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*

I. The Question and the Answer

Multi-task learning often treats "gradient conflict," the sign of the cosine between per-task parameter gradients, as a task property. Yet a parameter-space cosine presupposes coordinates, hence a metric. PCGrad [3] acts on this sign; CAGrad [4], Nash-MTL [6], MGDA [5], and Aligned-MTL [7] act on the underlying Gram matrix; none specifies the metric. (Their -UB variants use the shared representation, fixed by the symmetry below.) We ask: *how much conflict comes from the tasks, and how much from the coordinates?* For bias-free ReLU trunks, the answer is exact. A layerwise rescaling with compensated step sizes preserves the entire trajectory, but preserves the sign iff the layerwise task-gradient inner products contain no strictly opposite pair. This condition fails in 37.7% of our training states. Even so, ordinary layerwise conventions flip only about 4% of states, and under uniform step size the sign agrees with that of the gauge-invariant $\sum_l \eta_l a_l$. Gauge dependence therefore appears insufficient to explain failures to improve on loss summation [1], [2]; in our full-batch censuses, the step-size regime instead moves the conflict *rate* by over 50 points (Sec. III).

II. Conjugacy, and When the Sign Is Invariant

A bias-free depth-$L \geq 2$ ReLU trunk with task-specific linear heads is trained by full-batch GD on the summed loss, with layerwise step sizes η_l. For a gauge c, $c_l > 0$ and $\prod_l c_l = 1$, set $W_l \mapsto c_l W_l$ and $\eta_l \mapsto c_l^2 \eta_l$; leave the heads fixed. **Conjugacy (standard).** Positive homogeneity preserves the function, while $\partial L/\partial(c_l W_l) = c_l^{-1} g_l$ for $g_l = \partial L/\partial W_l$. Thus $W_l^{(c)}(t) = c_l W_l(t)$ at every step (weights agree to 4×10^{-15}; losses to 4×10^{-16}). With $g_i^l = \partial L_i/\partial W_l$ and $a_l = \langle g_1^l, g_2^l \rangle$, the gauge-$c$ trunk cosine has numerator $N(c) = \sum_l c_l^{-2} a_l$ and a positive, gauge-dependent denominator. Only its sign can be invariant.

Observation. *At any fixed step, the reported sign is identical in every gauge iff no two a_l have strictly opposite signs.*

Invariant form. The optimizer-weighted form $\langle g_1, g_2 \rangle_\eta = \sum_l \eta_l a_l$ is gauge-invariant (flat to 2×10^{-17}) and, to first order, equals the lookahead affinity of [8]. The same argument holds entrywise under the finer neuron-wise symmetry [9], [11]; in every cell, the resulting neuron-wise inner products include both signs.

The precondition is common. Width-64, depth-4 trunks, 400 GD steps ($\eta = 0.05$), target correlation $\rho \in \{-0.5, -0.2, 0, 0.3, 0.7\}$ and checkpoints $0, 99, \ldots, 399$; each (seed, ρ, checkpoint) is a *cell*. The a_l include both signs in 37.7% of cells [95% CI 36.8, 38.5] over 1000 seeds (42% in the 5-seed grid below). Across width, depth, step-size, step-count, and ρ sweeps it ranges from 10% (ρ near 1) to 68% (depth 8; monotone in depth). These runs lie at the edge of stability [15]: after step 99, median $\eta\lambda_{\max} = 1.99$.

PCGrad. Compensated GD is gauge-equivariant; trunk PCGrad is not. Its update u_{PC} differs from u_{sum} exactly for negative Euclidean cosine, with $\|u_{\mathrm{PC}} - u_{\mathrm{sum}}\|/\|u_{\mathrm{sum}}\| = |\cos\theta|$. For cells flipped within factors $\sqrt{2}$, the correction has median 1.5% (q90 4%; >10% in under 1% of 12,500 cells). From the flipped gauge, PCGrad yields median final-loss ratios 1.00/1.00 relative to summed GD; same-norm random kicks yield 1.00/0.99. Across 106 of 160 flipped cells, the paired PCGrad-minus-kick difference is $+0.012$ $[+0.004, +0.029]$: no benefit from the flip is detectable.

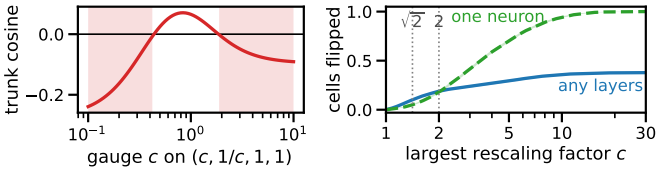


Fig. 1. **Left:** trunk cosine along $(c, 1/c, 1, 1)$ for seed 4, $\rho = 0.3$, step 99, the widest flipped interval among 26 of 125 cells on this path. Retrained conjugates keep losses flat to 1.2×10^{-16}; shading marks PCGrad activation. **Right:** fraction of 12,500 cells (500 seeds) flipped by some rescaling with largest factor $\leq c$, with 95% seed-bootstrap bands: layerwise, $\prod_l c_l = 1$ (solid; saturating at 38% mixed-sign cells), or one hidden neuron's fan-in/fan-out (dashed; reaching all cells).

