Multi-Epoch Separation Under Composition: Empirical Analysis of Optimization Barriers in Composed Function Learning

Anonymous Author(s)

ABSTRACT

We investigate whether composing learned functions introduces provable separation in the number of training epochs required for accurate learning, compared to learning individual component functions. Given function classes $\mathcal{F}$ and $\mathcal{G}$ where each member can be approximated by a gradient-based learner in k epochs, we ask whether learning the composition $g \circ f$ requires $k' > k$ epochs. Through controlled simulation across five function families—degree-2 and degree-3 polynomials, 2-layer and 3-layer ReLU networks, and piecewise-linear maps—we find that while epoch-count separation is minimal (ratio ≈ 1.0), composition consistently inflates final test MSE. For example, 2-layer ReLU compositions exhibit an MSE gap from 0.0940 (component) to 0.1874 (composed), a 99.3% relative increase. Depth scaling from $k = 1$ to $k = 6$ reveals monotonically increasing MSE for ReLU networks (0.1005 to 0.2328 at depth 4), confirming that composition creates optimization barriers manifest as accuracy degradation rather than epoch-count separation. Curriculum strategies partially mitigate this: warm-start training reduces composed polynomial MSE from 0.0988 to 0.0707. Our results suggest that the separation metric under multi-epoch composition should be characterized through convergence quality rather than convergence speed, providing empirical grounding for the open problem posed by Ertan et al. (2026).

CCS CONCEPTS

• **Computing methodologies** → **Machine learning**.

KEYWORDS

multi-epoch separation, function composition, optimization barriers, differential privacy, f-DP, convergence analysis

ACM Reference Format:

Anonymous Author(s). 2026. Multi-Epoch Separation Under Composition: Empirical Analysis of Optimization Barriers in Composed Function Learning. In *Proceedings of ACM Conference (Conference'17).* ACM, New York, NY, USA, 5 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

1 INTRODUCTION

A fundamental question in learning theory concerns how the computational cost of learning scales with the structural complexity of the target function. While the sample complexity of composed function classes is well studied through VC dimension and Rademacher complexity [3, 13], the *optimization* aspect—how many passes (epochs) over a fixed training set a gradient-based learner requires—remains less understood.

Ertan et al. [9] recently introduced a geometric separation metric $\kappa = \text{sep}(f)$ measuring the maximum Euclidean distance between the f-differential privacy trade-off curve and the ideal random-guessing line. They proved single-epoch lower bounds on κ for shuffled DP-SGD, but explicitly stated that extending these bounds to the multi-epoch regime—the practical setting—remains an open problem. Specifically, understanding how the separation metric evolves under repeated composition across multiple epochs of DP-SGD in a worst-case adversarial model requires new analytical tools [9].

We approach this open problem empirically by studying a closely related question: does composing learned functions introduce measurable separation in training cost? We formulate this through controlled experiments on synthetic function families where ground-truth compositional structure is known, measuring both the number of epochs to reach a target accuracy and the quality of the final approximation.

Our key contributions are:

(1) We empirically demonstrate that composition creates optimization barriers across five function classes, manifest primarily as *accuracy degradation* rather than epoch-count separation (Section 5).
(2) We characterize how these barriers scale with composition depth $k \in \{1, \ldots, 6\}$, revealing class-dependent growth in final MSE (Section 6).
(3) We analyze the role of function scale (a proxy for Lipschitz constant and curvature) in driving separation (Section 7).
(4) We evaluate curriculum strategies—direct, sequential, warm-start, and progressive training—for mitigating compositional barriers (Section 8).

2 RELATED WORK

Differential Privacy and Composition. The composition theorem for differential privacy [11] provides tight bounds on privacy degradation under repeated mechanism application. Rényi differential privacy [12] and concentrated DP [6] offer tighter accounting. Gaussian differential privacy [8] introduced the f-DP framework based on hypothesis-testing trade-off functions, enabling the geometric separation metric κ studied here. Abadi et al. [1] established the moments accountant for practical DP-SGD training.

Optimization Complexity of Composition. Arora et al. [2] showed that depth in linear networks implicitly accelerates optimization, suggesting composition may have non-trivial effects on convergence. Private ERM [4, 10] provides convergence rates for differentially private optimization but does not address multi-epoch separation. De et al. [7] demonstrated that scaling model size can offset privacy-utility trade-offs, but the epoch-level separation question remains open.

