]$, $s \in [0, 1]$, is monotone.*

(b) *The pattern cell $\mathcal{K}_\mathcal{X} = \mathcal{P}_\mathcal{X} \setminus \bigcup_{t \in O} H_t$ is convex. Its interior is $Q = \{\tilde{v} : \langle \tilde{v}, \tilde{x}_t\rangle > \theta \ (t \in O),\ \langle \tilde{v}, \tilde{x}_t\rangle < \theta \ (t \notin O)\}$, which contains $\tilde{w} - \varepsilon e_{d+1}$ for every ε with $0 < \varepsilon < \min_{t \in O}(z_t - \theta)$ (every $\varepsilon > 0$ if $O = \varnothing$) and is a chamber of the arrangement $\{H_t\}_{t \in [T]}$; and $\overline{\mathcal{K}_\mathcal{X}} = \mathcal{P}_\mathcal{X}$.*

(c) *The set $\mathcal{P}_\mathcal{X}$ is a full-dimensional polyhedron, $\partial\mathcal{P}_\mathcal{X} = \bigcup_t (\mathcal{P}_\mathcal{X} \cap H_t)$, every boundary point lies on a facet, and every facet of $\mathcal{P}_\mathcal{X}$ is $\mathcal{P}_\mathcal{X} \cap H_t$ for a token t: an active facet if $t \in O$, an inactive one if not, a label that does not depend on which such t is chosen.*

Proof. (a) By definition token t is open at $\tilde{v}$ if and only if $\langle \tilde{v}, \tilde{x}_t\rangle > \theta$, that is, $\tilde{v} \in H_t^>$; likewise $x_t \in \mathcal{W}_{\tilde{v}}$ if and only if $\tilde{v} \in H_t$. Token t changes cell if and only if its gate bits at $\tilde{w}$ and $\hat{w}$ differ, that is, if and only if exactly one of the two points lies in $H_t^>$. It closes, $t \in \mathcal{C}_> \cap \hat{\mathcal{C}}_\leq = S_{+-}$, if and only if it is open at $\tilde{w}$ and closed at $\hat{w}$, that is, if and only if the point in $H_t^>$ is $\tilde{w}$. Along the segment, $f(s) = \langle (1 - s)\tilde{w} + s\hat{w}, \tilde{x}_t\rangle = (1 - s)z_t + s\hat{z}_t$ is affine in s. If $\hat{z}_t \geq z_t$, then f is non-decreasing and so is the bit $b(s) = \mathbb{1}[f(s) > \theta]$; if $\hat{z}_t < z_t$, both are non-increasing. A monotone function with values in $\{0, 1\}$ changes value at most once: there are no $s_1 < s_2 < s_3$ with $b(s_1) = b(s_3) \neq b(s_2)$.

(b) The set $\mathcal{K}_\mathcal{X}$ is the intersection of the open half-spaces $H_t^>$, $t \in O$, and the closed half-spaces $\{\langle \cdot, \tilde{x}_t\rangle \leq \theta\}$, $t \notin O$, hence convex. The points of $\mathcal{P}_\mathcal{X} \setminus \bigcup_{t \in O} H_t$ satisfy $\geq \theta$ and $\neq \theta$ on O and $\leq \theta$ off O, which is the definition of $\mathcal{K}_\mathcal{X}$.

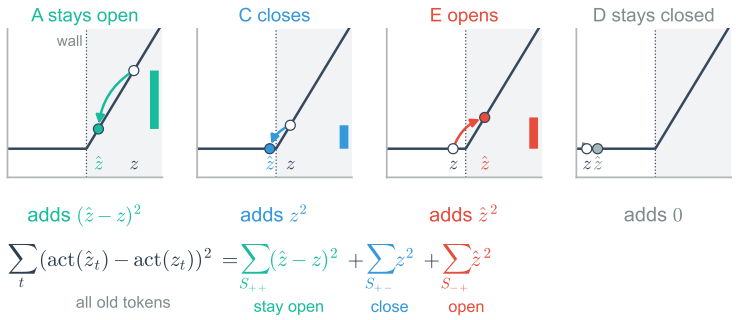


Figure 5: **Theorem 5, one token at a time.** The four old tokens of Figure 1d under the update of unit 1 from w to $\hat{w}$. Each panel draws the activation $\mathrm{act}(z) = 0 \oplus z$ against the pre-activation; shading marks the open side of the wall (dotted). The hollow dot is the token's pre-activation z at w, the filled dot its pre-activation $\hat{z}$ at $\hat{w}$, and the bar on the right is the change in the unit's output. The term under each panel is what that token adds to Eq. (2), and the words under the formula name the tokens each sum runs over. A stays open, moving from $z = 1.50$ to $\hat{z} = 0.38$; its output falls by 1.12, and it contributes the in-cell error $(\hat{z} - z)^2 = 1.26$. C closes, moving from 0.45 to -0.20; its output falls from 0.45 to 0, whatever $\hat{z}$ is below the wall, so it contributes its old output squared, $z^2 = 0.20$. E opens, moving from -0.40 to 0.60; its old output was 0, so it contributes its new output squared, $\hat{z}^2 = 0.36$. D moves from -2.17 to -1.83 but stays closed, so its output is 0 before and after and it contributes nothing, however far it moves. The three terms add to 1.82, which is the total squared change of the unit's output over the four tokens, with no remainder.

The point $\tilde{w} - \varepsilon e_{d+1}$. The last coordinate of every $\tilde{x}_t$ is 1, so $\langle \tilde{w} - \varepsilon e_{d+1}, \tilde{x}_t \rangle = z_t - \varepsilon$. For $t \in O$ and $0 < \varepsilon < \min_{s \in O}(z_s - \theta)$ this exceeds θ, and for $t \notin O$ it is at most $\theta - \varepsilon < \theta$. Hence $\tilde{w} - \varepsilon e_{d+1} \in Q$, and $Q \neq \varnothing$.

The interior of $\mathcal{K}_\mathcal{X}$ *is* Q. The set Q is open, as a finite intersection of open half-spaces, and $Q \subseteq \mathcal{K}_\mathcal{X}$, so Q lies in the interior of $\mathcal{K}_\mathcal{X}$. Conversely, let $\tilde{v}$ be an interior point of $\mathcal{K}_\mathcal{X}$. Then $\langle \tilde{v}, \tilde{x}_t \rangle > \theta$ for $t \in O$. Suppose, for a contradiction, that $\langle \tilde{v}, \tilde{x}_t \rangle = \theta$ for some $t \notin O$. For every $\varepsilon > 0$, $\langle \tilde{v} + \varepsilon \tilde{x}_t, \tilde{x}_t \rangle = \theta + \varepsilon \|\tilde{x}_t\|^2 > \theta$ because $\tilde{x}_t \neq 0$, so $\tilde{v} + \varepsilon \tilde{x}_t \notin \mathcal{K}_\mathcal{X}$, and $\tilde{v}$ is not interior. Hence $\langle \tilde{v}, \tilde{x}_t \rangle < \theta$ for every $t \notin O$, and $\tilde{v} \in Q$.

The set Q *is a chamber.* Let $A = \mathbb{R}^{d+1} \setminus \bigcup_{t \in [T]} H_t$ and, for $\sigma \in \{+1, -1\}^T$, $Q_\sigma = \{\tilde{v} : \sigma_t(\langle \tilde{v}, \tilde{x}_t \rangle - \theta) > 0 \text{ for all } t\}$. Every point of A has $\langle \tilde{v}, \tilde{x}_t \rangle \neq \theta$ for all t, which fixes one σ, so the sets Q_σ are pairwise disjoint, open and convex, and their union is A. Let Γ be a connected component of A and $\tilde{v} \in \Gamma \cap Q_\sigma$. The sets $\Gamma \cap Q_{\sigma'}$ are disjoint, relatively open in Γ and cover it, so connectedness of Γ gives $\Gamma \subseteq Q_\sigma$. Conversely, Q_σ is convex, hence connected, contains $\tilde{v}$ and lies in A, so $Q_\sigma \subseteq \Gamma$. The chambers are therefore exactly the nonempty sets Q_σ, and $Q = Q_{\sigma^0}$ with $\sigma^0_t = +1$ for $t \in O$ and $\sigma^0_t = -1$ otherwise is nonempty, hence a chamber.

The closure. Since $Q \subseteq \mathcal{K}_\mathcal{X} \subseteq \mathcal{P}_\mathcal{X}$ and $\mathcal{P}_\mathcal{X}$ is closed, $\overline{\mathcal{K}_\mathcal{X}} \subseteq \mathcal{P}_\mathcal{X}$. Let $y \in \mathcal{P}_\mathcal{X}$, $q \in Q$ and $s \in (0, 1]$, and put $y_s = (1 - s)y + sq$. For $t \in O$, $\langle y_s, \tilde{x}_t \rangle = (1 - s)\langle y, \tilde{x}_t \rangle + s\langle q, \tilde{x}_t \rangle > (1 - s)\theta + s\theta = \theta$, because $\langle y, \tilde{x}_t \rangle \geq \theta$, $\langle q, \tilde{x}_t \rangle > \theta$ and $s > 0$; for $t \notin O$ the same computation gives $\langle y_s, \tilde{x}_t \rangle < \theta$. Hence $y_s \in Q$, and $y_s \to y$ as $s \to 0$. It follows that $\mathcal{P}_\mathcal{X} \subseteq \overline{Q} \subseteq \overline{\mathcal{K}_\mathcal{X}} \subseteq \mathcal{P}_\mathcal{X}$.

(c) Put $g_t(\tilde{v}) = \theta - \langle \tilde{v}, \tilde{x}_t \rangle$ for $t \in O$ and $g_t(\tilde{v}) = \langle \tilde{v}, \tilde{x}_t \rangle - \theta$ for $t \notin O$. Then $\mathcal{P}_\mathcal{X} = \{g_t \leq 0 \text{ for all } t\}$, each g_t is affine with linear part $\pm \tilde{x}_t \neq 0$, hence non-constant, and $\{g_t = 0\} = H_t$. The polyhedron $\mathcal{P}_\mathcal{X}$ contains the nonempty open set Q of (b), so it has nonempty interior and is full-dimensional, and Lemma 18 gives the description of $\partial \mathcal{P}_\mathcal{X}$, the facet through every boundary point, and the form $\mathcal{P}_\mathcal{X} \cap H_t$ of every facet. For the label, let F be a facet with $F = \mathcal{P}_\mathcal{X} \cap H_t = \mathcal{P}_\mathcal{X} \cap H_s$. The facet has dimension d and lies in $H_t \cap H_s$; if $H_t \neq H_s$, this intersection is empty or an affine subspace of dimension $d - 1$, which cannot contain F, so $H_t = H_s$. The affine functions $\langle \cdot, \tilde{x}_t \rangle - \theta$ and $\langle \cdot, \tilde{x}_s \rangle - \theta$ then have the same zero hyperplane, so their linear parts are proportional, $\tilde{x}_s = c\tilde{x}_t$ for a real $c \neq 0$; the last coordinates, both equal to 1, give $c = 1$. Hence $\tilde{x}_s = \tilde{x}_t$, so $z_s = z_t$, and s and t have the same base gate bit. □

Corollary 6: the weight-space reading.

Corollary (Full form of Corollary 6)**.** *Let $\hat{w} \in \mathbb{R}^{d+1}$, and let O, $\mathcal{P}_{\mathcal{X}}$ and $\mathcal{K}_{\mathcal{X}}$ be those of the base weights $\tilde{w}$.*

(a) *The set $S_{+-} \cup S_{-+}$ consists of the tokens whose hyperplanes separate $\tilde{w}$ from $\hat{w}$ in the sense of Proposition 4a, that is, with exactly one of the two points in $H_t^{>}$, and S_{+-} consists of those with $\tilde{w} \in H_t^{>}$.*

(b) *On every set $W \subseteq \mathbb{R}^{d+1}$ of weights with a fixed gate pattern on $\mathcal{X}$ (every $\hat{w} \in W$ gives each token the same gate bit), L_{post} is the restriction to W of a polynomial of degree at most two in $\hat{w}$.*

(c) *On $\mathcal{K}_{\mathcal{X}}$, L_{post} is the in-cell quadratic $\|\tilde{X}_O(\hat{w} - \tilde{w})\|^2$, and at $\theta = 0$ it equals that quadratic on all of $\mathcal{P}_{\mathcal{X}}$.*

Proof. (a) By definition $t \in S_{+-} \cup S_{-+}$ if and only if t changes cell between $\tilde{w}$ and $\hat{w}$, and $t \in S_{+-}$ if and only if it closes; Proposition 4a translates both conditions into the stated ones.

(b) Each of the four index sets of Theorem 5 is determined by the base gate bits and the new gate bits, so all four are the same for every $\hat{w} \in W$. For such $\hat{w}$ the in-cell term $\sum_{t \in S_{++}} \langle \hat{w} - \tilde{w}, \tilde{x}_t \rangle^2$ and the closed→open term $\sum_{t \in S_{-+}} \langle \hat{w}, \tilde{x}_t \rangle^2$ are polynomials of degree at most two in $\hat{w}$, and the open→closed term $\sum_{t \in S_{+-}} z_t^2$ does not depend on $\hat{w}$. By Theorem 5 their sum is L_{post}.

(c) Let $\hat{w} \in \mathcal{K}_{\mathcal{X}}$. Every token then has its base gate bit, so $S_{+-} = S_{-+} = \varnothing$ and $S_{++} = O$, and Theorem 5 gives $L_{\mathrm{post}} = \sum_{t \in O} \langle \hat{w} - \tilde{w}, \tilde{x}_t \rangle^2 = \|\tilde{X}_O(\hat{w} - \tilde{w})\|^2$. Now let $\theta = 0$ and $\hat{w} \in \mathcal{P}_{\mathcal{X}} \setminus \mathcal{K}_{\mathcal{X}}$. Since $\hat{w} \in \mathcal{P}_{\mathcal{X}}$, $\hat{z}_t \geq 0$ for $t \in O$ and $\hat{z}_t \leq 0$ for $t \notin O$, and since $\mathcal{K}_{\mathcal{X}} = \mathcal{P}_{\mathcal{X}} \setminus \bigcup_{t \in O} H_t$ by Proposition 4b, the set $O_0 = \{t \in O : \hat{z}_t = 0\}$ is nonempty. The tokens of $O \setminus O_0$ lie in S_{++} and contribute $(\hat{z}_t - z_t)^2$. The tokens of O_0 lie in S_{+-} and contribute z_t^2, which equals $(\hat{z}_t - z_t)^2$ because $\hat{z}_t = 0$. The tokens off O have $z_t \leq 0$ and $\hat{z}_t \leq 0$, so they lie in S_{--} and contribute 0. Hence $L_{\mathrm{post}} = \sum_{t \in O} (\hat{z}_t - z_t)^2 = \|\tilde{X}_O(\hat{w} - \tilde{w})\|^2$. □

Proposition 8: exactness and the ridge residual. Here unit i, the base inputs and an integer cap $C \geq 0$ are fixed, and S_i, k_i, A_i and $P_i(\rho)$ are those of Definition 7. We write $z_t = z_i(x_t)$; for $u \in \mathbb{R}^{d+1}$ and $\rho \geq 0$ we put $u' = P_i(\rho)u$, the new row is $\tilde{w}_i + u'$, and $\hat{z}_t = z_t + \langle u', \tilde{x}_t \rangle$. The sets H_t, $\mathcal{P}_{\mathcal{X}}$ and $\mathcal{K}_{\mathcal{X}}$ are those of unit i, that is, of Definition 3 with $O = O_i$. Since $\theta \geq 0$, $h_t^{(i)} > 0$ if and only if $t \in O_i$; hence $S_i \subseteq O_i$ and $k_i = \min(C, |O_i|)$.

Proposition (Full form of Proposition 8)**.** ***(a)*** *The matrix $P_i(0)$ is the orthogonal projector onto $\ker A_i$: at $\rho = 0$, $A_i u' = 0$, and u' is the point of $\ker A_i$ nearest to u. Let $k_i \geq 1$ (then $A_i \neq 0$, since every row ends in 1) and let σ_+ be the smallest positive singular value of A_i. For $\rho > 0$ the residual satisfies $\|A_i u'\| \leq \frac{\rho}{\sigma_+^2 + \rho} \|A_i u\|$, with equality for a right singular vector of A_i belonging to σ_+.*

(a′) *Let $k_i \geq 1$ and $\rho > 0$. Then $P_i(\rho) = \rho(A_i^\top A_i + \rho I)^{-1}$; the vector u' is the unique minimiser of $\|v - u\|^2 + \rho^{-1}\|A_i v\|^2$ over $v \in \mathbb{R}^{d+1}$; and $\|P_i(\rho) - P_i(0)\|_2 = \rho/(\sigma_+^2 + \rho)$, so $P_i(\rho) \to P_i(0)$ as $\rho \to 0$. The bound of (a) depends on duplicate or dependent tokens only through σ_+.*

(b) *At the base inputs and $\rho = 0$, every pair in S_i stays in S_{++} and contributes zero to all three terms, and every pair closed at base contributes only to closed→open. Unit i's in-cell and open→closed terms are sums over $O_i \setminus S_i$, which is empty when $C \geq |O_i|$.*

(c) *Let $C \geq |O_i|$ and $\rho = 0$. The flat $F = \tilde{w}_i + \ker \tilde{X}_{O_i}$, which contains every new row $\tilde{w}_i + u'$, is parallel to every active hyperplane H_t, $t \in O_i$, and lies in $\bigcap_{t \in O_i} H_t^{>}$, so it meets $\mathcal{K}_{\mathcal{X}}$ exactly where it meets $\mathcal{P}_{\mathcal{X}}$. Every point where F meets $\partial\mathcal{P}_{\mathcal{X}}$ lies on a facet, and every facet through such a point is inactive (Figure 1f). In particular, every continuous path in F that starts in $\mathcal{K}_{\mathcal{X}}$ and ends outside $\mathcal{K}_{\mathcal{X}}$ meets $\partial\mathcal{P}_{\mathcal{X}}$.*

(d) *Let $S \subseteq [T]$ be any selected set, open or closed, and $u' = (I - \tilde{X}_S^{+}\tilde{X}_S)u$. At the base inputs every token of S keeps its pre-activation and contributes zero to L_{post}, and a closed token t with $\theta - z_i(x_t) \geq \|u'\| \, \|\tilde{x}_t\|$ contributes zero whether it is in S or not.*

Proof. If $k_i = 0$, then A_i has no rows, $\ker A_i = \mathbb{R}^{d+1}$, and $P_i(\rho) = I = P_i(0)$ by Definition 7; so $P_i(0)$ is the orthogonal projector onto $\ker A_i$ and $u' = u$ is the nearest point, which is the first sentence of (a); the rest of (a), and (a′), assume $k_i \geq 1$. The proofs of (b), (c) and (d) below use only the first sentence of (a) and therefore also hold when $k_i = 0$.

