Low-Rank Networks Recover Weight and Functional Symmetry Better

Anonymous Author(s) [1]

Abstract

We study an empirical phenomenon in low-rank random-feature networks: even when stochastic training sees one non-symmetrized sample at a time, the learned model can recover a globally symmetric function and, in successful regimes, symmetric internal partial functions. The effect is not explained by output invariance alone. Low-loss counterexamples exist where the output is nearly symmetric but the learned partial functions remain asymmetric. We therefore analyze symmetry simultaneously in function space, representation space, and weight space. Across one-dimensional sum-of-cosines reruns, multidimensional batch-size-one experiments, and new positive-bias stress tests, RF-LR networks exhibit a distinctive regime where symmetry is visible in late partial functions and in mirror-paired first-layer atoms with matched outgoing weights. We isolate the provable core of this observation: mirror-paired atoms with matched outgoing coefficients exactly generate invariant partial functions, and approximate mirror pairing gives an explicit defect bound.

1. Question

Weight-space symmetries are usually discussed as exact parameter transformations that preserve a network function, especially in model merging and mode connectivity (Garipov et al., 2018; Frankle et al., 2020; Entezari et al., 2022; Ainsworth et al., 2023). Here we ask a different question: can optimization discover a *data-dependent* symmetry in weight space when the training procedure itself is maximally asymmetric? Our motivating setting uses a target invariant under $x \mapsto -x$ or related finite transformations, but trains with stochastic batch size one and without paired samples. The surprising observation is that RF-LR models often learn not only an invariant output but also invariant low-rank partial functions. Our weight-space test asks whether this functional symmetry is accompanied by an organization of the trained parameters themselves: first-layer atoms should appear in approximate mirror pairs, and the low-rank outgoing coefficients attached to paired atoms should agree.

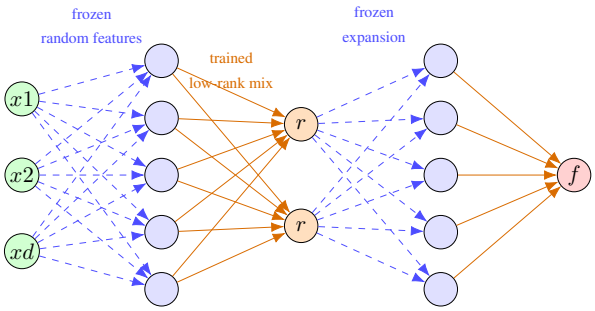

Figure 1. RF-LR alternates wide random-feature maps and low-rank bottlenecks. Symmetry is measured at the output, at bottleneck partials, and through mirror structure in first-layer atoms.

For a learned scalar function f, we measure

$$D_{\text{out}}(T) = \frac{\mathbb{E}_x\left[(f(x) - f(Tx))^2\right]}{\mathbb{E}_x\left[f(x)^2\right] + \varepsilon},$$

and apply the same relative defect to active internal partial functions. For first-layer ReLU atoms $\sigma(a_j^\top x + b_j)$, we also compare each atom to its nearest mirror $(-a_j, b_j)$ and measure the relative mismatch and correlation of their outgoing low-rank weights. Small mismatch means that symmetry is not only visible after evaluating the network; it is also encoded as paired atoms in weight space.

2. Low-Rank Architecture

We use the finite-width RF-LR model corresponding to the multi-component and multi-layer neural network line of work (Zhang et al., 2024). Let $z^{(0)}(x) = x \in \mathbb{R}^d$. For block $\ell = 1, \ldots, L$, the network applies

$$u^{(\ell)}(x) = \sigma\left(A^{(\ell)} z^{(\ell-1)}(x) + b^{(\ell)}\right),$$
$$z^{(\ell)}(x) = B^{(\ell)} u^{(\ell)}(x).$$

where $A^{(\ell)} \in \mathbb{R}^{W \times r_{\ell-1}}$ and $b^{(\ell)} \in \mathbb{R}^W$ are frozen i.i.d. random features, while $B^{(\ell)} \in \mathbb{R}^{r_\ell \times W}$ is trained. The ranks are $r_0 = d$, $r_L = 1$, and $r_\ell = r \ll W$ for hidden

[1]Anonymous Institution. Correspondence to: Anonymous Author <anonymous@example.com>.