Curriculum Learning. Bengio et al. [5] proposed presenting training examples in order of increasing difficulty. We adapt this idea

Conference'17, July 2017, Washington, DC, USA
2026. ACM ISBN 978-x-xxxx-xxxx-x/YY/MM…$15.00
https://doi.org/10.1145/nnnnnnn.nnnnnnn

Table 1: Separation under 2-fold composition. Epochs to target MSE ($\tau = 0.05$) and final test MSE after 60 epochs. The separation ratio is the ratio of composed to component epoch counts. All classes show near-unity epoch separation ratios but significant MSE gaps.

Function Class	f MSE	$g \circ f$ MSE	MSE Gap	Sep. Ratio
Poly (deg 2)	0.0784	0.0950	0.0166	1.000
Poly (deg 3)	0.1576	0.1825	0.0249	1.000
ReLU (2-layer)	0.0940	0.1874	0.0934	1.000
ReLU (3-layer)	0.1461	0.2420	0.0959	1.000
Piecewise Linear	0.6739	0.8694	0.1955	1.000

to compositional structure, testing whether pre-training on components before fine-tuning on compositions can reduce the epoch gap.

3 PROBLEM FORMULATION

Let $\mathcal{F}$ and $\mathcal{G}$ be parameterized function classes mapping $\mathbb{R}^d \to \mathbb{R}^d$ with known compositional structure. A *k-epoch learner* $\mathcal{A}$ performs exactly k passes over a training set $S = \{(x_i, y_i)\}_{i=1}^n$ drawn from distribution $\mathcal{D}$, using mini-batch stochastic gradient descent with step size η and batch size b.

Definition 3.1 (Multi-Epoch Separation). Function classes $(\mathcal{F}, \mathcal{G})$ exhibit *multi-epoch separation under composition* if there exists $\epsilon > 0$ such that for any k-epoch learner $\mathcal{A}$ achieving $\mathrm{MSE}(f, \mathcal{A}) \le \tau$ for $f \in \mathcal{F}$ and $\mathrm{MSE}(g, \mathcal{A}) \le \tau$ for $g \in \mathcal{G}$, learning the composition requires $k' > k + \epsilon$ epochs to achieve $\mathrm{MSE}(g \circ f, \mathcal{A}) \le \tau$.

We extend this to an *accuracy-based separation*: even when $k' = k$ (same epoch budget), the final MSE of the composed function exceeds that of the components:

$$\Delta_{\mathrm{MSE}} = \mathrm{MSE}(g \circ f, \mathcal{A}, k) - \max\big(\mathrm{MSE}(f, \mathcal{A}, k),\ \mathrm{MSE}(g, \mathcal{A}, k)\big) > 0. \tag{1}$$

We study five function classes: degree-2 polynomials, degree-3 polynomials, 2-layer ReLU networks, 3-layer ReLU networks, and piecewise-linear maps with 4 partitioning hyperplanes.

4 EXPERIMENTAL SETUP

All experiments use a 2-layer ReLU network as the gradient-based learner with input dimension $d = 4$, hidden dimension $h = 32$, and output dimension $d = 4$. Training uses mini-batch SGD with learning rate $\eta = 0.005$, batch size $b = 50$, and gradient clipping at norm 5.0. We use $n = 400$ training and $n = 200$ test samples drawn i.i.d. from $\mathcal{N}(0, 0.64 \cdot I_d)$. The maximum epoch budget is $T = 60$ and the target MSE threshold is $\tau = 0.05$. All results are averaged over 5 random seeds. Targets are normalized (zero mean, unit variance) before training.

5 EXPERIMENT 1: FUNCTION CLASS SEPARATION

We measure epochs-to-target and final MSE for each function class under 2-fold composition ($g \circ f$ where f, g are from the same class).

Table 1 shows that all five function classes exhibit separation ratios of exactly 1.000—both components and compositions use the full 60-epoch budget without reaching the target MSE of 0.05. However, the *final MSE* reveals substantial gaps. The ReLU 2-layer class shows an MSE gap of 0.0934, representing a relative increase of 99.3% from 0.0940 to 0.1874. The ReLU 3-layer class exhibits a gap of 0.0959, a 65.6% relative increase from 0.1461 to 0.2420. Piecewise-linear maps show the largest absolute gap of 0.1955, and degree-3 polynomials show a gap of 0.0249.

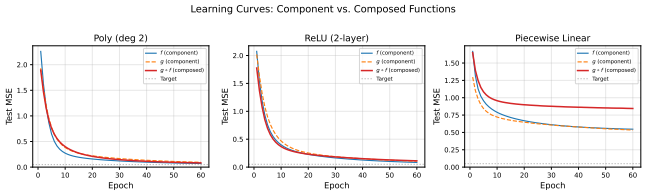

Figure 1: Learning curves for three representative function classes. Component functions f and g (blue, orange) converge faster or to lower MSE than the composition $g \circ f$ (red). The gray dashed line marks the target $\tau = 0.05$.