Let $k_i \geq 1$. Every row of A_i ends in 1, so $A_i \neq 0$. Let $r = \operatorname{rank} A_i \geq 1$ and let $A_i = \sum_{j=1}^{r} \sigma_j p_j q_j^\top$ be a compact singular value decomposition, with $\sigma_j > 0$ and orthonormal families $\{p_j\} \subseteq \mathbb{R}^{k_i}$ and $\{q_j\} \subseteq \mathbb{R}^{d+1}$, so that $\sigma_+ = \min_j \sigma_j$. Let $\Pi = \sum_j q_j q_j^\top$, the orthogonal projector onto the row space $\operatorname{span}\{q_j\}$ of A_i; then $I - \Pi$ is the orthogonal projector onto its orthogonal complement, which is $\ker A_i$.

(a), the projector. The pseudo-inverse is $A_i^+ = \sum_j \sigma_j^{-1} q_j p_j^\top$, so $A_i^+ A_i = \sum_{j,l} \sigma_j^{-1} \sigma_l q_j (p_j^\top p_l) q_l^\top = \Pi$ and $P_i(0) = I - \Pi$. By the projection theorem, $(I - \Pi)u$ is the unique point of $\ker A_i$ nearest to u, and $A_i(I - \Pi)u = 0$ because the range of $I - \Pi$ is $\ker A_i$.

The rest of (a), and (a′). Let $\rho > 0$ and $E = \sum_j p_j p_j^\top$. Then $A_i A_i^\top + \rho I = \sum_j (\sigma_j^2 + \rho) p_j p_j^\top + \rho(I - E)$ acts as multiplication by $\sigma_j^2 + \rho > 0$ on each p_j and by $\rho > 0$ on the range of $I - E$, so it is invertible, with inverse $\sum_j (\sigma_j^2 + \rho)^{-1} p_j p_j^\top + \rho^{-1}(I - E)$. Since $A_i^\top = \sum_j \sigma_j q_j p_j^\top$ vanishes on the range of $I - E$,

$$A_i^\top (A_i A_i^\top + \rho I)^{-1} A_i = \sum_j \frac{\sigma_j^2}{\sigma_j^2 + \rho} q_j q_j^\top, \qquad \text{hence} \qquad P_i(\rho) = (I - \Pi) + \sum_j \frac{\rho}{\sigma_j^2 + \rho} q_j q_j^\top.$$

The same basis diagonalises $A_i^\top A_i + \rho I = \sum_j (\sigma_j^2 + \rho) q_j q_j^\top + \rho(I - \Pi)$, and ρ times its inverse is $(I - \Pi) + \sum_j \rho(\sigma_j^2 + \rho)^{-1} q_j q_j^\top = P_i(\rho)$. The objective $\phi(v) = \|v - u\|^2 + \rho^{-1}\|A_i v\|^2$ is a quadratic with Hessian $2I + 2\rho^{-1} A_i^\top A_i$, which is positive definite, so its unique minimiser is the unique zero of $\nabla\phi(v) = 2(v - u) + 2\rho^{-1} A_i^\top A_i v$, namely $v = (I + \rho^{-1} A_i^\top A_i)^{-1} u = \rho(A_i^\top A_i + \rho I)^{-1} u = u'$. Next, $A_i q_l = \sigma_l p_l$ and $A_i(I - \Pi) = 0$ give $A_i u' = \sum_j \frac{\rho \sigma_j}{\sigma_j^2 + \rho} (q_j^\top u)\, p_j$, and orthonormality of the p_j gives

$$\|A_i u'\|^2 = \sum_j \Big(\frac{\rho}{\sigma_j^2 + \rho}\Big)^2 \sigma_j^2 (q_j^\top u)^2 \leq \Big(\frac{\rho}{\sigma_+^2 + \rho}\Big)^2 \sum_j \sigma_j^2 (q_j^\top u)^2 = \Big(\frac{\rho}{\sigma_+^2 + \rho}\Big)^2 \|A_i u\|^2,$$

where the inequality holds because $s \mapsto \rho/(s + \rho)$ is decreasing on $s > 0$ and every $\sigma_j \geq \sigma_+$, and the last equality because $A_i u = \sum_j \sigma_j (q_j^\top u) p_j$. For $u = q_{j^*}$ with $\sigma_{j^*} = \sigma_+$, a right singular vector belonging to σ_+, both sides equal $(\rho/(\sigma_+^2 + \rho))^2 \sigma_+^2$. Finally, $P_i(\rho) - P_i(0) = \sum_j \rho(\sigma_j^2 + \rho)^{-1} q_j q_j^\top$ is symmetric with eigenvalues $\rho/(\sigma_j^2 + \rho)$ and 0, so its operator norm is $\max_j \rho/(\sigma_j^2 + \rho) = \rho/(\sigma_+^2 + \rho)$, which tends to 0 as $\rho \to 0$. Repeated or linearly dependent rows of A_i change r and the σ_j, and the bound sees them only through σ_+.

(b) At the base inputs the new pre-activation of token t is $\hat{z}_t = z_t + \langle u', \tilde{x}_t \rangle$. Let $\rho = 0$ and $t \in S_i$. The number $\langle u', \tilde{x}_t \rangle$ is the entry of $A_i u'$ in the row of t, which is 0 by (a); hence $\hat{z}_t = z_t > \theta$, as $S_i \subseteq O_i$. The pair is therefore in S_{++} with in-cell contribution $(\hat{z}_t - z_t)^2 = 0$, and in neither crossing set. A pair closed at base has $z_t \leq \theta$, so it lies in $S_{-+} \cup S_{--}$ and can contribute only to closed→open. The in-cell and open→closed terms are sums over $S_{++} \cup S_{+-} = O_i$, and the pairs of S_i contribute zero to them, so they are sums over $O_i \setminus S_i$. Since $S_i \subseteq O_i$ and $|S_i| = k_i = \min(C, |O_i|)$, the set $O_i \setminus S_i$ is empty if and only if $C \geq |O_i|$.

(c) Let $C \geq |O_i|$ and $\rho = 0$. Then $S_i = O_i$ by (b), $A_i = \tilde{X}_{O_i}$, and $u' \in \ker \tilde{X}_{O_i}$ by (a), so $\tilde{w}_i + u' \in F$. The direction space of F is $\ker \tilde{X}_{O_i} = \bigcap_{t \in O_i} \tilde{x}_t^\perp$, which is contained in the direction space $\tilde{x}_t^\perp$ of each active hyperplane H_t: the flat is parallel to all of them. For $v \in \ker \tilde{X}_{O_i}$ and $t \in O_i$, $\langle \tilde{w}_i + v, \tilde{x}_t \rangle = z_t > \theta$, so $F \subseteq \bigcap_{t \in O_i} H_t^>$. A point of F thus satisfies every active constraint strictly, so it lies in $\mathcal{K}_\mathcal{X}$ if and only if it satisfies the inactive constraints, that is, if and only if it lies in $\mathcal{P}_\mathcal{X}$. Let $y \in F \cap \partial\mathcal{P}_\mathcal{X}$. By Proposition 4c, which rests on Lemma 18(iii), the point y lies on a facet, and every facet through y is $\mathcal{P}_\mathcal{X} \cap H_t$ for a token t with $y \in H_t$. Since $y \in H_s^>$ for every $s \in O_i$, y lies on no active hyperplane, so $t \notin O_i$ and the facet is inactive; by Proposition 4c the label does not depend on the choice of t. For the path, let $\alpha : [0, 1] \to F$ be continuous with $\alpha(0) \in \mathcal{K}_\mathcal{X}$ and $\alpha(1) \notin \mathcal{K}_\mathcal{X}$. By the first claim $\alpha(1) \notin \mathcal{P}_\mathcal{X}$, so $s^* = \inf\{s \in [0, 1] : \alpha(s) \notin \mathcal{P}_\mathcal{X}\}$

is well defined. Since $\mathcal{P}_\mathcal{X}$ is closed, $\alpha(0) \in \mathcal{P}_\mathcal{X}$ and $\alpha(s) \in \mathcal{P}_\mathcal{X}$ for every $s < s^*$, continuity gives $\alpha(s^*) \in \mathcal{P}_\mathcal{X}$; hence s^* is not in the set, and by the definition of the infimum there are $s_n > s^*$ with $s_n \to s^*$ and $\alpha(s_n) \notin \mathcal{P}_\mathcal{X}$. So $\alpha(s^*)$ is a point of $\mathcal{P}_\mathcal{X}$ that is a limit of points outside $\mathcal{P}_\mathcal{X}$, that is, a point of $F \cap \partial\mathcal{P}_\mathcal{X}$.

(d) If $S = \varnothing$, then $\tilde{X}_S^+ \tilde{X}_S = 0$ and $u' = u$, and the first claim is empty. If $S \neq \varnothing$, the argument of (a), which uses no property of A_i other than being a real matrix, applies to $\tilde{X}_S$ and gives $\tilde{X}_S u' = 0$. Hence $\hat{z}_t = z_t$ for $t \in S$, both activations agree, and the token contributes 0 to L_{post} (it lies in S_{++} with zero difference if it is open, and in S_{--} if it is closed). For a closed t with $\theta - z_t \geq \|u'\| \, \|\tilde{x}_t\|$, the Cauchy–Schwarz inequality gives $\hat{z}_t \leq z_t + \|u'\| \, \|\tilde{x}_t\| \leq \theta$, so $t \in S_{--}$ and contributes 0, whether or not $t \in S$. □

Theorem 9: budget separation. Here the base inputs are fixed. For a unit i and an update $u \in \mathbb{R}^{d+1}$ of row i, the in-cell, open→closed and closed→open terms of unit i are those of Theorem 5 with base weights $\tilde{w}_i$ and new weights $\tilde{w}_i + u$. A linear subspace $U \subseteq \mathbb{R}^{d+1}$ is *protecting* for unit i if unit i's in-cell term vanishes for every $u \in U$, and row i is *confined to* U if its update is required to lie in U. We write $r_i = \operatorname{rank} \tilde{X}_{O_i}$ and $O_\cup = \bigcup_i O_i$.

Theorem (Full form of Theorem 9)**.** ***(a)*** *A subspace U is protecting for unit i if and only if $U \subseteq \ker \tilde{X}_{O_i}$, and then unit i's open→closed term also vanishes for every $u \in U$. A protecting subspace removes at least $r_i \leq |O_i|$ directions from row i. If $C \geq |O_i|$, the projector $P_i(0)$ confines row i to $\ker \tilde{X}_{O_i}$, which is protecting for unit i, and removes exactly r_i directions.*

(b) *One subspace U shared by every row is protecting for all units if and only if $U \subseteq \ker \tilde{X}_{O_\cup}$, so it removes at least $\operatorname{rank} \tilde{X}_{O_\cup} \geq \max_i r_i$ directions from every row.*

(c) *Let $B \geq 0$ be an integer. If every row i is confined to a subspace U_i, with $\sum_i \operatorname{codim} U_i \leq B$, the set J of units i for which U_i is protecting satisfies $\sum_{i \in J} r_i \leq B$; conversely, for every $J \subseteq [m]$ with $\sum_{i \in J} r_i \leq B$ there are such U_i under which every unit of J is protected. Per-unit subspaces therefore protect up to $\max\{|J| : \sum_{i \in J} r_i \leq B\}$ units. If instead each row is either frozen ($U_i = \{0\}$, costing $d+1$) or free ($U_i = \mathbb{R}^{d+1}$, costing 0), with total cost at most B, the best choice protects exactly $\#\{i : r_i = 0\} + \min(\lfloor B/(d+1) \rfloor, \#\{i : r_i > 0\})$ units, which is never more than the per-unit count.*

Proof. Fix unit i, write $O = O_i$, and for an update u let $a = \tilde{X}_O u \in \mathbb{R}^{|O|}$, so that $\hat{z}_t = z_t + a_t$ for $t \in O$.

(a) If $O = \varnothing$, the in-cell and open→closed terms are sums over the empty set and vanish for every u, $\ker \tilde{X}_O = \mathbb{R}^{d+1}$ and $r_i = 0$, so every claim holds. Let $O \neq \varnothing$. If $\tilde{X}_O u = 0$, every $t \in O$ keeps $\hat{z}_t = z_t > \theta$, so $S_{++} = O$ with zero in-cell contribution and $S_{+-} = \varnothing$; both terms vanish on $\ker \tilde{X}_O$, and every subspace of it is protecting. Conversely, let U be protecting and suppose, for a contradiction, that some $u \in U$ has $a \neq 0$. Put $\delta = \min_{t \in O}(z_t - \theta)$, which is positive because O is finite and nonempty, and $s = \delta/(2\|a\|_\infty)$. Then $su \in U$, and every $t \in O$ has $|\hat{z}_t - z_t| = s|a_t| \leq \delta/2$, so $\hat{z}_t \geq z_t - \delta/2 > \theta$ and t stays open. The in-cell term of su is therefore $\sum_{t \in O} s^2 a_t^2 = s^2 \|a\|^2 > 0$, which contradicts protection. Hence $U \subseteq \ker \tilde{X}_O$, and by rank–nullity $\dim U \leq d + 1 - r_i$, that is, $\operatorname{codim} U \geq r_i$. Finally $r_i \leq |O|$, the number of rows of $\tilde{X}_O$.

(a), the projector. If $C \geq |O|$, then $S_i = O$ and $A_i = \tilde{X}_O$ by Proposition 8b, and by Proposition 8a the range of $P_i(0)$ is $\ker \tilde{X}_O$, which is protecting by the first claim of (a) and has codimension r_i by rank–nullity.

(b) By (a), a subspace U is protecting for every unit if and only if $U \subseteq \bigcap_i \ker \tilde{X}_{O_i}$. The rows of $\tilde{X}_{O_\cup}$ are, up to repetition, the rows of all the $\tilde{X}_{O_i}$, so $\bigcap_i \ker \tilde{X}_{O_i} = \ker \tilde{X}_{O_\cup}$. Hence $\operatorname{codim} U \geq \operatorname{rank} \tilde{X}_{O_\cup}$, and $\operatorname{rank} \tilde{X}_{O_\cup} \geq \operatorname{rank} \tilde{X}_{O_i} = r_i$ for every i, since the rows of $\tilde{X}_{O_i}$ are among those of $\tilde{X}_{O_\cup}$. The subspace $U = \ker \tilde{X}_{O_\cup}$ attains the bound.

(c) Let row i be confined to U_i for every i, with $\sum_i \operatorname{codim} U_i \leq B$, and let J be the set of protected units. By (a), $\operatorname{codim} U_i \geq r_i$ for $i \in J$, so $\sum_{i \in J} r_i \leq \sum_{i \in J} \operatorname{codim} U_i \leq B$. Conversely, let $\sum_{i \in J} r_i \leq B$, and choose $U_i = \ker \tilde{X}_{O_i}$ for $i \in J$ and $U_i = \mathbb{R}^{d+1}$ otherwise. The total codimen-

sion is $\sum_{i\in J} r_i \le B$, and every unit of J is protected by (a). Hence the largest number of units that per-unit subspaces protect is $\max\{|J| : \sum_{i\in J} r_i \le B\}$; the protected set of the construction also contains the units with $r_i = 0$, and it is itself admissible, so the maximum is attained.

Now let each row be frozen or free. A frozen row is confined to $\{0\} \subseteq \ker \tilde{X}_{O_i}$ and is protected; by (a) a free row is protected if and only if $\mathbb{R}^{d+1} \subseteq \ker \tilde{X}_{O_i}$, that is, if and only if $r_i = 0$. Let $Z = \{i : r_i = 0\}$ and let $\mathcal{F}$ be the set of frozen rows. The budget gives $(d+1)|\mathcal{F}| \le B$, that is, $|\mathcal{F}| \le \lfloor B/(d+1)\rfloor$, and the protected set is $Z \cup \mathcal{F}$, of size $|Z| + |\mathcal{F} \setminus Z| \le |Z| + \min(\lfloor B/(d+1)\rfloor, \#\{i : r_i > 0\})$. Freezing $\min(\lfloor B/(d+1)\rfloor, \#\{i : r_i > 0\})$ rows with $r_i > 0$ respects the budget and attains this bound, which proves the formula. The protected set $Z \cup \mathcal{F}$ of any admissible choice satisfies $\sum_{i\in Z\cup\mathcal{F}} r_i = \sum_{i\in\mathcal{F}\setminus Z} r_i \le (d+1)|\mathcal{F}| \le B$, because $r_i \le d+1$, the number of columns of $\tilde{X}_{O_i}$; so it is admissible for the per-unit count, which is therefore at least as large. □

Proposition 10: the budget-limited optimum. Here $\theta = 0$ and the inputs are the base inputs. For each row i, $U_i \subseteq \mathbb{R}^{d+1}$ is a fixed linear subspace, u_i is a random vector in $\mathbb{R}^{d+1}$ with $\mathbb{E}[u_i u_i^\top] = \sigma^2 I$ for a real $\sigma > 0$ that is the same for every row, and the applied update of row i is $P_{U_i} u_i$. We write $\Xi_i = \tilde{X}_{O_i}^\top \tilde{X}_{O_i}$, which is 0 if $O_i = \varnothing$, and $\lambda_j(\Xi_i)$, $j \in [d+1]$, for its eigenvalues in non-increasing order.

