How Much of *"Gradient Conflict"* Is Due to the Tasks, and How Much to the Coordinates?

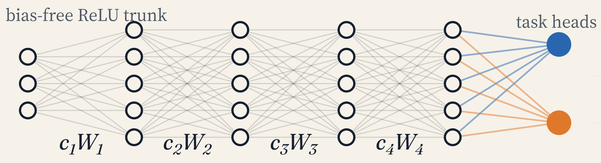
Tirth Joshi · Katz School of Science and Health, Yeshiva University · tjoshi1@mail.yu.edu

THE ANSWER, IN ONE LINE

*A change of ruler **can** flip the conflict sign, but rarely does. **The step size** moves the conflict rate by 74 points: at large steps, conflict hides behind one Hessian direction.*

1 WHAT

Rescale the layers: the network and its training run stay identical.

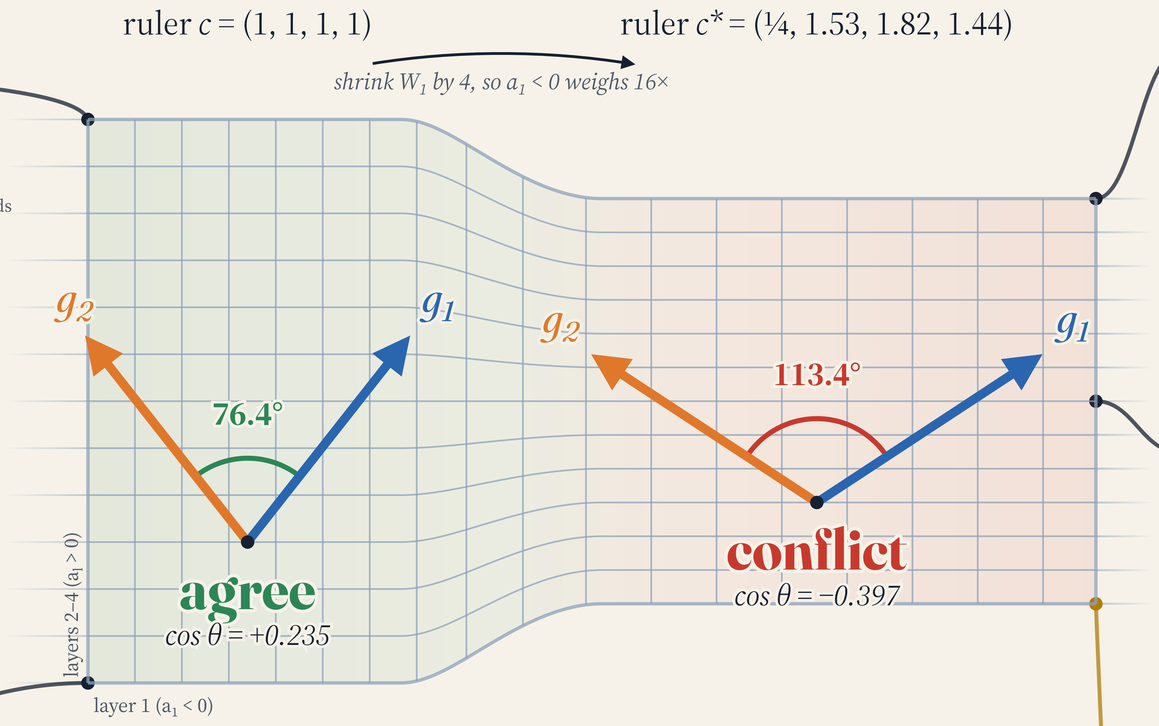


$$W_l \mapsto c_l W_l, \quad \eta_l \mapsto c_l^2\,\eta_l, \quad \prod_l c_l = 1$$

Positive homogeneity keeps the function; $\partial L/\partial(c_l W_l) = c_l^{-1} g_l$ keeps every GD step, so $W_l^{(c)}(t) = c_l\, W_l(t)\ \ \forall t$.
Measured: weights agree to **4×10⁻¹⁵**, losses to **4×10⁻¹⁶**.

But the reported **conflict**, the sign of the trunk cosine, is **not** invariant:

$$\cos_c(g_1, g_2) = \frac{\sum_l c_l^{-2}\, a_l}{\|g_1\|_c\, \|g_2\|_c}, \qquad a_l = \langle g_1^l, g_2^l \rangle$$



same network · same training run · different ruler

Fig. 1 A real training state, angles exact. One of the largest swings among 12,500 states; most flips sit near 90°.

2 HOW OFTEN CAN IT FLIP?

37.7% of training states have layer inner products of **both signs**

95% CI [36.8, 38.5] over 1000 seeds; it grows with depth:

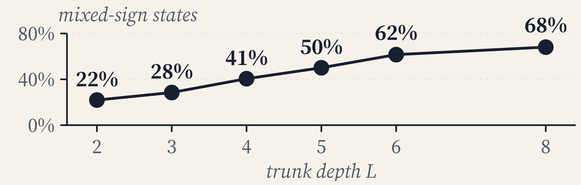


Fig. 2 Mixed-sign states by trunk depth L (50 seeds each, 95% intervals).

SO WHAT?

Gauge dependence appears insufficient to explain why these methods fail to beat loss summation (Kurin et al. '22; Xin et al. '22). The sign is coordinate-defined, but its frequency may primarily reflect optimization.