Workshop on Weight-Space Symmetries, held in conjunction with the 43^{rd} International Conference on Machine Learning, Seoul, South Korea. 2026.

bottlenecks. Thus each layer first expands to width W, then contracts through a trained rank-r channel mixer. The scalar prediction is $f_\theta(x) = z^{(L)}(x)$. Equivalently, $A^{(1)}$ is the frozen first feature map, and the entries of $B^{(\ell)}$ are the trainable channel weights.

The internal partial functions are the bottleneck coordinates

$$p_k^{(\ell)}(x) = z_k^{(\ell)}(x), \qquad k = 1, \dots, r_\ell.$$

These are the objects on which we measure active partial symmetry defects. The first-layer atom associated with row a_j and bias b_j is $\sigma(a_j^\top x + b_j)$; weight-space symmetry is probed by looking for mirror atoms $\sigma((-a_j)^\top x + b_j)$ and comparing their trained outgoing coefficients in $B^{(1)}$.

The architectural bias is not an explicit equivariance constraint. The first random-feature layer is sampled asymmetrically, and mini-batches contain a single point. Nevertheless, the trained low-rank channel weights can organize mirror-related atoms into an approximately symmetric representation.

Provable core. The empirical claim is about optimization: RF-LR training often finds mirror-organized weights. The clean theorem is the implication from mirror organization to internal symmetry. For $p(x) = \sum_{j=1}^{W} c_j \sigma(a_j^\top x + b_j)$, write $\bar{j}$ for the atom nearest to $(-a_j, b_j)$.

Proposition 1 (Mirror atoms imply symmetric partials)**.** *Assume the domain satisfies $\|x\| \le R$ and σ is 1-Lipschitz. If atoms are exactly paired, $a_{\bar{j}} = -a_j$, $b_{\bar{j}} = b_j$, and $c_{\bar{j}} = c_j$ for every pair, then $p(x) = p(-x)$. More generally, if $\|a_{\bar{j}} + a_j\| \le \delta_a$, $|b_{\bar{j}} - b_j| \le \delta_b$, $|c_{\bar{j}} - c_j| \le \eta$, and $|c_j| \le C$, then*

$$|p(x) - p(-x)| \le \eta \sum_{j=1}^{W} |\sigma(a_j^\top x + b_j)| + CW(\delta_a R + \delta_b).$$

The same statement applies coordinatewise to each bottleneck partial.

The proof is a direct pairing argument plus Lipschitz continuity of ReLU, and the exact statement propagates through RF-LR blocks by induction. A full optimization theorem can therefore target the remaining step: under a symmetric distribution, show that the bottleneck mixer suppresses the anti-symmetric channel component. The experiments test the two measurable premises of that theorem: small partial defects and small mirror-coefficient mismatch.

Table 1. Multidimensional batch-one protocol and representative results. No $(x, -x)$ pair is inserted during training; symmetry metrics are post-hoc on a separate symmetric grid.

Data	model	rank	test MSE	D_{out}	active D_p
2D Gauss.	RF-LR	8	1.20×10^{-5}	6.91×10^{-5}	6.62×10^{-4}
2D Quad.	RF-LR	8	2.17×10^{-5}	7.80×10^{-5}	9.98×10^{-4}
3D Quad.	RF-LR	8	1.92×10^{-4}	9.76×10^{-4}	1.81×10^{-3}
2D Quad.	RF-LR ctrl.	128	1.60×10^{-4}	3.07×10^{-4}	1.37×10^{-2}
3D Quad.	RF-LR ctrl.	128	2.52×10^{-4}	1.27×10^{-3}	7.38×10^{-3}

All rows: $W = 512$, Adam 10^{-3}, $10k$ epochs, batch 1; frozen $A^{(\ell)}, b^{(\ell)}$, train only $B^{(\ell)}$.

3. Evidence

One-dimensional sum of cosines. We train low-loss even sum of cosines configurations with width 1024 and adaptive SGD. The target family is $g_m(x) = \sum_{q=1}^{m} \alpha_q \cos(\omega_q x)$ with fixed frequencies and coefficients, following the high-frequency motivation of MMNN/FMMNN models (Zhang et al., 2024; 2025); every term is invariant under $x \mapsto -x$, but training samples are not duplicated with mirrored partners. This makes the task a controlled probe of whether RF-LR discovers even symmetry from optimization rather than explicit augmentation. Across six reliable low-loss runs, final test MSE ranges from $1.94 \cdot 10^{-5}$ to $3.26 \cdot 10^{-3}$. Output even defects range from $4.1 \cdot 10^{-6}$ to $8.0 \cdot 10^{-4}$, while active last-layer partial even defects range from $6.24 \cdot 10^{-5}$ to $5.22 \cdot 10^{-4}$. This is the cleanest evidence that the learned symmetry is internal, not only an output coincidence.