Proposition (Full form of Proposition 10)**.** *Unit i's in-cell plus open→closed terms are at most $\|\tilde{X}_{O_i} P_{U_i} u_i\|^2$, with equality when no open token closes, and $D_i(U) := \mathbb{E}\|\tilde{X}_{O_i} P_U u_i\|^2 = \sigma^2 \operatorname{tr}(\Xi_i P_U)$ for every subspace U.*

***(a)** For every integer q with $0 \le q \le d+1$, $\min_{\operatorname{codim} U = q} D_i(U) = \sigma^2 \sum_{j>q} \lambda_j(\Xi_i)$, attained when $U^\perp$ is a top-q eigenspace of the unit's own open-token Gram Ξ_i.*

***(b)** For one U shared by all rows, $\sum_i D_i(U) = \sigma^2 \operatorname{tr}(\bar{\Xi} P_U)$ with $\bar{\Xi} = \sum_i \Xi_i = \sum_t n_t \tilde{x}_t \tilde{x}_t^\top$, where n_t is the number of units open on token t; over $\operatorname{codim} U = q$ it is minimised when $U^\perp$ is a top-q eigenspace of $\bar{\Xi}$, with minimum $\sigma^2 \sum_{j>q} \lambda_j(\bar{\Xi})$.*

***(c)** For every integer q with $0 \le q \le d+1$, the minimum of $\sum_i D_i(U_i)$ over per-unit subspaces with $\sum_i \operatorname{codim} U_i \le mq$ is at most $\sigma^2 \sum_i \sum_{j>q} \lambda_j(\Xi_i)$, the value of the allocation $\operatorname{codim} U_i = q$ for every i, which is at most the shared minimum of (b).*

***(d)** Let $U_i = \ker \tilde{X}_{S_i}$, the subspace to which the gate projector $P_i(0)$ confines row i, so that $U_i^\perp = \operatorname{span}\{\tilde{x}_t : t \in S_i\}$. If these tokens span the row space of $\tilde{X}_{O_i}$, in particular if $C \ge |O_i|$, then $\operatorname{codim} U_i = r_i$ and $D_i(U_i) = 0$, the minimum of D_i over all subspaces.*

Lemma 19. *Let $\theta = 0$, fix unit i at the base inputs, and let $u \in \mathbb{R}^{d+1}$. Unit i's in-cell plus open→closed terms for the update u are at most $\|\tilde{X}_{O_i} u\|^2$, with equality when no token of O_i closes.*

Proof. Let $t \in O_i$, so that $z_t > 0$ and $\hat{z}_t = z_t + \langle u, \tilde{x}_t\rangle$. If t stays open, it contributes $(\hat{z}_t - z_t)^2 = \langle u, \tilde{x}_t\rangle^2$ to the in-cell term. If t closes, then $\hat{z}_t \le 0 < z_t$, so $z_t - \hat{z}_t \ge z_t > 0$, and its contribution z_t^2 to the open→closed term is at most $(z_t - \hat{z}_t)^2 = \langle u, \tilde{x}_t\rangle^2$. Tokens off O_i enter neither term. Summing over $t \in O_i$ gives the bound $\sum_{t\in O_i} \langle u, \tilde{x}_t\rangle^2 = \|\tilde{X}_{O_i} u\|^2$, and every step is an equality when no token of O_i closes. □

Lemma 20 (Ky Fan)**.** *Let $\Xi \in \mathbb{R}^{n\times n}$ be symmetric positive semidefinite with eigenvalues $\lambda_1 \ge \cdots \ge \lambda_n$, and let $Q \in \mathbb{R}^{n\times n}$ be an orthogonal projector of rank q, $0 \le q \le n$. Then $\operatorname{tr}(\Xi Q) \le \sum_{j\le q} \lambda_j$, with equality when Q projects onto a top-q eigenspace of Ξ.*

Proof. Let $e_1, \ldots, e_n$ be orthonormal eigenvectors with $\Xi e_j = \lambda_j e_j$, so that $\Xi = \sum_j \lambda_j e_j e_j^\top$ and $\operatorname{tr}(\Xi Q) = \sum_j \lambda_j c_j$ with $c_j = e_j^\top Q e_j$. Since Q is symmetric and idempotent, $c_j = \|Q e_j\|^2 \in [0,1]$, and $\sum_j c_j = \operatorname{tr} Q = q$. If $q = 0$, then $Q = 0$ and both sides are 0. If $q \ge 1$,

$$\sum_j \lambda_j c_j - \sum_{j\le q} \lambda_j = \sum_{j\le q} \lambda_j (c_j - 1) + \sum_{j>q} \lambda_j c_j \le \lambda_q \Big(\sum_{j\le q} (c_j - 1) + \sum_{j>q} c_j \Big) = \lambda_q \Big(\sum_j c_j - q \Big) = 0,$$

where for $j \le q$ we used $c_j - 1 \le 0$ and $\lambda_j \ge \lambda_q$, and for $j > q$ we used $c_j \ge 0$ and $\lambda_j \le \lambda_q$. If Q projects onto the span of orthonormal eigenvectors $f_1, \dots, f_q$ with eigenvalues $\lambda_1, \dots, \lambda_q$, then $\operatorname{tr}(\Xi Q) = \sum_{l \le q} f_l^\top \Xi f_l = \sum_{l \le q} \lambda_l$. □

Proof of the proposition. The bound and its equality case are Lemma 19 applied to the update $P_{U_i} u_i$. For the expectation, let U be a subspace. Since P_U is symmetric and idempotent, $\|\tilde{X}_{O_i} P_U u_i\|^2 = u_i^\top P_U \Xi_i P_U u_i = \operatorname{tr}(P_U \Xi_i P_U \, u_i u_i^\top)$, and linearity of the trace and of the expectation gives

$$D_i(U) = \operatorname{tr}\big(P_U \Xi_i P_U \, \mathbb{E}[u_i u_i^\top]\big) = \sigma^2 \operatorname{tr}(P_U \Xi_i P_U) = \sigma^2 \operatorname{tr}(\Xi_i P_U^2) = \sigma^2 \operatorname{tr}(\Xi_i P_U),$$

where the third equality is cyclicity of the trace.

(a) Let $\operatorname{codim} U = q$ and $Q = I - P_U$, the orthogonal projector onto $U^\perp$, of rank q. By Lemma 20, $\operatorname{tr}(\Xi_i P_U) = \operatorname{tr} \Xi_i - \operatorname{tr}(\Xi_i Q) \ge \sum_j \lambda_j(\Xi_i) - \sum_{j \le q} \lambda_j(\Xi_i) = \sum_{j > q} \lambda_j(\Xi_i)$, with equality when $U^\perp$, the range of Q, is a top-q eigenspace of Ξ_i. Such a subspace exists, since Ξ_i has an orthonormal eigenbasis, so the minimum is attained and equals $\sigma^2 \sum_{j>q} \lambda_j(\Xi_i)$.

(b) For a shared U, linearity gives $\sum_i D_i(U) = \sigma^2 \sum_i \operatorname{tr}(\Xi_i P_U) = \sigma^2 \operatorname{tr}(\bar{\Xi} P_U)$. Exchanging the order of summation, $\bar{\Xi} = \sum_i \sum_{t \in O_i} \tilde{x}_t \tilde{x}_t^\top = \sum_t \#\{i : t \in O_i\} \, \tilde{x}_t \tilde{x}_t^\top = \sum_t n_t \tilde{x}_t \tilde{x}_t^\top$. The matrix $\bar{\Xi}$ is symmetric positive semidefinite, and the argument of (a) with $\bar{\Xi}$ in place of Ξ_i gives the minimiser and the minimum.

(c) The allocation $\operatorname{codim} U_i = q$ for every i has total codimension mq, so it is admissible, and choosing each U_i by (a) gives the total $\sigma^2 \sum_i \sum_{j>q} \lambda_j(\Xi_i)$; the minimum over admissible allocations is at most this value. For every subspace U of codimension q, $\sum_i D_i(U) \ge \sum_i \min_{\operatorname{codim} U_i = q} D_i(U_i) = \sigma^2 \sum_i \sum_{j>q} \lambda_j(\Xi_i)$, because each term is at least its own minimum; taking the minimum over U gives the second inequality. Since the allocation used has total codimension exactly mq, the comparison also holds among per-unit subspaces of the same total codimension as the shared one.

(d) Since $S_i \subseteq O_i$, $\operatorname{span}\{\tilde{x}_t : t \in S_i\}$ is contained in the row space of $\tilde{X}_{O_i}$, and by hypothesis the two are equal. Hence $U_i = \ker \tilde{X}_{S_i} = (\operatorname{span}\{\tilde{x}_t : t \in S_i\})^\perp$ equals the orthogonal complement of the row space of $\tilde{X}_{O_i}$, which is $\ker \tilde{X}_{O_i}$, of codimension r_i by rank–nullity. Every $v \in U_i$ has $\tilde{X}_{O_i} v = 0$, so $\tilde{X}_{O_i} P_{U_i} = 0$, hence $\Xi_i P_{U_i} = 0$ and $D_i(U_i) = 0$. Since $D_i(U) \ge 0$ for every U, this is the minimum. If $C \ge |O_i|$, then $S_i = O_i$ by Proposition 8b, and the hypothesis holds. □

Proposition 11: the freed term is a near-wall term. Here unit i, the base inputs and an update $u \in \mathbb{R}^{d+1}$ of row i are fixed, so that token t has new pre-activation $\hat{z}_t = z_i(x_t) + a_t$ with $a_t = \langle u, \tilde{x}_t \rangle$. For a closed token $t \notin O_i$, $\gamma_t = \theta - z_i(x_t) \ge 0$ is its margin. A closed token *opens* if $\hat{z}_t > \theta$, and the closed→open term is $\sum_{t \in S_{-+}} \hat{z}_t^2$.

Proposition (Full form of Proposition 11)**.** ***(a)*** *A closed token t opens if and only if $a_t > \gamma_t$, and the closed→open term equals $\sum_{t \notin O_i,\, a_t > \gamma_t} (\theta + a_t - \gamma_t)^2$.*

(b) *Only near-wall tokens, those of $N(u) = \{t \notin O_i : \gamma_t < \|u\| \, \|\tilde{x}_t\|\}$, can open.*

(c) *At $\theta = 0$ the term is $\sum_{t \notin O_i} (a_t - \gamma_t)_+^2 \le \|\tilde{X}_{N(u)} u\|^2$.*

Proof. (a) For a closed token, $\hat{z}_t = z_i(x_t) + a_t = \theta - \gamma_t + a_t$, which exceeds θ if and only if $a_t > \gamma_t$. The opened tokens form S_{-+}, and each contributes $\hat{z}_t^2 = (\theta + a_t - \gamma_t)^2$.

(b) By the Cauchy–Schwarz inequality $a_t \le \|u\| \, \|\tilde{x}_t\|$, so an opened token, which has $a_t > \gamma_t$, satisfies $\gamma_t < \|u\| \, \|\tilde{x}_t\|$ and lies in $N(u)$.

(c) At $\theta = 0$, (a) gives the term $\sum_{a_t > \gamma_t} (a_t - \gamma_t)^2 = \sum_{t \notin O_i} (a_t - \gamma_t)_+^2$. An opened token has $0 < a_t - \gamma_t \le a_t$, because $\gamma_t \ge 0$, and lies in $N(u)$ by (b); so the term is at most $\sum_{\text{opened}} a_t^2 \le \sum_{t \in N(u)} a_t^2 = \|\tilde{X}_{N(u)} u\|^2$. □

Proposition 12: first-order criteria inherit the partition. Here the layer's inputs $\tilde{x}_t$ are fixed; the matrix $Z = (z_{t,i}) \in \mathbb{R}^{T\times m}$ has $z_{t,i} = \langle \tilde{w}_i, \tilde{x}_t \rangle$, the post-activations are $h = (h_{t,i}) = \mathrm{act}_\theta(Z)$ entrywise, and $L = g(h)$ for a function $g : \mathbb{R}^{T\times m} \to \mathbb{R}$, the downstream network per token or per sequence, whose partial derivatives exist at h. With the convention $\mathrm{act}'_\theta(\theta) := 0$ of automatic differentiation, $\partial L/\partial z_{t,i} := \mathrm{act}'_\theta(z_{t,i})\, \partial g/\partial h_{t,i}$ and $\nabla_{\tilde{w}_i} L := \sum_t (\partial L/\partial z_{t,i})\, \tilde{x}_t$ are the values that automatic differentiation returns.

Proposition (Full form of Proposition 12)**.** ***(a)*** *The derivative is $\partial L/\partial z_{t,i} = \mathbb{1}[z_{t,i} > \theta]\, \partial g/\partial h_{t,i}$, the true partial derivative wherever $z_{t,i} \neq \theta$; it vanishes on every closed pair.*

(b) *The row gradient $\nabla_{\tilde{w}_i} L = \sum_{t\in O_i} (\partial L/\partial z_{t,i})\, \tilde{x}_t$ lies in the row space of $\tilde{X}_{O_i}$, and it is 0 for a unit open on no token.*

(c) *The Fisher $F_i = \mathbb{E}[\nabla_{\tilde{w}_i} \ell\, \nabla_{\tilde{w}_i} \ell^\top]$ of row i, over per-sample losses ℓ of this form, annihilates $\ker \tilde{X}_{O_i}$, with O_i taken over the Fisher data.*

(d) *EWC's diagonal $F_{i,jj} = \mathbb{E}[(\sum_{t\in O_i} \partial \ell/\partial z_{t,i}\, \tilde{x}_{t,j})^2]$, $j \in [d+1]$, and its per-token form depend on the data only through open pairs: every term they contain is indexed by an open pair of unit i. Both vanish for a unit open on no token.*

(e) *A selector that takes the k_i largest $|\partial L/\partial z_{t,i}|$ of unit i picks only open pairs whenever at least k_i open pairs of unit i have $\partial g/\partial h_{t,i} \neq 0$.*

Proof. The inputs $\tilde{x}_t$ of the layer do not depend on its own weights, and L depends on Z only through $h = \mathrm{act}_\theta(Z)$.

(a) On $\{v \in \mathbb{R} : v \neq \theta\}$ the scalar map act_θ is locally affine with derivative $\mathbb{1}[v > \theta]$; at $v = \theta$ it jumps if $\theta > 0$ and has a kink if $\theta = 0$. Let $z_{t,i} \neq \theta$. For $|\eta| < |z_{t,i} - \theta|$, changing $z_{t,i}$ to $z_{t,i} + \eta$ changes only the entry $h_{t,i}$ of h, and by $\eta \mathbb{1}[z_{t,i} > \theta]$, so the chain rule gives $\partial L/\partial z_{t,i} = \mathbb{1}[z_{t,i} > \theta]\, \partial g/\partial h_{t,i}$, which is 0 when $z_{t,i} < \theta$. With $\mathrm{act}'_\theta(\theta) := 0$ the same formula holds at pairs with $z_{t,i} = \theta$ and gives 0 there, so the derivative vanishes on the whole closed set $\{z_{t,i} \leq \theta\}$.

(b) By definition and (a), $\nabla_{\tilde{w}_i} L = \sum_t (\partial L/\partial z_{t,i}) \tilde{x}_t = \sum_{t\in O_i} (\partial L/\partial z_{t,i}) \tilde{x}_t$, a linear combination of the rows of $\tilde{X}_{O_i}$, and the empty combination 0 when $O_i = \varnothing$.

(c) Each sample has its own tokens and downstream map, the expectation defining F_i is assumed to exist, and O_i is the set of tokens, over all samples, on which unit i is open, so that $\ker \tilde{X}_{O_i} = \{v : \langle v, \tilde{x}_t \rangle = 0 \text{ for all } t \in O_i\}$. For such v, (b) applied to each sample gives $\langle \nabla_{\tilde{w}_i} \ell, v \rangle = 0$ for every sample, hence $F_i v = \mathbb{E}[\nabla_{\tilde{w}_i} \ell\, \langle \nabla_{\tilde{w}_i} \ell, v \rangle] = 0$.

(d) The j-th diagonal entry of F_i is $\mathbb{E}[(\nabla_{\tilde{w}_i} \ell)_j^2]$, which by (b) equals $\mathbb{E}[(\sum_{t\in O_i} \partial \ell/\partial z_{t,i}\, \tilde{x}_{t,j})^2]$, with O_i the open set of the sample; every term is indexed by an open pair, and the entry is 0 when unit i is open on no token. In the per-token form $\mathbb{E}_t[(\partial \ell_t/\partial z_{t,i}\, \tilde{x}_{t,j})^2]$ the terms of closed pairs are 0 by (a).

(e) By (a), every closed pair has $|\partial L/\partial z_{t,i}| = 0$, and an open pair has $|\partial L/\partial z_{t,i}| = |\partial g/\partial h_{t,i}|$. If at least k_i open pairs have $\partial g/\partial h_{t,i} \neq 0$, then the k_i-th largest value is positive, so every one of the k_i selected pairs, whatever the tie-breaking, has a positive value and is open. □

Lemma 21 (Gated units)**.** *Let $\theta \geq 0$, and let unit i of a gated layer output $y_t = \mathrm{act}_\theta(g_t)\, v_t$ with $g_t = \langle \tilde{w}^g, \tilde{x}_t \rangle$ and $v_t = \langle \tilde{w}^v, \tilde{x}_t \rangle$ for gate and up weights $\tilde{w}^g, \tilde{w}^v \in \mathbb{R}^{d+1}$, and let $O = \{t : g_t > \theta\}$. If the updates $u^g, u^v \in \mathbb{R}^{d+1}$ satisfy $\tilde{X}_O u^g = \tilde{X}_O u^v = 0$, then at the base inputs $\hat{y}_t = y_t$ for every $t \in O$, and for $t \notin O$, $\hat{y}_t \neq y_t = 0$ only if $\hat{g}_t > \theta$.*

Proof. At the base inputs $\hat{g}_t = g_t + \langle u^g, \tilde{x}_t \rangle$, $\hat{v}_t = v_t + \langle u^v, \tilde{x}_t \rangle$ and $\hat{y}_t = \mathrm{act}_\theta(\hat{g}_t)\hat{v}_t$. For $t \in O$ the hypothesis gives $\hat{g}_t = g_t > \theta$ and $\hat{v}_t = v_t$, so $\hat{y}_t = g_t v_t = y_t$. For $t \notin O$, $g_t \leq \theta$ gives $y_t = 0$, and $\hat{y}_t = \mathrm{act}_\theta(\hat{g}_t)\hat{v}_t$ is 0 unless $\hat{g}_t > \theta$. □

C Fine-tuning setup and implementation

Models and text. Every run updates only the `fc1` matrix (weights and bias) of every decoder block of the pre-LN OPT-125m or OPT-1.3b model (Zhang et al., 2022), in `fp32`; its rows are the

wall units of Definition 1. On OPT-6.7b it updates the `fc1` matrices of the last four of the 32 blocks, in `fp32`, with the rest of the model in `bfloat16`; its units fire on 6.9% of the support tokens (a mean of 140.7 of 2 048, against 72.1 on OPT-1.3b), and cap 128 gives $\bar{k} = 68.7$ and preserves 67.4% of the active post-activation energy exactly. Text sets are disjoint character ranges. The new task trains on characters $[0,\ 400\,000)$ of its corpus file and is evaluated on $[600\,000,\ 1\,200\,000)$ (64 windows of 128 tokens); the old support X_{old}, on which every projector is built, is MiniPile $[0,\ 400\mathrm{k})$ (Kaddour, 2023), $16 \times 128 = 2\,048$ tokens, and the old probe, on which forgetting is scored, is MiniPile $[600\mathrm{k},\ 1.0\mathrm{M})$, $8 \times 128 = 1\,024$ tokens.

Training and measurements. The new task is 512 fresh 128-token windows presented once, four per AdamW step (accumulation 4, 128 steps), with 10 warmup steps, cosine decay to zero and peak learning rate 10^{-4} on every corpus and method. Data order is permuted under the seed, so two methods on one seed see the same windows in the same order; seeds are 0–2, and 0–5 on the settings with $n = 6$. Every number is read at step 128: *forgetting* is the old probe's NLL minus its base value and *learning* the held-out new-task NLL minus its base value (more negative is better), and every run logs the layer form of Eq. (2) at every evaluation (Appendix J). Base-model difficulty is 0.80 nats per character on WikiText-103 (Merity et al., 2017) (`wt103`) and 1.02 on `python`, Python files from the validation split of CodeParrot-clean (CodeParrot, 2022). On `wt103` unconstrained fine-tuning of OPT-1.3b forgets $+0.2684 \pm 0.0192$ and learns -0.5878 ± 0.0141 over three seeds, and OPT-125m forgets $+0.1958$; on `python` it forgets $+0.1154$, while every constrained method ends below base on the old probe, so there a more negative forgetting number is a larger improvement.