PAPER, CODE, REAL-DATA RESULTS & FULL REFERENCES

tirth5828.github.io/gradient-conflict-gauge

3 THE CONDITION

The verdict is coordinate-free ⟺ no two layers disagree.

a_1	a_2	a_3	a_4
−	+	+	+
×16	×0.43	×0.30	×0.49

Table 1 The Fig. 1 state: signs of a_l and weights $c_l^{*,-2}$. The lone negative layer is amplified 16×.

Observation. At a fixed step, sign $\cos_c(g_1, g_2)$ is the same for every gauge c iff no two a_l have strictly opposite signs.

Proof. The denominator is positive. ($\Leftarrow$) $N(c) = \sum_l c_l^{-2} a_l$ is a positive combination of same-sign terms. ($\Rightarrow$) If $a_i < 0 < a_j$, set $c_i = \varepsilon,\ c_j = \varepsilon^{-1}$: $N \to -\infty$ as $\varepsilon \to 0$; swap i, j for $N \to +\infty$. ■

Invariant form. $\langle g_1, g_2 \rangle_\eta = \sum_l \eta_l a_l$ is gauge-invariant and, to first order, the lookahead affinity of TAG (Fifty et al. '21); with uniform η the Euclidean sign equals it.

4 HOW HARD MUST YOU STRETCH?

Nearest flipping gauge: median factor **2.0**× per layer, **3.4**× for one neuron.

Ordinary conventions rarely get there: input-layer step ×1/16...×16 flips **4.0%**; random $\sqrt{2}$ factors, **1.2%**; trained layerwise step decay, $\leq$ **4.3** of 125 states.

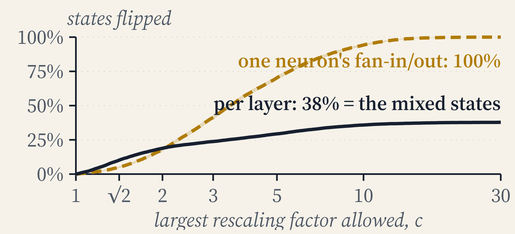


Fig. 3 Share of 12,500 states flipped by some gauge whose largest factor is at most c; 95% seed-bootstrap bands.

5 WHY IT MATTERS

Multi-task methods act on this quantity, in a metric they never state.

$$G = \begin{pmatrix} \|g_1\|^2 & \boxed{\langle g_1, g_2 \rangle} \\ \langle g_2, g_1 \rangle & \|g_2\|^2 \end{pmatrix}$$

PCGrad (Yu et al. '20): the sign of the boxed entry.
CAGrad, **MGDA**, **Nash-MTL**, **Aligned-MTL**: all of G.

PCGrad moves the step iff $\cos\theta < 0$, by exactly $|\cos\theta|\ \|u_{\text{sum}}\|$, a median **1.5%** where a $\sqrt{2}$ gauge flips the sign: **no benefit from the flip is detectable.**

6 THE RESULT

The ruler barely moves conflict. The step size moves it 74 points.

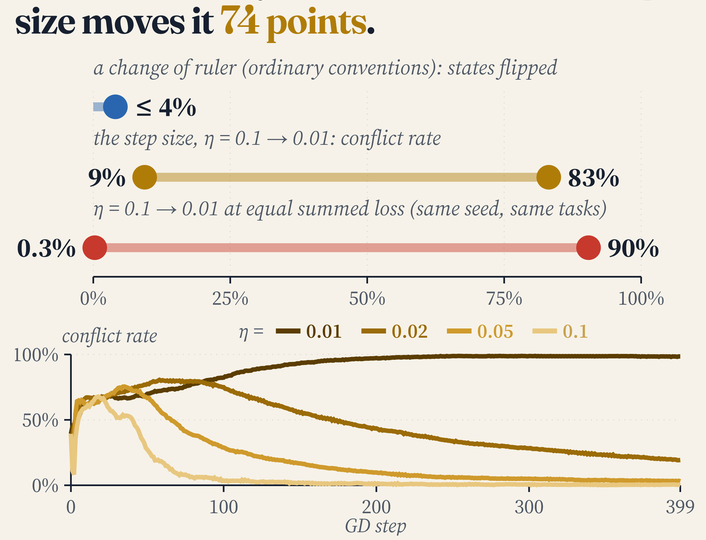
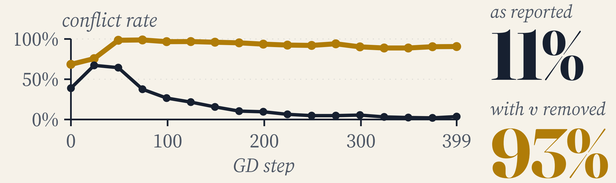


Fig. 4 Top: each knob, one scale. Bottom: conflict over training, 200 seeds × 5 ρ.

7 THE TWIST

Conflict hides behind one direction.

At the edge of stability (Cohen et al. '21; $\eta\lambda_{\max} \to 2$), **98–99%** of $\|g_1+g_2\|^2$ oscillates along the top Hessian eigenvector v. Project it out:



$$g_1 + g_2 \approx s\,v \ \Rightarrow\ P_{v^\perp} g_1 \approx -P_{v^\perp} g_2$$

Fig. 5 $\eta = 0.05$, 200 seeds, post-init. Off v the anti-alignment is bookkeeping; the finding is that the summed gradient collapses onto v at $\eta = 0.05$ and not at $\eta = 0.01$.