Counterexamples are useful. The completed checkpoint set contains 149 one-dimensional RF-LR reruns: 12 partial-symmetric, 28 output-only/asymmetric, 100 intermediate, and 9 underfit runs. For example, one rank-5 depth-3 f_2 run reaches test MSE $9.53 \cdot 10^{-4}$ with partial defect $2.62 \cdot 10^{-5}$. In contrast, a rank-10 depth-1 f_1 run reaches test MSE $1.17 \cdot 10^{-4}$ but partial defect 2.17. Thus low output loss does not force internal symmetry; the positive low-rank regime is a real representation phenomenon.

Batch-one multidimensional stress tests. We then stress-test the strongest claim in two and three dimensions: symmetry can emerge from a fully asymmetric stochastic stream. The main runs use batch size one, non-paired uniform training samples, Adam, width 512, and bottlenecks with rank 8 or 16, with rank-$64/128$ RF-LR and matched MLP controls. The training algorithm is deliberately simple: sample all random-feature matrices $A^{(\ell)}$ and biases $b^{(\ell)}$ once, freeze them, initialize only the channel mixers $B^{(\ell)}$, then run gradient descent on MSE through the full RF-LR forward map while updating only $B^{(\ell)}$. For the multidimensional runs each optimizer step sees one unpaired point x, never the pair $(x, -x)$; best checkpoints are selected by held-out MSE, and all output, partial, and mirror-symmetry defects are computed only after training on a separate symmetric evaluation grid. The 2D Gaussian run with rank 8 and depth 3 reaches test MSE $1.20 \cdot 10^{-5}$, output defect $6.91 \cdot 10^{-5}$ under $x \mapsto -x$, and active last-partial defect $6.62 \cdot 10^{-4}$.

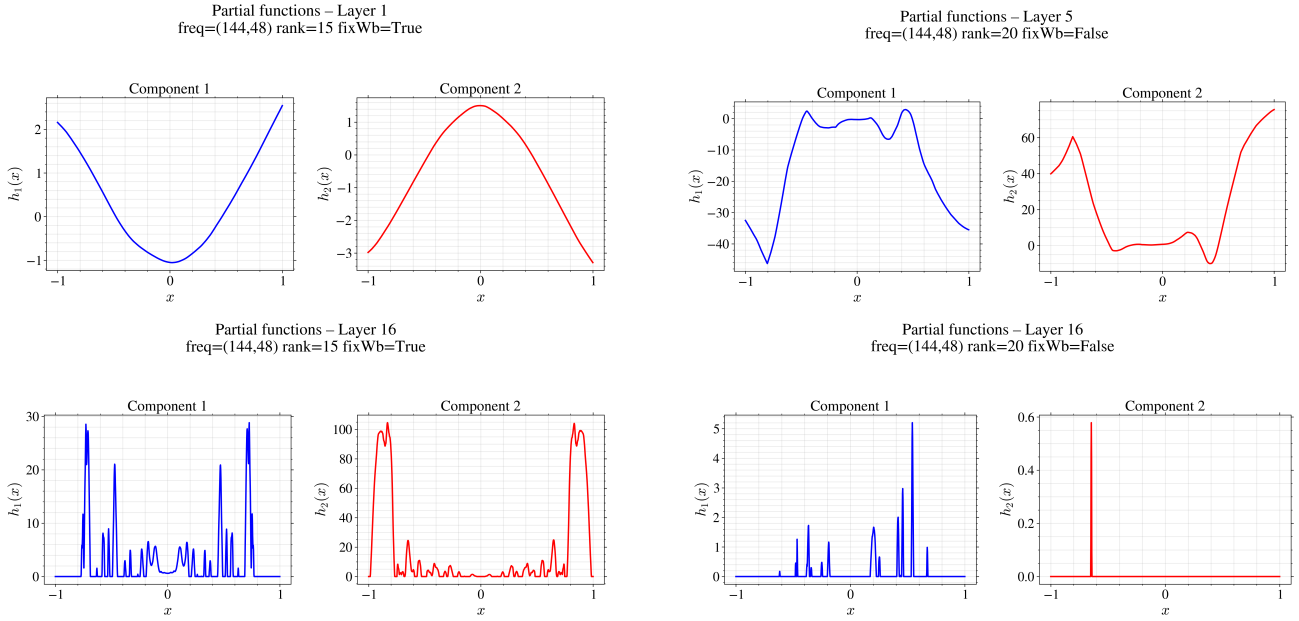

Figure 2. RF-LR partial functions on the even high-frequency target and no-RF MLP asymmetry controls. The two layer-16 panels are enlarged to show that RF-LR keeps mirror structure at depth, while the no-RF MLP control remains visibly asymmetric.