The projector. It is built once from the base model. For each wall the base inputs on X_{old}, with a ones column appended so that the bias is projected with the weights, form $X_a \in \mathbb{R}^{T\times(d+1)}$, $T = 2\,048$. For unit i, with base post-activations $h_t^{(i)} = \mathrm{act}_\theta(z_i(x_t))$ and open set $O_i = \{t : z_i(x_t) > \theta\}$, S_i holds the tokens of the C largest positive $h_t^{(i)}$, so $k_i = |S_i| = \min(C, |O_i|)$ as in Eq. (3), and $A_i = X_a[S_i, :]$. The Gram $\Gamma_i = A_i A_i^\top$ is gathered from $X_a X_a^\top$ into a $C \times C$ array $\widetilde{\Gamma}_i = \mathrm{diag}(\Gamma_i, I)$ padded by the identity, so that its factorisation is defined for every unit, and the ridge is $\rho_i = 10^{-4}\, C^{-1} \operatorname{tr} \widetilde{\Gamma}_i$, that is 10^{-4} times the mean of the padded diagonal. The Cholesky factor of $\widetilde{\Gamma}_i + \rho_i I$ is stored per wall in `fp16`, $2mC^2$ bytes, and cast to `fp32` for each solve; caps 256 and 512 re-factorise inside each step in blocks of 1 024 units instead, and at $C = 128$ the two paths agree to 12 decimal places. After each optimiser step, with D the cumulative displacement of the wall and its bias (row $u_i^\top = (dW_i, db_i)$), unit i's entries of $DX_a^\top$ at S_i, padded with zeros, form r_i; $y_i = (\widetilde{\Gamma}_i + \rho_i I)^{-1} r_i$ is scattered back to the positions S_i of a matrix Y; and $D \leftarrow D - \alpha\, Y X_a$. Because the padding block is decoupled and carries a zero right-hand side, row i of this update is

$$u_i \;\leftarrow\; u_i - \alpha\, A_i^\top \big(\Gamma_i + \rho_i I\big)^{-1} A_i\, u_i,$$

which is Eq. (4), with $\alpha = 1$ for every gate-aware run. The residual the ridge leaves at the stored base inputs, the summed $|\hat{z} - z|$ on the constrained pairs as a share of the total $|\hat{z} - z|$ mass, is 5.4×10^{-9} to 4.0×10^{-8} on OPT-1.3b at caps 16 to 128, between 3.8×10^{-9} and 3.5×10^{-7} over all 48 cap-ladder runs on both models, and at most 1.3×10^{-8} on every run of the necessity comparison (§4.2). At fixed inputs every pair in S_i keeps its base post-activation, so the cap fixes the share $\sum_i \sum_{t\in S_i} (h_t^{(i)})^2 \,/\, \sum_i \sum_t (h_t^{(i)})^2$ of the active post-activation energy preserved exactly (per wall, averaged over walls): 59.6% at $C = 64$ and 71.7% at $C = 128$ on OPT-1.3b.

Baselines and run names. Adam-NSCL (Wang et al., 2021), restricted to the walls, diagonalises the uncentred second moment $T^{-1} X_a^\top X_a$ of the same augmented support inputs, forbids its top k_{sh} eigendirections in every row, and after each optimiser step replaces the cumulative displacement by $(1 - \alpha)D + \alpha DP$, with P the projector onto their complement. The random-subspace control replaces the data-derived subspace by a random orthonormal one of exactly the same dimension. Tables name the unconstrained run `B1`, a shared-subspace run by its dimension (`B16k24`: $k_{\mathrm{sh}} = 24$) and a gate-aware run by its cap (`B17c64`: $C = 64$); the selector controls, which keep each unit's k_i and change only which support tokens fill it, carry a one-letter suffix (Appendix E).

C.1 THE FREED TERM ON HELD-OUT TEXT

On the old probe, which neither the projector nor the new task sees, every constrained method lowers the freed closed→open mass to about a fifth of unconstrained fine-tuning's or less, and lowers its share of the post-activation change, on all three seeds (Table 2).

Table 2: **Closed→open mass on the held-out old probe.** OPT-1.3b, WikiText-103 at lr 10^{-4}, `fc1`, seeds 0–2. "co mass" is the run's mean level; the difference columns are mean paired per-seed differences against the unconstrained run, negative on 3 of 3 seeds in every row. "Share" is of the probe's total squared post-activation change.

run	constraint	dims/row	co mass	Δ mass	Δ share
`B1`	none	0	9 239	—	—
`B17c16`	gate, $C = 16$	9.1	2 286	−6 953	−0.0537
`B17c64`	gate, $C = 64$	23.5	1 785	−7 454	−0.0433
`B17c128`	gate, $C = 128$	34.4	1 695	−7 544	−0.0366
`B16k9`	shared, $k_{\mathrm{sh}} = 9$	9.0	2 196	−7 043	−0.0772
`B16k24`	shared, $k_{\mathrm{sh}} = 24$	24.0	1 622	−7 617	−0.0910
`B16k34`	shared, $k_{\mathrm{sh}} = 34$	34.0	1 506	−7 733	−0.0927

C.2 STATISTICS

Two methods on one model and corpus are compared by a paired t-test on their per-seed differences (mean, 95% interval, p, number of seeds with the predicted sign), one-sided where the construction predicts the direction (ours below the shared subspace at a matched budget; the gate below a control selector at identical k_i) and two-sided otherwise, including every learning comparison. Forgetting removed is the paired reduction against the unconstrained run of the same model, divided by that run's forgetting. Equivalence is two one-sided tests (TOST) with margin $\delta = 0.0068$ nats, which is 2.5% of the unconstrained forgetting of OPT-1.3b on WikiText-103 (0.2684) and 5.4% of that of OPT-6.7b (0.1256), and equals the probe noise floor of the repair experiments (Appendix I): the 90% interval of the paired difference lies inside $(-\delta, \delta)$. A table of several rows is also tested as a whole, by a sign-flip permutation test over all its seed-pairs (each difference standardised by its row's paired standard deviation, the sum referred to all 2^n sign assignments for $n \le 22$ pairs and to 2×10^6 random ones otherwise) and by a DerSimonian–Laird random-effects meta-analysis of the row means (pooled difference, 95% interval, between-row variance τ^2).

D MATCHED BUDGET AND CAPACITY IN FULL

Table 3 pairs each cap, with mean per-unit dimension $\bar{k}$ (Eq. (3)), with Adam-NSCL at $k_{\mathrm{sh}} = \mathrm{round}(\bar{k})$, so both remove the same number of directions per row (training setup and tests of Appendix C). On OPT-1.3b the ratio of the two forgetting levels, means over each row's seeds, grows from 1.1 at 9 dimensions per row to 4.3 at 60: between 15 and 47 dimensions per row Adam-NSCL stays at 0.046–0.050 nats (seeds 0–2), while ours falls from 0.047 to 0.011 at 47 and 0.010 at 60, and learning differs by at most 0.0026 nats. On OPT-6.7b, cap 128 at 68.7 dimensions per row forgets 0.0192 against Adam-NSCL's 0.0184 at 69 (TOST at $\delta = 0.0068$: 90% CI $[-0.0001, +0.0018]$, equivalent) and learns 0.0109 nats more on all three seeds.

D.1 CAPACITY

On OPT-1.3b we set our caps against Adam-NSCL at a uniform $k_{\mathrm{sh}} = 259$ in every row (`B16k259`), the mean dimension GPM's energy criterion selects at 0.95 on the same support (259.2), so ours constrains 7.5× fewer dimensions per row at cap 128 ($\bar{k} = 34.4$) and 5.5× fewer at cap 256 ($\bar{k} = 47.3$). At cap 256 ours forgets 0.0110 against Adam-NSCL's 0.0255 on WikiText-103, a paired difference of −0.0145 (95% CI $[-0.0229, -0.0062]$, one-sided $p = 0.0087$, 3 of 3), and learns 0.0518 nats more ($[-0.0576, -0.0460]$, $p = 0.0007$). At cap 128, on Python ours forgets 0.0216 nats less (95% CI $[-0.0343, -0.0090]$, two-sided $p = 0.0180$) and learns 0.0265 nats more

Table 3: **Matched constrained dimensions per row.** Ours at cap C (mean per-unit dimension $\bar{k}$) against Adam-NSCL at k_{sh}, paired by seed; Δ is ours minus Adam-NSCL, so a negative Δ is ours ahead. Forgetting: one-sided paired t-test; learning: two-sided; $\Delta < 0$ counts the seed-pairs on which ours forgets less. Bold: the 95% CI excludes zero with ours ahead. WikiText-103, `fc1`, lr 10^{-4}, $T = 2\,048$; each block has its own no-constraint run, and the pooled lines combine the rows named.

				forgetting			learning		
C	$\bar{k}$	k_{sh}	n	Δ	95% CI	p $\Delta<0$	Δ	95% CI	p
OPT-1.3b (gate density 0.0392, ambient dim. 2 049; no constraint forgets $+0.2684$)									
16	9.067	9	6	-0.0068	$[-0.0246, +0.0109]$	0.1839	4/6	$+0.0011$ $[-0.0011, +0.0033]$	0.2691
32	15.048	15	3	-0.0027	$[-0.0126, +0.0073]$	0.1845	2/3	$+0.0011$ $[-0.0082, +0.0104]$	0.6679
64	23.519	24	6	$\mathbf{-0.0111}$	$[-0.0147, -0.0074]$	0.0003	6/6	$+0.0015$ $[-0.0002, +0.0033]$	0.0769
128	34.421	34	6	$\mathbf{-0.0195}$	$[-0.0247, -0.0143]$	0.0001	6/6	$+0.0025$ $[-0.0002, +0.0051]$	0.0654
256	47.257	47	3	$\mathbf{-0.0357}$	$[-0.0481, -0.0232]$	0.0033	3/3	$+0.0011$ $[-0.0004, +0.0025]$	0.0844
512	60.332	60	3	$\mathbf{-0.0325}$	$[-0.0485, -0.0164]$	0.0065	3/3	$+0.0026$ $[+0.0012, +0.0040]$	0.0146
OPT-125m (gate density 0.0606, ambient dim. 769; no constraint forgets $+0.1958$)									
8	5.700	6	3	$+0.0348$	$[+0.0279, +0.0417]$	0.9990	0/3	$+0.0066$ $[+0.0008, +0.0124]$	0.0389
32	17.022	17	6	$+0.0007$	$[-0.0039, +0.0052]$	0.6373	2/6	$\mathbf{-0.0226}$ $[-0.0249, -0.0203]$	$< 10^{-4}$
128	42.080	42	6	$\mathbf{-0.0047}$	$[-0.0073, -0.0021]$	0.0027	6/6	$\mathbf{-0.0084}$ $[-0.0100, -0.0069]$	$< 10^{-4}$
256	62.047	62	6	$\mathbf{-0.0102}$	$[-0.0138, -0.0067]$	0.0003	6/6	-0.0003 $[-0.0024, +0.0019]$	0.7599
OPT-6.7b, last four blocks (gate density 0.0677, ambient dim. 4 097; no constraint forgets $+0.1256$)									
128	68.724	69	3	$+0.0008$	$[-0.0006, +0.0023]$	0.9359	0/3	$\mathbf{-0.0109}$ $[-0.0115, -0.0103]$	0.0001

Pooled forgetting (sign-flip permutation over seed-pairs, one-sided; random effects, DerSimonian–Laird). OPT-1.3b, six rows: $24/27$ lower, $p = 1.5 \times 10^{-6}$ (2×10^6 Monte Carlo); $\Delta = -0.0182$ $[-0.0279, -0.0086]$, $\tau^2 = 1.3 \times 10^{-4}$. OPT-125m from 17 dims/row, three rows: $14/18$ lower, exact $p = 5.1 \times 10^{-4}$; $\Delta = -0.0049$ $[-0.0102, +0.0005]$.

($[-0.0380, -0.0151]$, $p = 0.0099$), each on 3 of 3 seed-pairs; on WikiText-103, over 6 seed-pairs, it learns 0.0523 nats more ($[-0.0546, -0.0499]$, $p < 10^{-4}$, 6 of 6) at a forgetting difference of -0.0065 ($[-0.0136, +0.0006]$, two-sided $p = 0.0645$), negative on 5 of 6. Against no constraint, cap 256 removes 95.9% of OPT-1.3b's forgetting, where Adam-NSCL removes 82.6% at 47 dimensions per row and 90.5% at 259, and 97.7% of OPT-125m's, each method on 3 of 3 seeds.

D.2 The reachable budget

Table 4: **The reachable budget: base gate density and per-unit open sets.** p is the fraction of (token, unit) pairs open at the base weights on the 1 024-token old probe. $\mathrm{open}_i = |O_i|$ is counted directly on the $T = 2\,048$-token support that builds the projector; its mean $\overline{\mathrm{open}}$ over units is the right-hand side of Eq. (6) and bounds $\bar{k}$ (pT is about 11% higher because p is measured on the probe). "Ambient" is the row dimension $d + 1$.

					open_i over units				
model	p	ambient	$\overline{\mathrm{open}}$	pT	median	p25	p99	never open	mean/median
OPT-1.3b	0.0392	2 049	72.1	80	11	1	954	19.5%	6.6
OPT-125m	0.0606	769	111.3	124	16	2	1 320	14.8%	7.0

Since $k_i \le |O_i|$ for every unit, the mean constrained dimension per row is bounded by the mean open-set size on the support,

$$\bar{k} \;=\; \frac{1}{m}\sum_{i=1}^{m} k_i \;\le\; \frac{1}{m}\sum_{i=1}^{m} |O_i| \;=\; p_{\mathrm{sup}}\, T, \tag{6}$$

with p_{sup} the base gate density counted on the support: at $T = 2\,048$ this reachable budget is 72.1 dimensions per row on OPT-1.3b and 111.3 on OPT-125m (Table 4), while a shared k_{sh} ranges up to $d + 1$. The cap binds on a heavy tail: on OPT-1.3b the 90th percentile unit fires on 187 support tokens and the 99th on 954, and 7.5% of units fire on more than 259. The budget is fixed by the data

alone: from a single source at $T = 2\,048$ the projector's setup returns 259.2 dimensions per row for the shared subspace at energy fraction 0.95 and 34.4 for our cap 128, and our cap 64 rises from 23.5 dimensions per row at $2\,048$ tokens to 55.8 at $49\,152$.

E What the criterion is: the selector controls in full

Figure 6 draws, on one random layer at one mean budget per row, the (token, unit) pairs on which each constraint family of §4.3 spends its budget, and in panel (e) the tokens that each selector control of §4.2 picks for one unit. The controls hold each unit's k_i at the gate's value and vary only which of its support tokens fill it; they share one no-constraint run `B1` (seeds 0–2), run on seeds 0–5 (`openrand` on 0–2) and use the tests of Appendix C.2. `gate` takes the k_i largest base post-activations (Eq. (3)); `openrand` draws k_i tokens uniformly from the unit's open set; `grad` takes the largest $|\partial L_{\text{old}}/\partial z|$ at the base model; `absz` the largest $|z - \theta|$ over all tokens, which lands almost entirely on closed pairs because the negative tail of z is far longer; `closed` the most closed tokens; and the random selector k_i tokens uniformly from the whole support.

Reading the table. In Table 5, off the partition the gate forgets less than each of the three selectors on all 18 seed-pairs; on it, `openrand` and `grad` are TOST-equivalent to the gate (Table 1) while preserving 35.2% and 35.8% of the active energy against 59.6%; and across all selectors the open share of the selected set orders forgetting while preserved energy does not (`absz` preserves more than the random selector and forgets more), so the partition matters, not a ranking inside it or the energy it covers. Learning differences across the five OPT-1.3b contrasts lie between -0.0024 and $+0.0045$ nats, and the gate learns more than `grad` on all six seeds, by 0.0024 nats (95% CI $[-0.0039, -0.0008]$, two-sided $p = 0.0107$). `grad` inherits the partition through the chain rule (Proposition 12), placing 99.6% of its pairs on open pairs, but needs a backward pass and a $T \times m$ gradient tensor per wall ($2\,048 \times 8\,192 \times 24$ in `fp32`, 1.6 GB), where the gate reads the sign of $z - \theta$ off one forward pass.

E.1 Freezing whole units

We freeze the `fc1` rows (weights and bias) of the units with the largest old-support activity $\sum_t h_i(t)^2$ and train the rest without constraint. A frozen row removes all $d + 1$ of its directions, so freezing 1.15% of units (94 per wall) costs 23.51 dimensions per row, the budget of cap 64, 1.68% (138 units) costs 34.52, that of cap 128, and the smallest top set carrying the 59.6% of active energy that cap 64 preserves exactly needs 2 to 464 units per wall, 69.34 dimensions per row. At equal budget freezing forgets about eight times as much as the per-unit projector on the shared seeds and learns 0.036–0.039 nats less, and at three times the budget it still forgets five times as much (Table 6).

F Comparison with other method families

Every family runs under the training setup of Appendix C on the `fc1` walls of OPT-1.3b with WikiText-103 as the new task; under a single transition AdNS (Kong et al., 2022) reduces to Adam-NSCL with a rank cap and one scalar strength. A random orthonormal subspace of exactly the data-derived dimension per wall (`B16e99r`), at matched $\|\Delta W\|_F$ (12.1 against 12.2), forgets $+0.2176$, against $+0.2684$ for no constraint and -0.0045 for the real projector at the same rank, and at the 259-dimension rank the corresponding random control forgets $+0.2318$ against the projector's $+0.0227$. Two-sided activation anchoring on the walls (LDIFS (Mukhoti et al., 2024), $\lambda = 100$) forgets $+0.1720$, fresh old-corpus replay (Ibrahim et al., 2024) $+0.1721$ on a quarter of the micro-batches and $+0.0314$ on every one, and replaying the $2\,048$ support tokens themselves $+0.7247$. Relaxing Adam-NSCL at $k_{\text{sh}} = 259$ to $(1 - \alpha)D + \alpha DP$ with $\alpha \in \{1, 0.8, 0.5\}$ moves its learning from -0.5671 to -0.5697 and -0.5719 (seeds 0–2), recovering 0.005 of the 0.054 nats by which our cap 128 learns more (Table 7).