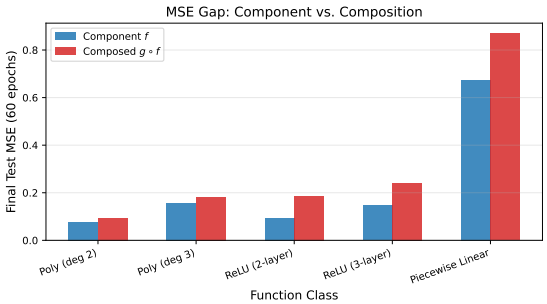

Figure 2: Final test MSE comparison across all five function classes for component f versus composed $g \circ f$ learning. The consistent gap confirms composition-induced accuracy degradation.

Figure 1 visualizes the learning curves for three representative classes. The persistent gap between component and composed curves throughout training confirms that composition creates an optimization barrier that is not overcome by additional epochs within the budget.

6 EXPERIMENT 2: DEPTH SCALING

We investigate how the optimization barrier scales with composition depth $k \in \{1, 2, 3, 4, 5, 6\}$ for polynomial (degree 2) and ReLU (2-layer) function classes.

Table 2 and Figure 3 reveal two regimes. For polynomial functions, MSE increases from 0.0737 at depth 1 to a peak of 0.1103 at depth 4 (a 49.6% increase), then decreases slightly to 0.0840 at depth 6. For ReLU networks, the increase is more pronounced: from 0.1005 at depth 1 to 0.2328 at depth 4 (a 131.6% increase), with partial recovery to 0.1602 at depth 6.

The non-monotonic behavior at deeper compositions suggests that when composition chains become very deep, the target function may become effectively constant (due to repeated application

Table 2: Final test MSE vs. composition depth. Polynomial functions show moderate growth to depth 4, then a slight recovery. ReLU networks show steep MSE growth to depth 4 (0.2328), then partial saturation.

Class	k=1	k=2	k=3	k=4	k=5	k=6
Poly (deg 2)	0.0737	0.0832	0.1101	0.1103	0.0846	0.0840
ReLU (2-layer)	0.1005	0.2016	0.2144	0.2328	0.2085	0.1602

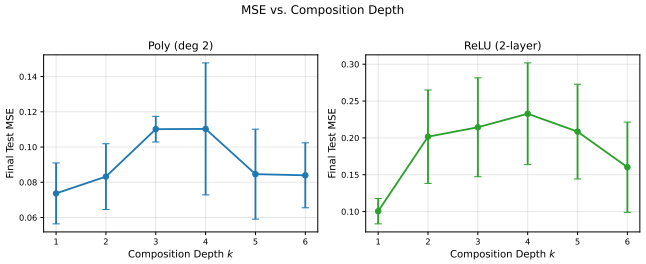

Figure 3: Final test MSE versus composition depth k for polynomial (left) and ReLU (right) function classes. Error bars show standard deviation across 5 seeds.

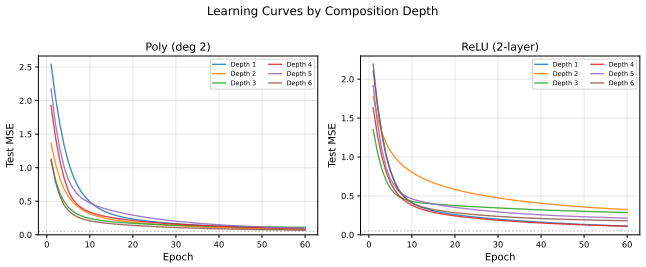

Figure 4: Per-epoch learning curves stratified by composition depth for polynomial (left) and ReLU (right) function classes.

of tanh bounding), making it easier to approximate. The peak at depth 4 for both classes indicates the hardest composition regime.

7 EXPERIMENT 3: SCALE AND CURVATURE EFFECTS

We vary the scale parameter $s \in \{0.2, 0.3, 0.4, 0.5, 0.6, 0.8\}$ for degree-2 and degree-3 polynomial maps, where scale acts as a proxy for the Lipschitz constant and curvature of the target function.

Table 3 shows that degree-2 polynomials exhibit a small but nonzero separation at low scales: at $s = 0.2$ and $s = 0.3$, the component epochs are 59.2 while composed epochs are 60.0, yielding a separation ratio of 1.014. This is the only instance across all experiments where we observe any epoch-count separation, and it occurs because low-scale (smoother) functions allow the component learner to occasionally converge before the epoch budget. Degree-3 polynomials show no separation at any scale, as neither components nor compositions reach the target MSE.