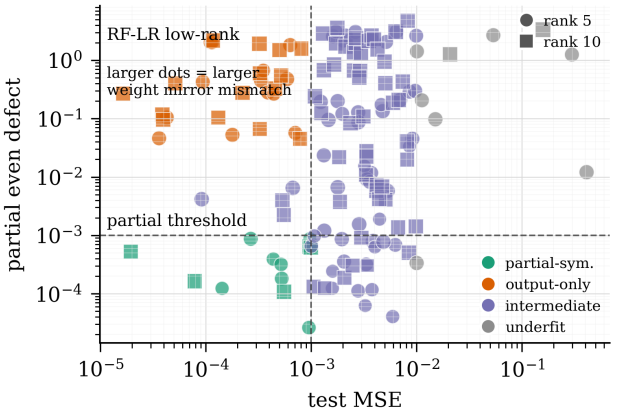

Figure 3. Weight-space symmetry diagnostic for completed one-dimensional RF-LR low-rank checkpoints. The axes are test MSE and active partial even defect; colors show functional regimes, circles/squares are rank $5/10$, and marker size is the median outgoing-weight mismatch between nearest mirror atoms (a, b) and $(-a, b)$. Thus the plot should be read as a joint function/weight-space map: lower-left points are low-loss, internally symmetric, and have smaller mirror mismatch, while output-only points show that fitting the target does not force symmetric partials or organized mirror weights.

The 3D quadratic run with rank 8 and depth 3 reaches test MSE $1.92 \cdot 10^{-4}$, output defect $9.76 \cdot 10^{-4}$, and active last-partial defect $1.81 \cdot 10^{-3}$. On the 2D quadratic target, increasing rank to 128 keeps the output fit strong but worsens active partial symmetry to $1.37 \cdot 10^{-2}$, while a matched MLP reaches test MSE $7.5 \cdot 10^{-5}$ with active partial defect $2.4 \cdot 10^{-1}$. Thus the phenomenon survives beyond the original one-dimensional setting while using rank-to-width ratio $8/512 = 1.56\%$.

We also ran 24 new positive-bias stress tests with width 256, $N = 128$, two seeds, 2D/3D Gaussian and quadratic targets, RF-LR ranks 8 and 128, and matched MLPs. The training distribution deliberately concentrates samples in the positive orthant, with mean nearest-mirror distance 0.22 in 2D and 0.54–0.60 in 3D. Among low-loss positive-bias runs, MLPs reach median test MSE $3.28 \cdot 10^{-4}$ but median active partial defect 1.30. RF-LR reaches comparable median test MSE ($3.59 \cdot 10^{-4}$ for rank 8, $2.93 \cdot 10^{-4}$ for rank 128) with active partial defects $1.13 \cdot 10^{-2}$ and $6.57 \cdot 10^{-3}$. So the new controls support the representation claim sharply: ordinary MLPs can fit the same asymmetric stream, but their hidden features remain asymmetric by two orders of magnitude.

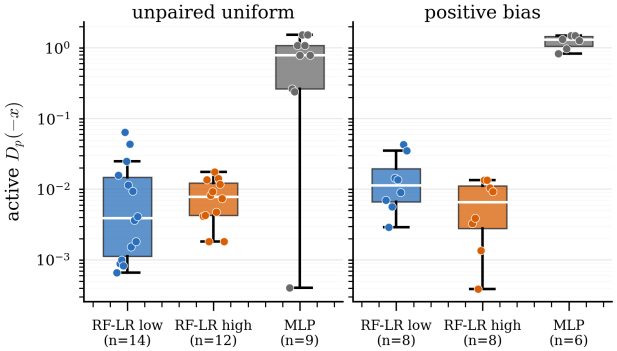

Figure 4. Aggregate multidimensional evidence over low-loss runs. RF-LR low rank means rank 8 or 16; RF-LR high rank means rank 64 or 128. Under both unpaired-uniform and positive-bias training, RF-LR partial functions are far more symmetric than MLP hidden features at comparable loss.