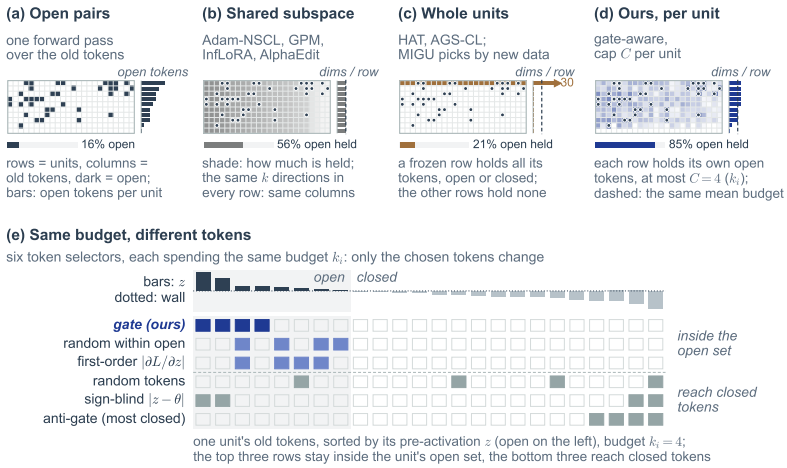


Figure 6: **Which (token, unit) pairs a constraint holds, at one budget.** A random ReLU layer with $d = 30$ inputs, $m = 10$ units and $T = 24$ old tokens. **How to read a grid.** Each row is one unit, that is one row of the layer's first weight matrix (`fc1` in OPT), and each column is one old token, so the cell in row i and column t is the pair (t, i). Rows are sorted by how many tokens the unit fires on, most active on top. Columns are sorted by how much the shared subspace of (b) holds them, most held on the left. The order is the same in every grid, so a given cell sits at the same place in (a)–(d). **(a)** The gate pattern of the base model, read off one forward pass over the old tokens: a cell is dark when unit i is open on token t, that is $z_i(x_t) > \theta$, with θ the gate threshold (0 for ReLU). The bar to the right of a row counts that unit's open tokens. The top unit fires on 8 of the 24 tokens, the next on 6, and the bottom unit on none. Only 38 of the 240 pairs, 16%, are open; at OPT-1.3b the share is 4%. **(b)–(d)** A cell's shade is the fraction of an isotropic weight update's effect on that pair's pre-activation that the constraint removes: white means the pair is left free, and full colour means it is held exactly. The dots repeat the open pairs of (a). The bar to the right of a row is the number of directions the constraint forbids in that row, and the dashed line marks the mean, which is 3 for every family. The bar under a grid is the mean shade over the open pairs. Holding a closed pair guards only against its opening. **(b)** The shared subspace of Adam-NSCL, GPM, InfLoRA and AlphaEdit forbids the top 3 eigendirections of the old inputs' second moment, and it forbids them in every row. Every row therefore spends 3 directions, and a cell's shade depends only on its column: the leftmost token is held at 83% in every row and the rightmost at 8%. The top unit's 8 open tokens are held at between 9% and 78%. Overall 56% of the open pairs' effect is held. **(c)** Methods that protect whole units freeze entire rows; HAT and AGS-CL choose them by old-task use, and MIGU by new-data magnitude. A frozen row spends all 30 of its directions (arrow), so a mean of 3 buys one frozen row out of ten. The top row is frozen. All 24 of its tokens are held, including the 16 on which it is closed, and no other row holds anything, so 21% of the open pairs' effect is held. **(d)** Our projector gives unit i the budget $k_i = \min(C, \#\text{open})$ with cap $C = 4$, and spends it on that unit's own open tokens with the largest activations. The top unit therefore holds its 4 most active open tokens exactly, and its other 4 open tokens only in part, at 54%, 22%, 14% and 10%. A unit with 3 open tokens holds all 3 exactly, and the bottom unit, which never fires, receives $k_i = 0$ and holds nothing. The budgets, top to bottom, are $4, 4, 4, 4, 4, 3, 3, 3, 1, 0$, which also average 3, and 85% of the open pairs' effect is held. **(e)** The selector controls of §4.2, drawn for the top unit. Its 24 old tokens are sorted by pre-activation z, shown by the bars with the open tokens on the left, and the dotted line is the wall. Eight are open, and its budget is $k_i = 4$. Each row marks the 4 tokens that one selector spends the budget on. The gate takes the 4 most active open tokens. Random-within-open takes 4 of the 8 open tokens at random, here sharing one with the gate. The first-order criterion takes the 4 tokens with the largest $|\partial L/\partial z|$; since that derivative is zero on closed tokens, it also stays inside the open set. Random tokens draws 4 of all 24, and here lands on 1 open and 3 closed tokens. The sign-blind criterion takes the 4 largest $|z - \theta|$, which are the 2 most open and the 2 most deeply closed tokens, and the anti-gate takes the 4 most deeply closed. Every row spends the same k_i; only the tokens change.

Table 5: **The selector controls at identical per-unit constraint dimension.** OPT-1.3b, wt103, $C = 64$, $\bar{k} = 23.5$. **(a)** The selected set on the base model (*overlap* with the gate set, share *open*, share of the active post-activation *energy* preserved exactly) and outcomes, means over n seeds. **(b)** Gate minus rank-matched random tokens on OPT-125m, on Python (OPT-1.3b) and on OPT-6.7b ($\bar{k} = 68.7$), paired by seed, 95% interval and one-sided p; the first row pools the three off-partition contrasts of Table 1 (random-effects difference, exact sign-flip p).

(a)

		selected set, base model				outcome	
run	selector	overlap	open	energy	n	forget	learn
B1	no constraint	—	—	—	3	+0.2684	−0.5878
B17c64	gate	100%	100%	59.6%	6	+0.0247	−0.6240
B17c64o	openrand	69.9%	100%	35.2%	3	+0.0324	−0.6223
B17c64g	grad	70.1%	99.6%	35.8%	6	+0.0222	−0.6216
B17c64r	random	2.3%	8.3%	2.8%	6	+0.0563	−0.6285
B17c64n	absz	3.0%	3.0%	10.4%	6	+0.0821	−0.6270
B17c64x	closed	0.0%	0.2%	0.1%	6	+0.0906	−0.6275

(b)

setting	n	Δforget [95% CI]	test	seeds	Δlearn [95% CI]
1.3b, pooled	18	−0.0515 [−0.0740, −0.0289]	$p = 3.81\times10^{-6}$	18/18	—
125m, $C = 32$	3	−0.0114 [−0.0162, −0.0067]	$p = 0.0046$	3/3	+0.0254 [+0.0154, +0.0355]
python, $C = 64$	6	−0.0224 [−0.0320, −0.0129]	$p = 0.0009$	6/6	+0.0069 [+0.0034, +0.0104]
6.7b, $C = 128$	3	−0.0098 [−0.0150, −0.0047]	$p = 0.0072$	3/3	+0.0082 [+0.0070, +0.0094]

Table 6: **Freezing whole units against the per-unit projector.** OPT-1.3b, WikiText-103, fc1, lr 10^{-4}. Δ is ours minus the freezing run, paired by seed; negative means ours forgets less (one-sided p) or learns more (two-sided p); the fraction counts seed pairs with $\Delta < 0$.

freezing setting	matched to	dims/row	n	forget, learn	Δ forget (p)	Δ learn (p)
top 1.15% of units	cap 64 (23.5)	23.51	6	+0.2101, −0.5851	−0.1855 (6/6, $< 10^{-4}$)	−0.0389 (6/6, $< 10^{-4}$)
top 1.68% of units	cap 128 (34.4)	34.52	3	+0.1951, −0.5853	−0.1713 (3/3, 0.0014)	−0.0357 (3/3, 0.027)
59.6% of active energy	cap 64 (energy)	69.34	3	+0.1719, −0.5871	−0.1380 (3/3, 0.0006)	−0.0367 (3/3, 0.054)

F.1 EWC, MIGU, LoRA and LoRA-Null

Table 7 gives every operating point of Figure 3, each paired by seed against our cap 128. The figure shows EWC across $\lambda \in \{4, \ldots, 4\times10^5\}$, LoRA at ranks 8, 32, 128 and learning rates 2×10^{-4} and 10^{-3} (and rank 32 at 5×10^{-4}), LoRA-Null at ranks 32 and 128 (2×10^{-4}), MIGU at its default mask ratio 0.7, Adam-NSCL across k_{sh} and with its dimension set by GPM's energy criterion at 0.95, and ours at caps $C \in \{16, 64, 128, 256, 512\}$. Five operating points lie outside its range, each learning less than unconstrained fine-tuning or than ours by a wide margin: MIGU at mask ratio 0.9 (old-probe NLL falls by 0.027; 0.17 nats less learning than ours), Adam-NSCL at $k_{\text{sh}} = 585$ and at GPM energy 0.99 (forgetting 0.004 and −0.005; 0.054 and 0.046 nats less learning than unconstrained), EWC at $\lambda = 4\times10^6$ (−0.011; 0.021 nats less), and L2-to-init at $\lambda = 1$ (old-probe NLL falls by 0.021; 0.148 nats less learning than ours).

Implementations. *EWC* adds $\frac{\lambda}{2}\sum_j F_j(w_j - w_j^0)^2$ over the fc1 weights, the diagonal Fisher estimated on the 16 old-support windows ($\lambda = 0$ reproduces B1), and *L2-to-init* sets $F_j = 1$; *MIGU* masks, on each batch, the gradient of the fraction τ of units with the smallest mean $|z_i|$; *LoRA* trains $s\,BA$ with $s = \alpha/r$, $\alpha = 2r$, B initialised at zero and the bias trained, which O-LoRA equals at the first task; *LoRA-Null* initialises the adapter from the thin SVD of $W_0UU^\top$, U the r smallest-eigenvalue eigenvectors of $X_{\text{old}}^\top X_{\text{old}}$, on a frozen residual that makes the effective weight W_0 at initialisation.

Reading the table. EWC's forgetting falls with λ, from +0.27 at $\lambda \le 40$ to +0.094 at $\lambda = 4\times10^4$; up to there ours forgets less at $p \le 0.019$ and learns at least as well. At $\lambda = 4\times10^5$ EWC reaches our frontier (+0.0175, −0.6241; neither paired difference is significant against cap 128 or cap 256),

Table 7: **Method families on `fc1`.** OPT-1.3b, WikiText-103, seeds 0–2, forgetting and learning as in Appendix C. Δ is our cap 128 minus the row, paired by seed; negative means ours forgets less (one-sided p) or learns more (two-sided p).

family	setting	lr	forget	learn	Δ forget (p)	Δ learn (p)
none	—	10^{-4}	$+0.2684$	-0.5878	-0.2447 (0.0018)	-0.0332 (0.0436)
EWC	$\lambda = 4$	10^{-4}	$+0.2695$	-0.5850	-0.2457 (0.0007)	-0.0360 (0.0261)
EWC	$\lambda = 40$	10^{-4}	$+0.2743$	-0.5863	-0.2505 (0.0001)	-0.0347 (0.0118)
EWC	$\lambda = 400$	10^{-4}	$+0.2578$	-0.5863	-0.2341 (0.0004)	-0.0347 (0.0051)
EWC	$\lambda = 4 \times 10^3$	10^{-4}	$+0.2092$	-0.5962	-0.1854 ($< 10^{-4}$)	-0.0248 (0.0047)
EWC	$\lambda = 4 \times 10^4$	10^{-4}	$+0.0937$	-0.6144	-0.0700 (0.0190)	-0.0066 (0.3399)
EWC	$\lambda = 4 \times 10^5$	10^{-4}	$+0.0175$	-0.6241	$+0.0063$ (0.7618)	$+0.0032$ (0.2396)
EWC	$\lambda = 4 \times 10^6$	10^{-4}	-0.0114	-0.5668	$+0.0352$ (0.9866)	-0.0541 (0.0069)
MIGU	mask 0.7	10^{-4}	$+0.0486$	-0.5698	-0.0248 (0.0039)	-0.0512 (0.0005)
MIGU	mask 0.9	10^{-4}	-0.0267	-0.4484	$+0.0505$ (0.9896)	-0.1726 (0.0003)
LoRA	$r = 8$	2×10^{-4}	$+0.1360$	-0.5553	-0.1122 (0.0003)	-0.0657 (0.0080)
LoRA	$r = 8$	10^{-3}	$+0.3601$	-0.6567	-0.3363 (0.0007)	$+0.0357$ (0.0360)
LoRA	$r = 32$	2×10^{-4}	$+0.2558$	-0.6320	-0.2320 (0.0007)	$+0.0110$ (0.1448)
LoRA	$r = 32$	5×10^{-4}	$+0.3319$	-0.6611	-0.3081 (0.0004)	$+0.0401$ (0.0101)
LoRA	$r = 32$	10^{-3}	$+0.3875$	-0.6563	-0.3637 (0.0039)	$+0.0353$ (0.0015)
LoRA	$r = 128$	2×10^{-4}	$+0.2841$	-0.6626	-0.2604 (0.0002)	$+0.0416$ (0.0168)
LoRA	$r = 128$	10^{-3}	$+0.5328$	-0.5927	-0.5090 (0.0027)	-0.0283 (0.0351)
LoRA-Null	$r = 32$	2×10^{-4}	$+0.0853$	-0.5602	-0.0615 (0.0028)	-0.0608 (0.0026)
LoRA-Null	$r = 128$	2×10^{-4}	$+0.1221$	-0.6325	-0.0983 (0.0055)	$+0.0115$ (0.0551)
L2-to-init	$\lambda = 10^{-3}$	10^{-4}	$+0.2648$	-0.5944	-0.2410 (0.0003)	-0.0266 (0.0309)
L2-to-init	$\lambda = 10^{-2}$	10^{-4}	$+0.1893$	-0.6328	-0.1655 (0.0022)	$+0.0118$ (0.0425)
L2-to-init	$\lambda = 0.1$	10^{-4}	$+0.0558$	-0.6252	-0.0321 (0.0517)	$+0.0042$ (0.5019)
L2-to-init	$\lambda = 0.3$	10^{-4}	$+0.0103$	-0.5694	$+0.0134$ (0.9557)	-0.0516 (0.0169)
L2-to-init	$\lambda = 1$	10^{-4}	-0.0207	-0.4733	$+0.0445$ (0.9947)	-0.1477 (0.0005)
Adam-NSCL	$k_{\mathrm{sh}} = 34$	10^{-4}	$+0.0462$	-0.6229	-0.0224 (0.0075)	$+0.0019$ (0.2644)
Adam-NSCL	$k_{\mathrm{sh}} = 259$	10^{-4}	$+0.0255$	-0.5671	-0.0018 (0.2146)	-0.0539 (0.0002)
Adam-NSCL	GPM energy 0.95	10^{-4}	$+0.0227$	-0.6346	$+0.0011$ (0.5669)	$+0.0136$ (0.0217)
ours	$C = 64$	10^{-4}	$+0.0339$	-0.6237	-0.0101 (0.0181)	$+0.0027$ (0.1251)
ours	$C = 128$	10^{-4}	$+0.0238$	-0.6210	reference	
ours	$C = 256$	10^{-4}	$+0.0110$	-0.6189	$+0.0128$ (0.9731)	-0.0021 (0.1189)

Table 8: **WiSE-FT on each method.** OPT-1.3b, WikiText-103, seeds 0–2. *End of training* is $\alpha = 1$; *zero forgetting* is the learning each method's WiSE-FT curve reaches where the old-probe NLL returns to its base value, and its gain is paired by seed against WiSE-FT on unconstrained fine-tuning (negative = learns more; seeds on which it is negative).

		end of training		WiSE-FT, zero forgetting	
method	dims/row	forget	learn	learn	gain over none (seeds)
no constraint	—	$+0.2684$	-0.5878	-0.5904	—
ours, $C = 64$	23.5	$+0.0339$	-0.6237	-0.6289	-0.0385 (3/3)
ours, $C = 128$	34.4	$+0.0238$	-0.6210	-0.6262	-0.0359 (3/3)
Adam-NSCL, $k_{\mathrm{sh}} = 24$	24	$+0.0468$	-0.6252	-0.6215	-0.0311 (3/3)
Adam-NSCL, $k_{\mathrm{sh}} = 34$	34	$+0.0462$	-0.6229	-0.6178	-0.0275 (3/3)
Adam-NSCL, GPM energy 0.95	259.2	$+0.0227$	-0.6346	-0.6337	-0.0434 (3/3)

as expected of a criterion whose diagonal Fisher inherits the open/closed partition (Proposition 12). MIGU at $\tau = 0.7$ forgets $2\times$ more than ours and learns 0.051 nats less ($p = 0.0005$). Every LoRA configuration forgets 5.7–$22\times$ as much as ours; the four that learn significantly more do so by 0.035–0.042 nats, at 12–$16\times$ the forgetting, and LoRA at $r = 8$, lr 2×10^{-4} and at $r = 128$, lr 10^{-3} is dominated outright. LoRA-Null at lr 2×10^{-4} forgets $3.0\times$ less than LoRA at rank 32 (0.085 against 0.256) and $2.3\times$ less at rank 128 (0.122 against 0.284); against our cap 128 it forgets $3.6\times$ and $5.1\times$ more, is dominated at rank 32 (0.061 nats less learning, $p = 0.003$), and at rank 128 learns 0.012 nats more than ours ($p = 0.055$). L2-to-init at $\lambda = 0.3$ forgets the same as our cap 256 (paired difference $+0.0007$, TOST-equivalent at $\delta = 0.0068$) and learns 0.049 nats less ($3/3$ seeds, $p = 0.020$). Adam-NSCL with GPM's energy criterion at 0.95 (mean 259 dimensions per row) forgets as much as our cap 128 ($+0.0011$, $p = 0.57$) and learns 0.014 nats more; our cap 256,

at 47.3 dimensions per row, forgets 0.0117 less than it (3/3 seeds, one-sided $p = 0.03$) and learns 0.016 nats less ($p = 0.013$).