III. Where Ordinary Practice Sits on the Orbit

Coordinates. Input-layer step sizes in $[\eta/16, 16\eta]$ (weights c_l^{-2} on a_l) flip about 5 of 125 cells. Directly trained layerwise decay, $\eta_l = \eta d^{L-l}$, departs from the invariant sign in at most 4.3 of 125 cells over $d \in [0.5, 0.8]$ and 500 seeds. Adversarially, the nearest flipping layerwise gauge has median largest factor 2.0 (10% flip within $\sqrt{2}$); the single-neuron median is 3.4 (Fig. 1, right).

Trajectory. Optimization matters far more. Across 200–1000 seeds, checkpoint conflict is 17% at $\eta = 0.05$ (12–22% across ρ), 83% at $\eta = 0.01$, and 9% at $\eta = 0.1$ (7% divergent runs excluded). At matched summed loss for the same seed and tasks, it is 0.3% for $\eta = 0.1$ (reached by 74% of runs), 3.6% for $\eta = 0.05$, and 90% for $\eta = 0.01$ (2000 steps). At the edge of stability, 98–99% of the summed gradient's squared norm lies along the oscillatory top Hessian eigenvector v at steps 199–399. Off v the task gradients are anti-aligned by construction; projecting it out raises conflict from 12% to 93%. Task correlation shows through in the slow regime; near convergence, the shrinking summed gradient drives the cosine negative whatever the tasks. On MNIST (digit + parity heads; bias-free $784{\times}256^3$ trunk; minibatch SGD), conjugacy holds to $2{\times}10^{-15}$; a factor-3 gauge flips 8 of 300 steps. The invariance condition fails in 6% of cells, but conflict remains below 3%: this aligned-task, three-epoch run stays below the edge of stability.

Scope and prior work. Path-SGD [9] and G-SGD [10] make the *trajectory* invariant; we fix it and inspect the *diagnostic*. GradVac [13], Recon [12], and RotoGrad [14] act on layerwise or representation conflict (layerwise-, not neuron-wise-gauge invariant). Biases extend exactly. Under LayerNorm *every* layerwise rescaling is a symmetry of the function alone and the sign is *more* fragile (55% mixed). Adam with $\varepsilon = 0$ admits $\eta_l \mapsto c_l \eta_l$; weight decay remains open. Thus the sign is coordinate-defined, but its frequency may primarily reflect optimization.

REFERENCES

[1] V. Kurin, A. De Palma, I. Kostrikov, S. Whiteson, and P. K. Mudigonda, "In defense of the unitary scalarization for deep multi-task learning," in *NeurIPS*, 2022. arXiv:2201.04122.

[2] D. Xin, B. Ghorbani, A. Garg, O. Firat, and J. Gilmer, "Do current multi-task optimization methods in deep learning even help?" in *NeurIPS*, 2022. arXiv:2209.11379.

[3] T. Yu, S. Kumar, A. Gupta, S. Levine, K. Hausman, and C. Finn, "Gradient surgery for multi-task learning," in *NeurIPS*, 2020.

[4] B. Liu, X. Liu, X. Jin, P. Stone, and Q. Liu, "Conflict-averse gradient descent for multi-task learning," in *NeurIPS*, 2021.

[5] O. Sener and V. Koltun, "Multi-task learning as multi-objective optimization," in *NeurIPS*, 2018.

[6] A. Navon, A. Shamsian, I. Achituve, H. Maron, K. Kawaguchi, G. Chechik, and E. Fetaya, "Multi-task learning as a bargaining game," in *ICML*, 2022.

[7] D. Senushkin, N. Patakin, A. Kuznetsov, and A. Konushin, "Independent component alignment for multi-task learning," in *CVPR*, 2023. arXiv:2305.19000.

[8] C. Fifty, E. Amid, Z. Zhao, T. Yu, R. Anil, and C. Finn, "Efficiently identifying task groupings for multi-task learning," in *NeurIPS*, 2021. arXiv:2109.04617.

[9] B. Neyshabur, R. Salakhutdinov, and N. Srebro, "Path-SGD: Path-normalized optimization in deep neural networks," in *NIPS*, 2015. arXiv:1506.02617.

[10] Q. Meng, S. Zheng, H. Zhang, W. Chen, Q. Ye, Z.-M. Ma, N. Yu, and T.-Y. Liu, "$\mathcal{G}$-SGD: Optimizing ReLU neural networks in its positively scale-invariant space," in *ICLR*, 2019. arXiv:1802.03713.

[11] L. Dinh, R. Pascanu, S. Bengio, and Y. Bengio, "Sharp minima can generalize for deep nets," in *ICML*, 2017. arXiv:1703.04933.

[12] G. Shi, Q. Li, W. Zhang, J. Chen, and X.-M. Wu, "Recon: Reducing conflicting gradients from the root for multi-task learning," in *ICLR*, 2023. arXiv:2302.11289.

[13] Z. Wang, Y. Tsvetkov, O. Firat, and Y. Cao, "Gradient Vaccine: Investigating and improving multi-task optimization in massively multilingual models," in *ICLR*, 2021. arXiv:2010.05874.

[14] A. Javaloy and I. Valera, "RotoGrad: Gradient homogenization in multitask learning," in *ICLR*, 2022. arXiv:2103.02631.

[15] J. M. Cohen, S. Kaur, Y. Li, J. Z. Kolter, and A. Talwalkar, "Gradient descent on neural networks typically occurs at the edge of stability," in *ICLR*, 2021. arXiv:2103.00065.