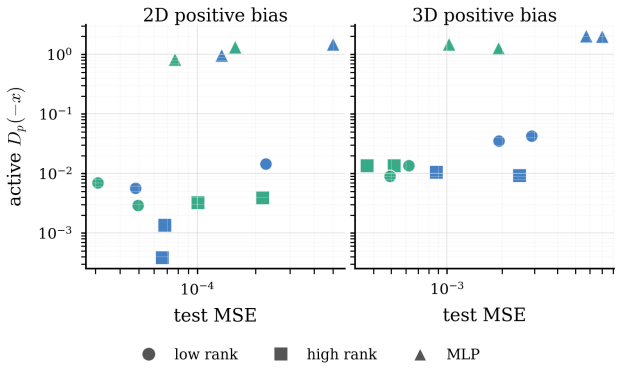

Figure 5. New positive-bias stress test. Circles are rank-8 RF-LR, squares are rank-128 RF-LR, and triangles are MLP controls. Colors distinguish Gaussian and quadratic targets. MLPs can fit the output but stay high in active partial defect; RF-LR stays in the low-defect regime.

Figure 3 is the main weight-space evidence. It does not plot weights directly; instead it compresses the first-layer mirror-pair test into point size, so a good run should sit in the lower-left and avoid large markers. This is precisely the regime where RF-LR has low loss, symmetric partials, and comparatively matched outgoing weights on mirror atoms. Figures 4 and 5 are the main multidimensional summaries. They show that batch-one asymmetric training can recover symmetric outputs and, for RF-LR, symmetric internal partial functions. The rank comparison is deliberately nuanced: more rank is not a monotone explanation, and under the positive-bias stress test rank 128 can also recover symmetric partials. The robust separation is between RF-LR bottleneck representations and unconstrained MLP hidden representations.

The key message for weight-space symmetries is that useful symmetries need not be imposed as exact architectural equivariances or explicit path-norm constraints (Neyshabur et al., 2015). In low-rank random-feature networks, the optimizer can discover approximate, data-dependent symmetry in the trained low-rank weights. This suggests a practical analysis program: compare models not only by their functions, but by whether the symmetry appears in their internal partials and mirror-organized parameters.

4. Conclusion and Limitations

Our final conclusion is that RF-LR networks recover functional symmetry better in the regimes where the target is actually learned, and that this recovery leaves a measurable trace in weight space. The strongest evidence is not just small output defect: successful RF-LR runs also show small active partial-function defects, meaning the symmetry is present inside the learned representation. The rank ablations show that simply increasing rank is not the whole explanation; the decisive separation is between RF-LR bottleneck/random-feature representations and ordinary MLP hidden features. The mirror-pair measurements show the corresponding parameter signature: nearest mirrored atoms have more compatible outgoing weights in the successful regime. Thus the RF-LR bottleneck acts as an implicit symmetry-recovery bias rather than merely as a parameter-count reduction. The limitation is that the mirror metric is nearest-neighbor based and only probes the first layer; it supports the weight-space interpretation but remains weaker than the partial-function evidence. Some harder 3D soft-axis targets still generalize poorly despite tiny training error, so the multidimensional claim is for the low-loss Gaussian and quadratic regimes, not all symmetric targets.

References

Ainsworth, S. K., Hayase, J., and Srinivasa, S. Git re-basin: Merging models modulo permutation symmetries. In *International Conference on Learning Representations*, 2023.

Entezari, R., Sedghi, H., Saukh, O., and Neyshabur, B. The role of permutation invariance in linear mode connectivity of neural networks. In *International Conference on Learning Representations*, 2022.

Frankle, J., Dziugaite, G. K., Roy, D. M., and Carbin, M. Linear mode connectivity and the lottery ticket hypothesis. In *International Conference on Machine Learning*, 2020.

Garipov, T., Izmailov, P., Podoprikhin, D., Vetrov, D. P., and Wilson, A. G. Loss surfaces, mode connectivity, and fast ensembling of DNNs. In *Advances in Neural Information Processing Systems*, 2018.

Neyshabur, B., Salakhutdinov, R. R., and Srebro, N. Path-SGD: Path-normalized optimization in deep neural networks. In *Advances in Neural Information Processing Systems*, 2015.

Zhang, S., Zhao, H., Zhong, Y., and Zhou, H. Structured and balanced multi-component and multi-layer neural networks. *arXiv preprint arXiv:2407.00765*, 2024.

Zhang, S., Zhao, H., Zhong, Y., and Zhou, H. Fourier multi-component and multi-layer neural networks: Unlocking high-frequency potential. *arXiv preprint arXiv:2502.18959*, 2025.