F.2 WiSE-FT AND ZERO-SHOT ACCURACY

Table 9: **Zero-shot accuracy after one fine-tuning transition.** OPT-1.3b fine-tuned on WikiText-103 (`fc1`, lr 10^{-4}), mean over seeds 0–2, in %. Accuracy for LAMBADA (OpenAI variant) and Winogrande, length-normalised accuracy for the others. *Mean drop* is the base model's mean accuracy over the five tasks minus the run's. Cap 128 against no constraint: $+1.92$ points of mean accuracy, positive on 3 of 3 seeds, one-sided paired $p = 0.0125$ ($n = 3$).

method	LAMBADA	HellaSwag	ARC-e	Winogrande	SciQ	mean drop
base	58.9	53.7	51.3	59.1	78.2	—
no constraint	53.7	51.1	48.8	58.6	74.4	2.94
Adam-NSCL, $k_{\mathrm{sh}} = 34$	56.1	52.6	51.3	58.5	77.3	1.08
ours, $C = 128$ ($\bar{k} = 34.4$)	57.3	52.6	50.8	58.1	77.4	1.02
Adam-NSCL, GPM energy 0.95 (259)	57.0	53.2	51.4	58.4	76.9	0.85

WiSE-FT (Wortsman et al., 2022) interpolates a finished run's `fc1` weights toward the base, $W_0 + \alpha(W - W_0)$. On the checkpoints of seeds 0–2 at $\alpha \in \{0.4, 0.5, 0.6, 0.7, 0.75, 0.8, 0.85, 0.9, 0.95, 1\}$ each curve is read at a fixed forgetting by linear interpolation between neighbouring α; α is swept on the evaluation sets, so this compares trade-off curves, not deployable choices of α. Every constrained run composes with it (Table 8), and the matched-budget ordering holds: at zero forgetting composed cap 64 learns 0.0074 nats more than composed $k_{\mathrm{sh}} = 24$ and composed cap 128 0.0084 more than composed $k_{\mathrm{sh}} = 34$, on 3/3 seeds each ($p = 0.08$), while Adam-NSCL with GPM's energy criterion, at 7.5–11× their dimensions, learns 0.005–0.008 nats more than either composed cap ($p = 0.19, 0.09$). Without interpolation, each projector learns more on the seed mean than WiSE-FT on the unconstrained run read at the same forgetting: by 0.006 and 0.008 nats at caps 64 and 128, by 0.001–0.003 for Adam-NSCL at the matched budgets and by 0.023 for its energy criterion (3 seeds, $p = 0.13$–0.88). Zero-shot accuracy (Table 9) is scored on the full evaluation splits of the five tasks.

G REPAIR AFTER PRUNING: THE FULL DEVELOPMENT

This appendix proves Lemma 13 and Theorem 15, whose body statements condense parts (b)–(c) and (a)–(f) of the full forms below, and derives the signatures they predict; Theorem 14 is proved in Appendix H, the experiments are in Appendix K, and Figure 7 reads the objectives of §5 one token at a time and on one pruned row of the toy layer of Figure 1.

G.1 NOTATION AND CURVATURE BLINDNESS

Notation for one row. Throughout this appendix and Appendix H we fix one unit of a wall matrix. Its dense weights are $w \in \mathbb{R}^d$, its bias is $b \in \mathbb{R}$ and its threshold is $\theta \geq 0$, so that $z_{\hat{w}}(x) = \langle \hat{w}, x \rangle + b$ for all $\hat{w}, x \in \mathbb{R}^d$. The mask $M \in \{0,1\}^d$ keeps the index set $K \subseteq \{1, \dots, d\}$, with $k = |K|$ and complement K^c; for $v \in \mathbb{R}^d$, the vectors $v_K \in \mathbb{R}^k$ and v_{K^c} are its restrictions to K and K^c. The repair moves only the kept weights, and the bias is held at b. A repaired row is a vector $\hat{w} \in \mathbb{R}^d$ with $\hat{w} = \hat{w} \odot M$, which we call *masked*. We parametrise it by $\delta := \hat{w}_K - w_K \in \mathbb{R}^k$, so that the mask-only weights $w \odot M$ are $\delta = 0$, and we write a function of a masked $\hat{w}$ as a function of δ where convenient. Because the bias is fixed, we write $\hat{w} \in \mathcal{P}_{\mathcal{X}}$ for $(\hat{w}, b) \in \mathcal{P}_{\mathcal{X}}$ and $\hat{w} \in \mathcal{K}_{\mathcal{X}}$ for $(\hat{w}, b) \in \mathcal{K}_{\mathcal{X}}$ (Definition 3); the cells $\mathcal{C}_>$ and $\mathcal{C}_\leq$ and the sets $\mathcal{P}_{\mathcal{X}}$ and $\mathcal{K}_{\mathcal{X}}$ are those of the dense unit (w, b) on the calibration tokens. These tokens are $x_1, \dots, x_n \in \mathbb{R}^d$, forming the multiset $\mathcal{X}$, and for every $\hat{w} \in \mathbb{R}^d$ we write

$$z_t = z_w(x_t), \qquad z_t^M = z_{w \odot M}(x_t), \qquad \hat{z}_t = z_{\hat{w}}(x_t),$$

so that $\hat{z}_t = z_t^M + \langle \delta, x_{t,K} \rangle$ for a masked $\hat{w}$, where $x_{t,K} \in \mathbb{R}^k$ is the restriction of x_t to K. Let $I_+ = \{t : z_t > \theta\}$, the tokens in $\mathcal{C}_>$, and $I_- = \{t : z_t \leq \theta\}$, with $n_\pm = |I_\pm|$. Let $X_K \in \mathbb{R}^{n \times k}$ have rows $x_{t,K}^\top$, and let $A \in \mathbb{R}^{n_+ \times k}$ and $\tilde{A} \in \mathbb{R}^{n_- \times k}$ consist of its rows indexed by I_+ and by I_-;

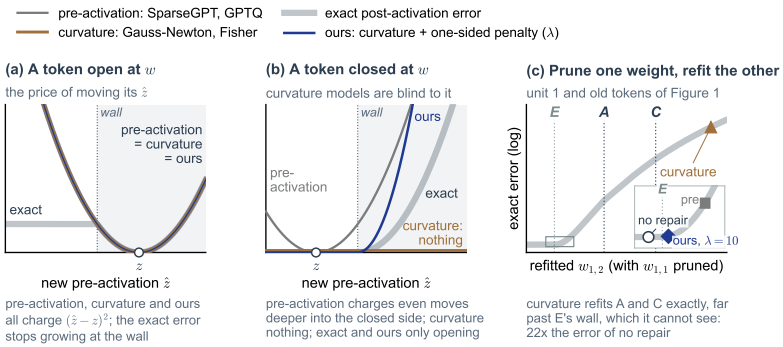


Figure 7: **What each repair objective charges for, on open and closed tokens.** Four objectives are compared, as in the legend: the pre-activation error $\sum_x (\hat{z}-z)^2$, which SparseGPT and GPTQ fit; the Gauss–Newton and Fisher models of the post-activation error, which reduce to the in-cell error on the open tokens, L_{gated} (Lemma 13); the exact post-activation error L_{post}; and ours, $L_{\text{gated_h}}$, which adds to that in-cell error a penalty of weight λ on closed tokens that open, the one-sided penalty of Theorem 15. **(a)** The price each objective charges one token that is open at w, here with $z = 1$, as its new pre-activation $\hat{z}$ moves; shading marks the open side of the wall. The pre-activation, curvature and our objectives charge the same parabola $(\hat{z} - z)^2$, drawn as three coincident curves. The exact error agrees with them while the token stays open, and stops growing once it closes. A move to $\hat{z} = -1$ costs 4 under the three quadratics but only 1 under the exact error, because the unit's output cannot fall below 0. **(b)** The same for a token closed at w, here with $z = -1$. The pre-activation objective still charges $(\hat{z} - z)^2$, so a move to $\hat{z} = -2$, deeper into the closed side and with no effect on the output, costs 1. The curvature models charge nothing anywhere, even after the token opens. The exact error charges only opening, $\max(\hat{z}, 0)^2$, and ours charges λ times that, with $\lambda = 3$ here. A move to $\hat{z} = 1$ thus costs 4 under pre-activation, 0 under curvature, 1 under the exact error and 3 under ours. **(c)** What each objective does to a pruned row. Take unit 1 of Figure 1, with $w = (1, 0.4)$, old tokens A and C open and E and D closed. Prune $w_{1,1}$ to 0 and let each objective refit $w_{1,2}$ alone. The horizontal axis is the refitted $w_{1,2}$, and the curve is the exact error along it (log scale). The dotted lines mark where E opens ($w_{1,2} = 0.50$) and where A and C open again (1.65 and 2.85). No point of this line keeps every gate: A and C need $w_{1,2} > 2.85$, while E needs $w_{1,2} \leq 0.50$. In this toy no refit therefore improves on leaving the pruned row alone at $w_{1,2} = 0.4$, where A, C and E are all closed and the exact error is 2.45; the objectives differ in how much they lose. The curvature fit sees only A and C and restores both exactly, by moving to $w_{1,2} = 4.13$. That opens E to $z_E = 7.26$, and the exact error there is 52.8, 22× that of no repair. The pre-activation fit, at 0.81, keeps A and C closed and opens E to 0.61, for an exact error of 2.83. Ours at $\lambda = 10$ stops at 0.54, just past E's wall with $z_E = 0.09$, and its exact error of 2.46 equals that of no repair to within 1%. The inset, on a linear scale, enlarges the three fits next to E's wall.

if $n_+ = 0$ or $n_- = 0$ the corresponding matrix is empty, with kernel $\mathbb{R}^k$. Put

$$\zeta := (z_t - z_t^M)_{t \in I_+} = \big(\langle w_{K^c}, x_{t,K^c} \rangle\big)_{t \in I_+} \in \mathbb{R}^{n_+}, \qquad c := (\theta - z_t^M)_{t \in I_-} \in \mathbb{R}^{n_-},$$

and let y_+ denote the componentwise positive part of a vector y. For a masked $\hat{w}$, $\hat{z}_t - z_t = (A\delta - \zeta)_t$ when $t \in I_+$ and $\hat{z}_t - \theta = (\tilde{A}\delta - c)_t$ when $t \in I_-$, so the objectives of §5 are

$$L_{\text{gated}} = \|A\delta - \zeta\|^2, \qquad \text{Esc} = \|(\tilde{A}\delta - c)_+\|^2,$$

and in particular $\|\zeta\|^2 = L_{\text{gated}}(w \odot M)$ and $E_0 = \text{Esc}(w \odot M) = \|(-c)_+\|^2$. The masked vectors that satisfy the inactive block of $\mathcal{P}_\mathcal{X}$ form $\mathcal{Q}_M = \{\delta : \tilde{A}\delta \leq c\}$, and $\mathcal{P}_\mathcal{X} \cap \{\hat{w} = \hat{w} \odot M\} = \{\delta : A\delta \geq \theta\mathbf{1} - z_{I_+}^M,\ \tilde{A}\delta \leq c\}$, so $\mathcal{Q}_M \supseteq \mathcal{P}_\mathcal{X} \cap \{\hat{w} = \hat{w} \odot M\}$. The escape is at most the closed→open term of Eq. (2), because both sum over the closed tokens with $\hat{z}_t > \theta$ and $(\hat{z}_t - \theta)^2 \leq \hat{z}_t^2$ there, with equality for every such token exactly when $\theta = 0$.

Lemma (Full form of Lemma 13)**.** *Let $\theta \geq 0$, let $(w, b) \in \mathbb{R}^d \times \mathbb{R}$ be one unit's dense weights and bias, and let $\mathcal{X} \subseteq \mathbb{R}^d$ be a finite multiset of n calibration tokens. Assume that no token lies on the wall, that is, $z_w(x) \neq \theta$ for every $x \in \mathcal{X}$. Put $m_\mathcal{X} := \min_{x \in \mathcal{X}} |z_w(x) - \theta|$, which is positive, and*

$R := \max_{x\in\mathcal{X}} \|x\|$, *with* $m_{\mathcal{X}}/R := +\infty$ *if* $R = 0$. *Throughout,* $\hat{w}$ *ranges over* $\mathbb{R}^d$ *with the bias held at b, and* $L_{\text{gated}}(\hat{w}) = \sum_{x\in\mathcal{X}\cap\mathcal{C}_>} (z_{\hat{w}}(x) - z_w(x))^2$.

(a) $L_{\text{post}} = L_{\text{gated}}$ *on* $\mathcal{K}_{\mathcal{X}}$.

(b) *Every* $\hat{w}$ *with* $\|\hat{w} - w\| < m_{\mathcal{X}}/R$ *lies in* $\mathcal{K}_{\mathcal{X}}$. *Hence* $L_{\text{post}} = L_{\text{gated}}$ *on the open ball* $B(w, m_{\mathcal{X}}/R)$, *and both crossing terms of Eq. (2) vanish on it.*

(c) L_{post} *and* L_{gated} *have the same derivatives of every order at* w; *in particular* $L_{\text{post}}(w) = 0$, $\nabla L_{\text{post}}(w) = 0$ *and* $\nabla^2 L_{\text{post}}(w) = 2\sum_{x\in\mathcal{X}\cap\mathcal{C}_>} xx^\top$. *The Gauss–Newton matrix of* L_{post} *at* w *is* $J^\top J = \sum_{x\in\mathcal{X}\cap\mathcal{C}_>} xx^\top = \frac{1}{2}\nabla^2 L_{\text{post}}(w)$, *where* J *has one row* $J_x = \mathbb{1}[x \in \mathcal{C}_>]\, x^\top$ *per token, and for every* $\sigma > 0$ *the Fisher matrix at* $\hat{w} = w$ *of the Gaussian model* $y \sim \mathcal{N}(\text{act}_\theta(z_{\hat{w}}(x)), \sigma^2)$, *summed over the tokens, is* $\sigma^{-2} J^\top J$. *K-FAC (Martens & Grosse, 2015) replaces* $J^\top J$ *by* $\big(\frac{1}{n}\sum_{x\in\mathcal{X}} \mathbb{1}[x \in \mathcal{C}_>]\big) \sum_{x\in\mathcal{X}} xx^\top$, *which approximates* $J^\top J$ *and differs from it in general.*

Proof. (a) Let $\hat{w} \in \mathcal{K}_{\mathcal{X}}$. Every token then keeps its dense gate bit, so the sets S_{+-} and S_{-+} of Theorem 5 are empty and $S_{++} = \mathcal{X} \cap \mathcal{C}_>$; Eq. (2) reduces to its in-cell term, which is $L_{\text{gated}}(\hat{w})$.

(b) Let $\|\hat{w} - w\| < m_{\mathcal{X}}/R$ and $x \in \mathcal{X}$. By the Cauchy–Schwarz inequality, $|z_{\hat{w}}(x) - z_w(x)| = |\langle \hat{w} - w, x\rangle| \le \|\hat{w} - w\|\,\|x\|$. If $x = 0$ this is 0; otherwise $0 < \|x\| \le R$, and it is less than $(m_{\mathcal{X}}/R)\, R = m_{\mathcal{X}}$. In both cases $|z_{\hat{w}}(x) - z_w(x)| < m_{\mathcal{X}} \le |z_w(x) - \theta|$, so $z_{\hat{w}}(x) - \theta$ is non-zero and has the sign of $z_w(x) - \theta$: the token keeps its gate bit. Hence $\hat{w} \in \mathcal{K}_{\mathcal{X}}$, part (a) gives $L_{\text{post}}(\hat{w}) = L_{\text{gated}}(\hat{w})$, and the crossing sets, hence both crossing terms, are empty.

(c) Two functions that agree on an open set containing w have the same derivatives of every order at w, because a derivative at a point depends only on the values on a neighbourhood of that point. By (b) this applies to L_{post} and L_{gated}. The function $L_{\text{gated}}(\hat{w}) = \sum_{x\in\mathcal{X}\cap\mathcal{C}_>} \langle \hat{w} - w, x\rangle^2$ is a quadratic with value 0, gradient 0 and Hessian $2\sum_{x\in\mathcal{X}\cap\mathcal{C}_>} xx^\top$ at w, which gives the three stated derivatives of L_{post}. For the Gauss–Newton matrix, write $L_{\text{post}}(\hat{w}) = \sum_x \rho_x(\hat{w})^2$ with residuals $\rho_x(\hat{w}) = \text{act}_\theta(z_{\hat{w}}(x)) - \text{act}_\theta(z_w(x))$. The map act_θ is differentiable at every $u \neq \theta$, with derivative $\mathbb{1}[u > \theta]$, and $z_w(x) \neq \theta$ by hypothesis, so ρ_x is differentiable at w with gradient $\mathbb{1}[x \in \mathcal{C}_>]\, x$. These gradients are the rows J_x, and $J^\top J = \sum_x \mathbb{1}[x \in \mathcal{C}_>]\, xx^\top = \frac{1}{2}\nabla^2 L_{\text{post}}(w)$. For the Gaussian model, the score of one observation is $\sigma^{-2}\big(y - \text{act}_\theta(z_{\hat{w}}(x))\big)\nabla_{\hat{w}} \text{act}_\theta(z_{\hat{w}}(x))$; at $\hat{w} = w$ its second moment under the model is $\sigma^{-4} \cdot \sigma^2 \cdot J_x^\top J_x = \sigma^{-2} J_x^\top J_x$, and summing over the tokens gives $\sigma^{-2} J^\top J$. Finally, the K-FAC matrix differs from $J^\top J$ in general. Take $d = 2$, $w = e_1$, $b = 0$, $\theta = 0$ and the two tokens e_1 and $e_2 - e_1$, whose pre-activations are 1 and -1. Then $J^\top J = e_1 e_1^\top$ has $(2,2)$ entry 0, while the K-FAC matrix $\frac{1}{2}\big(e_1 e_1^\top + (e_2 - e_1)(e_2 - e_1)^\top\big)$ has $(2,2)$ entry $\frac{1}{2}$. □

G.2 Pricing the faces back

Theorem (Full form of Theorem 15)**.** *Use the notation for one row of §G.1, let* $\lambda > 0$ *be real, and write*

$$F_\lambda(\delta) := L_{\text{gated_h}}(\hat{w}, \lambda) = \|A\delta - \zeta\|^2 + \lambda\,\|(\tilde{A}\delta - c)_+\|^2, \qquad \delta \in \mathbb{R}^k.$$

(a) F_λ *is convex and continuously differentiable, with Lipschitz gradient.*

(b) *Its penalty is the sum of the squared violations of the inactive constraints. It is zero exactly on* $\mathcal{Q}_M$, *which contains* $\mathcal{P}_{\mathcal{X}} \cap \{\hat{w} = \hat{w} \odot M\}$, *so it is an exterior penalty for the inactive block, finite everywhere.*

(c) F_λ *attains its minimum, and* $A\delta$ *and* $(\tilde{A}\delta - c)_+$ *take the same value at every minimiser; in particular* L_{gated} *and* Esc *take single values* $g(\lambda)$ *and* $e(\lambda)$ *on the minimiser set.*

(d) *Every* δ *with* $F_\lambda(\delta) \le F_\lambda(0)$ *(every minimiser, and the output of any method whose output has objective at most that of its starting point* $w \odot M$*) satisfies*

$$\sum_{x\in\mathcal{X}\cap\mathcal{C}_\le} \max\big(z_{\hat{w}}(x) - \theta, 0\big)^2 \;\le\; \frac{\|\zeta\|^2}{\lambda} + E_0, \qquad E_0 := \sum_{x\in\mathcal{X}\cap\mathcal{C}_\le} \max\big(z_{w\odot M}(x) - \theta, 0\big)^2. \quad (7)$$

If $\mathcal{Q}_M \neq \varnothing$, then $G^\star := \min_{\delta\in\mathcal{Q}_M} L_{\mathrm{gated}}(\delta)$ is attained, and every δ with $F_\lambda(\delta) \le G^\star$, every minimiser in particular, satisfies

$$\sum_{x\in\mathcal{X}\cap\mathcal{C}_\le} \max\big(z_{\hat w}(x)-\theta,0\big)^2 \;\le\; \frac{G^\star}{\lambda} \;\xrightarrow[\lambda\to\infty]{}\; 0. \tag{8}$$

***(e)** On $(0,\infty)$, $\lambda\mapsto e(\lambda)$ is non-increasing and $\lambda\mapsto g(\lambda)$ is non-decreasing.*

***(f)** $e(\lambda)\to 0$ as $\lambda\to\infty$ if and only if $\mathcal{Q}_M\neq\varnothing$. If $\mathcal{Q}_M=\varnothing$, then $\min_\delta \mathrm{Esc}(\delta)$ is attained and positive, and $e(\lambda)\ge\min_\delta \mathrm{Esc}(\delta)$ for every $\lambda>0$.*

For $\theta=0$ the escape is exactly the closed→open term of Eq. (2), and E_0 is that term at the mask-only weights.

The proof of (e) uses one elementary lemma.

Lemma 22 (Exchange)**.** *Let $\mathcal{D}$ be a set, let $g,P:\mathcal{D}\to\mathbb{R}$ be functions, and let $\lambda_1<\lambda_2$ be real numbers. If $x_i\in\mathcal{D}$ minimises $g+\lambda_i P$ over $\mathcal{D}$ for $i=1,2$, then $P(x_2)\le P(x_1)$. If in addition $\lambda_1\ge 0$, then $g(x_2)\ge g(x_1)$.*

Proof. Optimality of x_1 at λ_1 and of x_2 at λ_2 gives $g(x_1)+\lambda_1P(x_1)\le g(x_2)+\lambda_1P(x_2)$ and $g(x_2)+\lambda_2P(x_2)\le g(x_1)+\lambda_2P(x_1)$. Adding the two inequalities gives $(\lambda_2-\lambda_1)\big(P(x_2)-P(x_1)\big)\le 0$, and since $\lambda_2-\lambda_1>0$, $P(x_2)\le P(x_1)$. The first inequality can be rewritten as $g(x_1)-g(x_2)\le\lambda_1\big(P(x_2)-P(x_1)\big)$. When $\lambda_1\ge 0$ the right-hand side is the product of a non-negative and a non-positive number, hence at most 0, so $g(x_2)\ge g(x_1)$. □

Proof of Theorem 15. Throughout, $\psi(u):=u_+^2$ for $u\in\mathbb{R}$, so that the penalty is $\sum_{j\in I_-}\psi\big((\tilde A\delta-c)_j\big)$.

(a) The function ψ is convex and non-decreasing, being the composition of the convex map $u\mapsto u_+$ with the map $s\mapsto s^2$, which is convex and non-decreasing on $[0,\infty)$. It is continuously differentiable with $\psi'(u)=2u_+$, and $|\psi'(u)-\psi'(u')|\le 2|u-u'|$, so ψ' is 2-Lipschitz. Each penalty summand is ψ composed with an affine function of δ, hence convex and continuously differentiable with Lipschitz gradient, and $\|A\delta-\zeta\|^2$ is a convex quadratic whose gradient $2A^\top(A\delta-\zeta)$ is Lipschitz. A non-negative combination of such functions, here with weights 1 and $\lambda>0$, has the same three properties.

(b) Since $\psi\ge 0$ and $\psi(u)=0$ if and only if $u\le 0$, the penalty vanishes if and only if $(\tilde A\delta-c)_j\le 0$ for every j, that is, on $\mathcal{Q}_M$; elsewhere it is the sum of the squared positive violations. The containment $\mathcal{Q}_M\supseteq\mathcal{P}_\mathcal{X}\cap\{\hat w=\hat w\odot M\}$ holds because $\mathcal{Q}_M$ imposes only the inactive block of $\mathcal{P}_\mathcal{X}$ (§G.1).

(c), existence. Consider the quadratic programme

$$\min\big\{\|A\delta-\zeta\|^2+\lambda\|s\|^2:\ (\delta,s)\in\mathbb{R}^k\times\mathbb{R}^{n_-},\ s\ge\tilde A\delta-c,\ s\ge 0\big\}.$$

Its feasible set is a polyhedron, and it is non-empty because it contains $(0,(-c)_+)$; its objective is a quadratic bounded below by 0 on it. By the Frank–Wolfe theorem (Frank & Wolfe, 1956), a quadratic function bounded below on a non-empty polyhedron attains its infimum there, so the programme has a minimiser $(\bar\delta,\bar s)$. For fixed δ the constraints on s are the componentwise bounds $s_j\ge\max((\tilde A\delta-c)_j,0)$, and s_j^2 is increasing on $[0,\infty)$, so the minimum over s is attained at $s=(\tilde A\delta-c)_+$ and equals $F_\lambda(\delta)$. Hence, for every δ, $F_\lambda(\delta)\ge\|A\bar\delta-\zeta\|^2+\lambda\|\bar s\|^2\ge F_\lambda(\bar\delta)$, where the second inequality holds because $\bar s$ is feasible for $\bar\delta$. So $\bar\delta$ minimises F_λ.

(c), invariance. Let δ_1,δ_2 be minimisers, let $F^\star$ be the minimum and let $\delta_m=\frac12(\delta_1+\delta_2)$. Put $y_i=A\delta_i-\zeta$ and $u_i=\tilde A\delta_i-c$ for $i=1,2$. The parallelogram identity gives $\|\frac12(y_1+y_2)\|^2=\frac12\|y_1\|^2+\frac12\|y_2\|^2-\frac14\|y_1-y_2\|^2$, and for each j the midpoint gap $\nu_j:=\frac12\psi(u_{1j})+\frac12\psi(u_{2j})-\psi\big(\frac12(u_{1j}+u_{2j})\big)$ is non-negative because ψ is convex. Adding these,

$$F^\star\le F_\lambda(\delta_m)=F^\star-\tfrac14\|y_1-y_2\|^2-\lambda\sum_j\nu_j.$$

Since $\lambda > 0$ and every subtracted term is non-negative, $A\delta_1 = A\delta_2$ and $\nu_j = 0$ for every j. We show that $\nu_j = 0$ forces $(u_{1j})_+ = (u_{2j})_+$. Write $a = \min(u_{1j}, u_{2j})$ and $a' = \max(u_{1j}, u_{2j})$, and consider three exhaustive cases. If $a' \le 0$, both positive parts are 0. If $a \ge 0$, then $\psi(u) = u^2$ on $[a, a']$ and $\nu_j = \frac{1}{4}(a' - a)^2$, so $a = a'$. If $a < 0 < a'$, the midpoint $\bar{u} = \frac{1}{2}(a + a')$ satisfies $\bar{u} < \frac{1}{2}a'$, so $\psi(\bar{u}) \le \frac{1}{4}a'^2$ and $\nu_j \ge \frac{1}{2}a'^2 - \frac{1}{4}a'^2 > 0$, which is excluded. Hence $(\tilde{A}\delta_1 - c)_+ = (\tilde{A}\delta_2 - c)_+$, and L_{gated} and Esc, which are functions of $A\delta$ and of $(\tilde{A}\delta - c)_+$, take one value each on the minimiser set.

(d) At $\delta = 0$, $F_\lambda(0) = \|\zeta\|^2 + \lambda E_0$ (§G.1). Let $F_\lambda(\delta) \le F_\lambda(0)$. Since $L_{\text{gated}}(\delta) \ge 0$, $\lambda\,\text{Esc}(\delta) \le F_\lambda(\delta) \le \|\zeta\|^2 + \lambda E_0$, and dividing by $\lambda > 0$ gives (7). Now let $\mathcal{Q}_M \neq \varnothing$. Then $\mathcal{Q}_M$ is a non-empty polyhedron on which the quadratic L_{gated} is bounded below by 0, so by the Frank–Wolfe theorem $G^\star$ is attained, at some $\delta_Q \in \mathcal{Q}_M$; by (b) the penalty vanishes there, so $F_\lambda(\delta_Q) = G^\star$. Let $F_\lambda(\delta) \le G^\star$. Then $\lambda\,\text{Esc}(\delta) \le G^\star - L_{\text{gated}}(\delta) \le G^\star$, which is (8), and a minimiser δ_λ qualifies because $F_\lambda(\delta_\lambda) \le F_\lambda(\delta_Q) = G^\star$.

(e) Let $0 < \lambda_1 < \lambda_2$. By (c) both F_{λ_1} and F_{λ_2} have minimisers, and Lemma 22 with $\mathcal{D} = \mathbb{R}^k$, $g = L_{\text{gated}}$ and $P = \text{Esc}$ gives $e(\lambda_2) \le e(\lambda_1)$ and $g(\lambda_2) \ge g(\lambda_1)$; by (c) these values do not depend on which minimisers are chosen.

(f) If $\mathcal{Q}_M \neq \varnothing$, (d) gives $0 \le e(\lambda) \le G^\star/\lambda \to 0$. Conversely, let $\mathcal{Q}_M = \varnothing$. The programme $\min\{\|s\|^2 : (\delta, s) \in \mathbb{R}^k \times \mathbb{R}^{n_-},\ s \ge \tilde{A}\delta - c,\ s \ge 0\}$ attains its minimum by the Frank–Wolfe theorem, as in (c), and by the same partial minimisation over s that minimum equals $\min_\delta \text{Esc}(\delta)$, which is therefore attained. It is positive: a minimiser δ' with $\text{Esc}(\delta') = 0$ would satisfy $\tilde{A}\delta' \le c$, that is, $\delta' \in \mathcal{Q}_M = \varnothing$. Hence $e(\lambda) \ge \min_\delta \text{Esc} > 0$ for every $\lambda > 0$, and $e(\lambda)$ does not tend to 0.

Finally, for $\theta = 0$ the escape $\sum_{t \in I_-} (\hat{z}_t)_+^2$ is the sum of $\hat{z}_t^2$ over the closed tokens with $\hat{z}_t > 0$, which is the closed→open term, and at $\delta = 0$ it is E_0. □

Remark 23 (Certifying the feasibility hypothesis). The bound that tends to zero is (8), and its hypothesis is cheap to certify: $\mathcal{Q}_M \neq \varnothing$ is the feasibility of the linear system $\tilde{A}\delta \le c$, and the vector $\hat{w} = 0$ is a witness whenever the unit's bias satisfies $b \le \theta$, since every token then has $z_{\hat{w}}(x) = b \le \theta$. At $\theta = 0$ this holds for 81.2% of OPT-1.3b and 95.7% of OPT-125m wall units (on dense-model activations at 32×256 calibration sequences with 8 192 reconstruction tokens). This witness certifies membership of $\mathcal{Q}_M$, which asks only the inactive block; when both blocks are non-empty, $\hat{w} = 0$ lies in $\mathcal{P}_\mathcal{X}$ itself only when $b = \theta$, which is why the theorem is stated over $\mathcal{Q}_M$. By Theorem 15(f), a masked weight vector respecting the closed side is exactly what confinement requires.

Remark 24 (The $\theta > 0$ case, and why the penalty is written with $\max(u - \theta, 0)$). Writing the penalty as $\text{act}_\theta(z_{\hat{w}})^2$ instead, the form our implementation uses, gives $u^2 \mathbb{1}[u > \theta]$, which for $\theta > 0$ jumps from 0 to θ^2 at $u = \theta$ and is neither continuous nor convex. For $\theta = 0$ the two forms coincide, so every ReLU result optimises the analysed objective. For $\theta > 0$ the implemented penalty $P(\hat{w}) = \sum_{\mathcal{X} \cap \mathcal{C}_\le} \text{act}_\theta(z_{\hat{w}})^2$ is exactly the closed→open term of Eq. (2), it dominates the escape pointwise and it is zero on $\mathcal{Q}_M$. The argument of Theorem 15(d) uses only these facts and non-negativity, so for the implemented objective $L_{\text{gated}} + \lambda P$: every $\hat{w}$ with objective at most that of $w \odot M$ has closed→open mass at most $\|\zeta\|^2/\lambda + P(w \odot M)$, and, if $\mathcal{Q}_M \neq \varnothing$, every $\hat{w}$ with objective at most $G^\star$ has closed→open mass at most $G^\star/\lambda$. These statements use $\text{act}_\theta(u) = u\mathbb{1}[u > \theta]$ of Definition 1. Under the implementation's convention $u \ge \theta$, the penalty at a point of $\mathcal{Q}_M$ equals θ^2 times the number of inactive tokens on the wall, so it vanishes on $\{\tilde{A}\delta < c\}$; when that set is non-empty it is dense in $\mathcal{Q}_M$, so for every $\epsilon > 0$ some $\hat{w}$ has objective at most $G^\star + \epsilon$, and every such $\hat{w}$ has closed→open mass at most $(G^\star + \epsilon)/\lambda$.

The objectives $L_{\text{post_w}}(\lambda)$ price the exact objective in the same way, for the model's activation act: $L_{\text{post_w}}(\lambda)(\hat{w}) = \sum_{t \in I_+} (\text{act}(\hat{z}_t) - \text{act}(z_t))^2 + \lambda \sum_{t \in I_-} (\text{act}(\hat{z}_t) - \text{act}(z_t))^2$, so that $L_{\text{post_w}}(1)$ is the exact post-activation error; for $\text{act} = \text{act}_\theta$ it is $\sum_{t \in I_+} (\text{act}_\theta(\hat{z}_t) - z_t)^2 + \lambda \sum_{t \in I_-} \text{act}_\theta(\hat{z}_t)^2$ and $L_{\text{post_w}}(1) = L_{\text{post}}$. On a SwiGLU block the suffix v ($L_{\text{gated_hv}}$, $L_{\text{post_wv}}$) weights every token term by v_t^2, where v_t is the unit's up-projection output on token t, which the repair does not change.

G.3 Predicted activation-pattern signatures

With the mask fixed and only the functional varied, Theorem 14 leaves the inactive faces unpriced by L_{gated}, which should therefore show a closed→open share far above the other objectives, while Theorem 15 and Remark 24 bound that mass for $L_{\text{gated_h}}$, and every $\hat{w}$ with $L_{\text{post}}(\hat{w}) \le L_{\text{post}}(w \odot M)$ has closed→open mass at most $L_{\text{post}}(w \odot M)$, since each term of Eq. (2) is non-negative. Measured against the dense pattern (Table 13, OPT-1.3b, $s = 0.7$), L_{gated}'s crossings are closed→open by 1.15% to 0.18% of pairs, and L_{pre}'s run the other way, open→closed by 0.52% to 0.34%; the closed→open term holds 0.372 of L_{gated}'s residual, against 0.171 for $L_{\text{gated_h}}(1)$, 0.111 for L_{post} and 0.049 for $L_{\text{gated_h}}(30)$.

H Proof of Theorem 14

We use the notation for one row of §G.1: the bias is fixed at b, $\delta = \hat{w}_K - w_K \in \mathbb{R}^k$, $L_{\text{gated}} = \|A\delta - \zeta\|^2$ and $\text{Esc} = \|(\tilde{A}\delta - c)_+\|^2$ with $c = (\theta - z_t^M)_{t \in I_-}$. For a matrix B, B^+ denotes its Moore–Penrose pseudo-inverse.

Theorem (Full form of Theorem 14)**.** *Let $\theta \ge 0$, and let one unit, its mask and its calibration tokens be as in §G.1.*

(i) *The objective $L_{\text{gated}}(\delta) = \|A\delta - \zeta\|^2$ depends on the calibration tokens only through A and ζ. Changing the inactive rows $\tilde{A}$ or the inactive pre-activations $z_{I_-}^M$ changes neither L_{gated} nor its minimiser set, which is $A^+\zeta + \ker A$.*

(ii) *Assume $\ker A \not\subseteq \ker \tilde{A}$. This holds if $\operatorname{rank} X_K = k$ and $\operatorname{rank} A < k$, and $\operatorname{rank} A < k$ holds whenever $n_+ < k$. Then there is $v \in \ker A$ such that, for every minimiser δ of L_{gated} and every $\tau \ge 0$, the point $\delta + \tau v$ is a minimiser, and $\text{Esc}(\delta + \tau v) \to \infty$ as $\tau \to \infty$.*

(iii) *Let the calibration tokens $x_1, \dots, x_n$ be i.i.d. draws from a distribution D on $\mathbb{R}^d$ with $\mathbb{E}\|x\|^2 < \infty$, let the mask be fixed, and assume $\Sigma_+ := \mathbb{E}\big[\mathbb{1}[z_w(x) > \theta]\, x_K x_K^\top\big] \succ 0$. Put $h := \mathbb{E}\big[\mathbb{1}[z_w(x) > \theta]\, x_K \langle w_{K^c}, x_{K^c}\rangle\big] \in \mathbb{R}^k$ and $\delta^\star := \Sigma_+^{-1} h$, let $w^\star$ be the masked vector with $w_K^\star = w_K + \delta^\star$, and put*

$$e^\star := \mathbb{E}\Big[\mathbb{1}[z_w(x) \le \theta]\, \big(z_{w^\star}(x) - \theta\big)_+^2\Big].$$

Then, almost surely, A has full column rank for all large n, so that the minimiser $\delta_n = (A^\top A)^{-1} A^\top \zeta$ of L_{gated} is unique; $\delta_n \to \delta^\star$; and $\frac{1}{n}\text{Esc}(\delta_n) \to e^\star$. The point $w^\star$ depends on D only through its restriction to $\mathcal{C}_>$, while $e^\star$ depends on its restriction to $\mathcal{C}_\le$ as well. Finally, $e^\star \ge 0$, and $e^\star > 0$ if and only if $D\big(z_w(x) \le \theta < z_{w^\star}(x)\big) > 0$.

The body's affine subspace in (ii) is the minimiser set of (i), and in body (iii) "a point $w^\star$ the active tokens determine" means that $w^\star$ depends on D only through its restriction to $\mathcal{C}_>$.

Proof. (i) The formula $L_{\text{gated}}(\delta) = \|A\delta - \zeta\|^2$ involves A and ζ only. Its minimisers are the solutions of the normal equations $A^\top A\delta = A^\top\zeta$. The vector $A^+\zeta$ is one of them, because $A^\top A A^+ = A^\top$, and two solutions differ by an element of $\ker(A^\top A) = \ker A$, since $A^\top A u = 0$ implies $\|Au\|^2 = u^\top A^\top A u = 0$; so the minimiser set is $A^+\zeta + \ker A$. Neither $\tilde{A}$ nor $z_{I_-}^M$ occurs in any of this: they enter only the inactive constraints $(\tilde{A}\delta)_j \le \theta - z_j^M$, $j \in I_-$. Finally $\ker A \cap \ker \tilde{A} = \ker X_K$, because the rows of X_K are those of A together with those of $\tilde{A}$.

(ii) Since $\ker A \not\subseteq \ker \tilde{A}$, there is $v \in \ker A$ with $\tilde{A}v \ne 0$, and so an index $j \in I_-$ with $(\tilde{A}v)_j \ne 0$; replacing v by $-v \in \ker A$ if necessary, $(\tilde{A}v)_j > 0$. For the sufficient condition: if $\operatorname{rank} X_K = k$ and $\operatorname{rank} A < k$, then $\ker A$ contains some $u \ne 0$, and $u \notin \ker \tilde{A}$ because $\ker A \cap \ker \tilde{A} = \ker X_K = \{0\}$; and $\operatorname{rank} A \le n_+$, so $n_+ < k$ gives $\operatorname{rank} A < k$. Now let δ be a minimiser and $\tau \ge 0$. Since $A(\delta + \tau v) = A\delta$, the point $\delta + \tau v$ has the minimal value of L_{gated}, so it is a minimiser. Token j's pre-activation along the ray is $\hat{z}_j(\tau) = z_j^M + (\tilde{A}\delta)_j + \tau(\tilde{A}v)_j$, affine and strictly increasing in τ. Hence it exceeds θ for every τ beyond a finite threshold, where the ray leaves $\mathcal{Q}_M$, and $\mathcal{P}_\mathcal{X}$, through the j-th inactive face; and

$$\text{Esc}(\delta + \tau v) \;\ge\; \big(\hat{z}_j(\tau) - \theta\big)_+^2 \;\xrightarrow[\tau\to\infty]{}\; \infty.$$

(iii), the minimiser. By the strong law of large numbers, $\frac{1}{n}A^\top A = \frac{1}{n}\sum_t \mathbb{1}[z_t > \theta]\, x_{t,K} x_{t,K}^\top \to \Sigma_+$ and $\frac{1}{n}A^\top \zeta = \frac{1}{n}\sum_t \mathbb{1}[z_t > \theta]\, x_{t,K}\langle w_{K^c}, x_{t,K^c}\rangle \to h$ almost surely; both summands are integrable, since their norms are at most $\|x\|^2$ and $\|w\|\,\|x\|^2$ and $\mathbb{E}\|x\|^2 < \infty$. The smallest eigenvalue is a continuous function of a symmetric matrix and $\Sigma_+ \succ 0$, so $\frac{1}{n}A^\top A \succ 0$ for all large n. Then A has full column rank, L_{gated} is strictly convex, and its unique minimiser is $\delta_n = (\frac{1}{n}A^\top A)^{-1}\frac{1}{n}A^\top \zeta$, which converges to $\delta^\star$ because inversion is continuous on invertible matrices. The population objective $\mathbb{E}\big[\mathbb{1}[z_w(x) > \theta]\,(\langle \delta, x_K\rangle - \langle w_{K^c}, x_{K^c}\rangle)^2\big]$ is a convex quadratic in δ with Hessian $2\Sigma_+ \succ 0$ and unique minimiser $\delta^\star$; its indicator is evaluated at the dense pre-activation and does not depend on δ. Both Σ_+ and h are expectations of functions that vanish off $\mathcal{C}_>$, so $w^\star$ depends on D only through its restriction to $\mathcal{C}_>$; the indicator $\mathbb{1}[z_w(x) \le \theta]$ in $e^\star$ brings in the restriction to $\mathcal{C}_\le$.

(iii), the escape. For $x \in \mathbb{R}^d$ and $\delta \in \mathbb{R}^k$ let $f(x,\delta) = \mathbb{1}[z_w(x) \le \theta]\,\big(z_{w\odot M}(x) + \langle \delta, x_K\rangle - \theta\big)_+^2$, so that $\mathrm{Esc}(\delta) = \sum_t f(x_t, \delta)$. For reals a and a', $|a_+^2 - (a')_+^2| \le |a - a'|\,(|a| + |a'|)$, and $|z_{w\odot M}(x) - \theta| \le |b - \theta| + \|w\|\,\|x\|$. Hence, for $\delta, \delta' \in \mathbb{R}^k$,

$$|f(x,\delta) - f(x,\delta')| \le \|\delta - \delta'\|\,\|x\|\,\big(2|b-\theta| + (2\|w\| + \|\delta\| + \|\delta'\|)\|x\|\big) \le C(\delta,\delta')\,\|\delta - \delta'\|\,(1 + \|x\|^2),$$

with $C(\delta,\delta') = 2|b-\theta| + 2\|w\| + \|\delta\| + \|\delta'\|$, using $\|x\| \le 1 + \|x\|^2$ and $\|x\|^2 \le 1 + \|x\|^2$. Since $\delta_n \to \delta^\star$ almost surely, the sequence δ_n is bounded, so $C(\delta_n, \delta^\star) \le \bar{C}$ for a finite $\bar{C}$ and all large n, and

$$\Big|\frac{1}{n}\sum_t f(x_t, \delta_n) - \frac{1}{n}\sum_t f(x_t, \delta^\star)\Big| \le \bar{C}\,\|\delta_n - \delta^\star\|\,\frac{1}{n}\sum_t (1 + \|x_t\|^2) \;\longrightarrow\; 0$$

almost surely, because the average converges to $1 + \mathbb{E}\|x\|^2 < \infty$. The bound holds for every realisation of the data, so the dependence of δ_n on the same tokens does not matter. Finally $\frac{1}{n}\sum_t f(x_t, \delta^\star) \to \mathbb{E} f(x, \delta^\star) = e^\star$ by the strong law, since $0 \le f(x,\delta^\star) \le \big(|b-\theta| + (\|w\| + \|\delta^\star\|)\|x\|\big)^2$ is integrable.

(iii), positivity. The integrand of $e^\star$ is non-negative, so $e^\star \ge 0$, and the positivity clause is Lemma 25 below with $w^\circ = w^\star$. □

Lemma 25 (When the limit escape is positive)**.** *In the setting of (iii), let $w^\circ \in \mathbb{R}^d$ be any masked weight vector and $e(w^\circ) := \mathbb{E}\big[\mathbb{1}[z_w(x) \le \theta]\,(z_{w^\circ}(x) - \theta)_+^2\big]$, so that $e^\star = e(w^\star)$. Then $e(w^\circ) > 0$ if and only if $D\big(z_w(x) \le \theta < z_{w^\circ}(x)\big) > 0$.*

Proof. The integrand of $e(w^\circ)$ is non-negative and is positive exactly on the set $\{z_w \le \theta < z_{w^\circ}\}$; an expectation of a non-negative function is positive if and only if the function is positive on a set of positive measure, which gives the equivalence. □

The regime of our measurements. At OPT-1.3b, $s = 0.7$, the closed-form solver of Appendix K fits $57\,344$ tokens; a row of gate density 0.035 has $n_+ \approx 2\,007$ active tokens against $k \approx 614$ kept columns, yet 39.2% of rows have $n_+ < k$, so both (ii) and (iii) occur.

I Experimental setup and noise floors

Activation and masks. The activation is resolved from the model config and verified against the loaded `act_fn` on 8192 random inputs, and the LLaMA-family models require a BOS token (ProSparse scores 6.07 nats on our wiki probe without it and 3.54 with it). The mask is a Wanda-style saliency, $\mathrm{sal}_{ij} = |W_{ij}| \cdot \|X_j\|_2$, ranked within each output row and thresholded at the target sparsity. The $\|X_j\|_2$ statistics are accumulated in a single forward pass over the calibration set on the *dense* model, and the mask is frozen and shared by every objective in a cell. All `Linear` modules inside transformer blocks are masked, and a 2:4 mask keeps the top 2 of every 4 consecutive input channels.

Repair. Repair is applied to the wall matrices alone (`fc1` on OPT, `dense_h_to_4h` on GPT-NeoX, `gate_proj` on LLaMA-family blocks), the matrices whose output passes through the activation; attention and the output matrix are pruned and left mask-only for every objective. Every objective is solved identically: Adam, learning rate 3×10^{-4}, batch 1024 reconstruction tokens,

Table 10: Models. The *wall matrix* is the one whose output passes through the activation and which the repair modifies; θ is recovered at load time by bisection on the resolved `act_fn`; the density column is the mask-only gate density at sparsity s (Pythia-160m has none). Activations: ReLU, GELU (Hendrycks & Gimpel, 2016) and the thresholded FATReLU (Kurtz et al., 2020); SwiGLU blocks (Shazeer, 2020).

model	activation	block	θ	n_{wall}	s	mask-only density
OPT-125m (Zhang et al., 2022)	ReLU	plain FFN	0	12	0.7	0.0589
OPT-350m	ReLU	plain FFN	0	24	0.7	0.0848
OPT-1.3b	ReLU	plain FFN	0	24	0.7	0.0352
Pythia-160m (Biderman et al., 2023)	GELU	plain FFN	0	12	0.7	—
Pythia-410m	GELU	plain FFN	0	24	0.5	0.1428
MiniCPM-S-1B-sft (Song et al., 2025)	FATReLU	SwiGLU	0.03	52	0.5	0.1899
ProSparse-7B (Song et al., 2025)	FATReLU	SwiGLU	0.01	32	0.6	0.163
ReluLLaMA-7B (SparseLLM Team, 2023)	ReLU	SwiGLU	0	32	0.6	0.339

step cap 3000, early stopping on a held-out 20% slice of the calibration tokens scored with its *own* objective, the mask re-applied after every step so that all objectives optimise the same free parameters, and the loss normalised by the target's own mean square so that a shared learning rate does not favour one objective. The input X to layer ℓ is taken from the current, already pruned and repaired model (dense inputs make every objective worse than no repair by 0.37–1.60 nats), and the target is the dense weights on those same inputs, the convention of Eq. (2).

Data. Calibration is drawn from a slice of MiniPile (Kaddour, 2023), a subset of The Pile (Gao et al., 2020); the held-out in-domain probe comes from the same file past an offset, with a guard that the calibration span does not reach it, and the out-of-domain probe from a WikiText-103 slice. WikiText-2 is the raw test set at context 2048, cut into non-overlapping windows with no BOS per window, on which our evaluation code reproduces the published dense perplexities, 27.663 vs 27.65 (OPT-125m) and 14.624 vs 14.62 (OPT-1.3b).

Noise floors and evaluation. At calibration seeds 4000 and 91000 the no-repair baseline reproduces exactly, because the seed controls the reconstruction minibatch order, not the saliency; the largest spread of a repair objective on the 256-token probe, $L_{\text{gated_h}}(30)$ on OPT-1.3b at $s = 0.7$, is the probe floor of 0.0068 nats used throughout and the equivalence margin δ of Table 1. On WikiText-2 at context 2048, calibration seeds 4000, 77 000 and 91 000 give a spread over repair objectives with median 0.0079 and maximum 0.0284 nats, whose 90th percentile, 0.0150 nats, is the WikiText-2 floor. Evaluated inside the same run on identical repaired weights, the 256-token probe understates the repair effect of the published protocol by $2.02\times$ at $s = 0.4$ and $4.13\times$ at $s = 0.5$, with the same ordering of repair objectives at $s \leq 0.7$; WikiText-2 figures use the published protocol.

J VERIFICATION OF THE IDENTITY

Eq. (2) is an identity, so the check verifies the implementation: both sides are formed on real activations, and their relative gap is at floating-point level in every cell (Table 11).

Table 11: Identity gap $|\text{LHS} - \text{RHS}|/\text{LHS}$ for Eq. (2), maximum over the cells of each family.

family	models	s	cells	max. identity gap
OPT	OPT-125m, OPT-350m, OPT-1.3b	0.1–0.9	27	5.80×10^{-5}
7B ReLU family	ReluLLaMA-7B, ProSparse-7B	0.1–0.9	18	9.02×10^{-5}

K REPAIR EXPERIMENTS

Design. The objectives compared in a cell share one frozen Wanda-style mask and everything else of Appendix I, so only the functional changes; probe gaps are compared with the floor of 0.0068

Table 12: WikiText-2 perplexity at context 2048, OPT-1.3b, dense $= 14.62$. One frozen mask per row; only the functional changes. $L_{\text{post}} < L_{\text{pre}}$ at every sparsity, and at 3 of the 4 by more than the WikiText-2 noise floor 0.0150 nats.

s	no repair	L_{pre}	L_{gated}	L_{post}	$L_{\text{gated_h}}(30)$	$L_{\text{post_w}}(10)$
0.5	19.76	20.76	20.41	20.54	**20.34**	20.54
0.6	29.37	34.33	31.56	31.46	**30.04**	30.91
0.7	97.25	153.64	100.66	100.07	**84.50**	94.12
0.8	2643.77	2334.70	879.56	**740.52**	902.77	776.50

Table 13: Shares of the post-activation residual by term, and the fraction of (t, i) pairs crossing in each direction, OPT-1.3b $s = 0.7$, averaged over the 24 wall matrices.

	residual share			flips	
objective	in-cell	o→c	c→o	o→c	c→o
L_{gated}	.602	.027	**.372**	.0018	.0115
$L_{\text{gated_h}}(1)$	.767	.062	.171	.0027	.0070
L_{post}	.747	.143	.111	.0044	.0046
L_{pre}	.805	.137	.058	.0052	.0034
$L_{\text{post_w}}(10)$	.724	.224	**.053**	.0071	.0025
$L_{\text{gated_h}}(30)$	.802	.150	**.049**	.0065	.0028

nats and WikiText-2 gaps with WikiText-2's own (Appendix I), and Table 12 gives the WikiText-2 perplexities on OPT-1.3b.

The degeneracy is in the objective, not the solver. A degenerate minimiser is a claim about the objective, so it is tested on the minimiser. Solving L_{gated} in closed form in `fp64` (ridge 10^{-7} of the mean Gram diagonal), per row, with no step limit, costs $+1.9416$ nats against not repairing on OPT-1.3b at $s = 0.7$, and adds 237% of the error it was asked to remove. Inside the same solver, the identical `fp64` normal equations reproduce the iteratively solved pre-activation fit to -0.0029 nats. The early-stopped L_{gated} fit is at -0.01 nats. Results reported for this family in the pruning regime are therefore mediated by early stopping: the early-stopped fit has already left $\mathcal{P}_{\mathcal{X}}$, at $1.271\times$ the mask-only gate density, but halts long before the escape the closed-form solve reaches. L_{pre} stays accurate under the closed-form solve because it is a least-squares problem in every calibration token, not because it respects $\mathcal{P}_{\mathcal{X}}$; L_{gated} discards the inactive rows.

Movement and stopping controls. At the 3000-step budget on OPT-1.3b, $s = 0.7$, L_{post} reaches -0.1098 nats at relative weight movement 0.255, while `prematch`, which descends L_{pre} but halts at L_{post}'s movement, reaches $+0.2894$, and `pre_pval`, which descends L_{pre} but selects its checkpoint on the post-activation validation loss, reaches $+0.2258$.

Crossing directions. Table 13 is the source of the crossing rates of §G.3. Because Eq. (2) is exact, each objective's residual also splits by term: L_{gated} carries 37.2% of its residual in closed→open, $2.2\times$ the next-largest objective ($L_{\text{gated_h}}$ at $\lambda = 1$, 17.1%), and the *least* open→closed mass (2.7%). Both follow from what it charges: nothing for a closed→open token, since it sums only over $\mathcal{C}_{>}$, and, at $\theta = 0$, at least the true cost z_w^2 for an open→closed one, since it charges $(z_{\hat{w}} - z_w)^2 \geq z_w^2$ when $z_{\hat{w}} \leq 0 < z_w$.

L THE λ SWEEP AND THE $L_{\text{post}} < L_{\text{pre}}$ COUNTS

Of the configurations in which a $\lambda = 1$ and a $\lambda > 1$ member of one family are comparable, a $\lambda > 1$ member attains lower held-out loss in 10 of 11, in 8 of 11 by more than the noise floor, on plain-FFN ReLU, GELU and a SwiGLU block with a thresholded ReLU (Table 14); 5 of 9 rows are bold, at 5 to 27 times the floor, the second seed reproduces the sign in 6 of 6 cases, and the best λ runs from 3 to 30. Orderings are read within runs, because across runs that measure one configuration under different protocols $L_{\text{post}} - L_{\text{pre}}$ spreads by up to 1.04 nats.

Table 14: Held-out probe loss of the best $\lambda > 1$ member of an objective family against its $\lambda = 1$ member, sharing configuration, no-repair anchor, calibration seed and budgets. **Bold**: the gap exceeds $3\times$ the seed-noise floor (Appendix I) and a second calibration seed reproduces its sign; "seed 2 gap" is the gap of the same λ at that seed, "other groups" the gap in the other comparability groups of the configuration.

model	activation	$\lambda = 1$	best $\lambda > 1$	attained at	seed 2 gap	other groups
MiniCPM-S-1B, s=0.5	SwiGLU+FATReLU	3.4636	3.4594	$L_{\text{post_wv}}(10)$	—	+0.0043 to +0.0043
OPT-1.3b, s=0.5	ReLU	3.6847	3.6805	$L_{\text{post_w}}(10)$	—	—
OPT-1.3b, s=0.6	ReLU	4.0404	**3.9915**	$L_{\text{gated_h}}(30)$	+0.0437	+0.0283 to +0.0489
OPT-1.3b, s=0.7	ReLU	4.9442	**4.7631**	$L_{\text{gated_h}}(30)$	+0.1715	+0.1072 to +0.1811
OPT-1.3b, s=0.8	ReLU	6.4546	6.4381	$L_{\text{post_w}}(3)$	—	+0.0062 to +0.0165
OPT-125m, s=0.7	ReLU	5.6261	5.6121	$L_{\text{post_w}}(10)$	+0.0086	—
OPT-350m, s=0.7	ReLU	5.9919	**5.8906**	$L_{\text{gated_h}}(30)$	+0.0950	+0.0812 to +0.1014
Pythia-410m, s=0.5	GELU	4.6791	**4.6407**	$L_{\text{post_w}}(10)$	+0.0401	—
Pythia-410m, s=0.6	GELU	5.8639	**5.8273**	$L_{\text{post_w}}(10)$	+0.0727	—

population	$L_{\text{post}} < L_{\text{pre}}$	exception
probe cells, 256-token evaluation	46 of 47	OPT-125m s=0.3, +0.0058
configurations, under *every* protocol	23 of 24	the same one
WikiText-2 cells, 2048-token evaluation	15 of 15	—
7B ReLU-family cells	20 of 20	—

The probe cells span OPT-125m, OPT-350m, OPT-1.3b and two GELU models, the WikiText-2 cells the three OPT models and a SwiGLU block with a thresholded ReLU, and the 7B cells ReluLLaMA-7B and ProSparse-7B over two mask types, four sparsities and two calibration seeds. The 7B gaps run 0.0086–0.1594 nats, all beyond that family's seed-noise floor of 0.0061 nats; collapsing the two seeds, which share a mask, leaves 10 independent configurations, all negative, a sign test at $p = 2^{-10} = 9.8 \times 10^{-4}$.