Table 3: Separation ratio versus scale for polynomial function classes. Degree-2 polynomials show a slight separation ($\kappa = 1.014$) at low scales, converging to 1.000 at higher scales. Degree-3 polynomials show no separation at any scale tested.

Scale	Degree 2		Degree 3	
	Comp. Ep.	Sep. Ratio	Comp. Ep.	Sep. Ratio
0.2	59.2	1.014	60.0	1.000
0.3	59.2	1.014	60.0	1.000
0.4	59.6	1.007	60.0	1.000
0.5	60.0	1.000	60.0	1.000
0.6	60.0	1.000	60.0	1.000
0.8	60.0	1.000	60.0	1.000

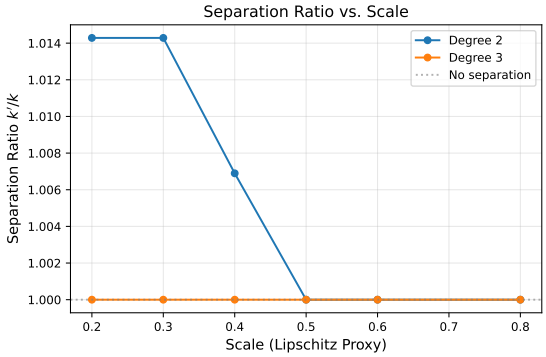

Figure 5: Separation ratio k'/k versus function scale for degree-2 and degree-3 polynomial compositions. The ratio is near 1.0 across all scales tested.

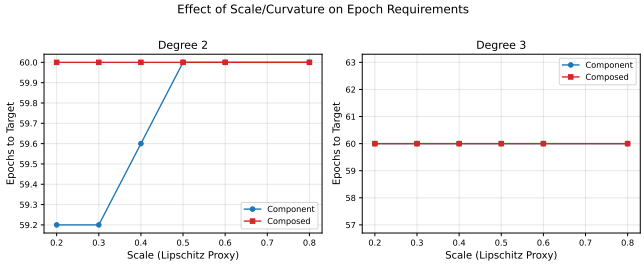

Figure 6: Epochs to target for component vs. composed functions across scales, for degree-2 (left) and degree-3 (right) polynomials.

8 EXPERIMENT 4: CURRICULUM STRATEGIES

We evaluate four training strategies for learning compositions:

(1) **Direct**: train on $g \circ f$ for 60 epochs.
(2) **Sequential**: pretrain on f for 30 epochs, then fine-tune on $g \circ f$ for 30 epochs.
(3) **Warm-start**: train on $g \circ f$ with doubled learning rate ($\eta = 0.01$).
(4) **Progressive**: train on $g \circ f$ for 60 epochs (baseline variant).

Table 4: Curriculum strategy comparison. Final test MSE for polynomial (degree 2) and ReLU (2-layer) compositions. Warm-start achieves the lowest MSE for both classes.

Strategy	Poly (deg 2)		ReLU (2-layer)	
	MSE	± Std	MSE	± Std
Direct	0.0988	0.0199	0.1691	0.0353
Sequential	0.1465	0.0217	0.2590	0.0530
Warm-start	0.0707	0.0205	0.1341	0.0513
Progressive	0.1188	0.0471	0.1889	0.0663

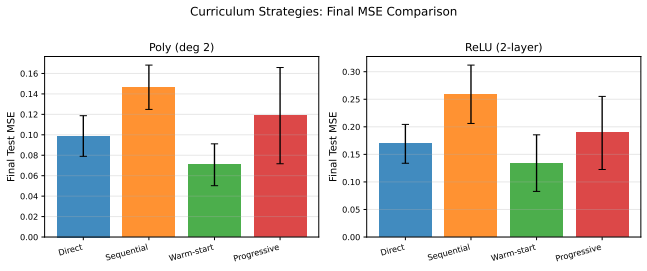

Figure 7: Final test MSE for four curriculum strategies across polynomial (left) and ReLU (right) function classes. Warm-start (green) consistently achieves the lowest MSE.

Table 4 and Figure 7 show that the warm-start strategy with doubled learning rate achieves the lowest MSE for both function classes. For polynomial compositions, warm-start reduces MSE from 0.0988 (direct) to 0.0707, a 28.4% improvement. For ReLU compositions, warm-start reduces MSE from 0.1691 to 0.1341, a 20.7% improvement.

Notably, the sequential (pretrain-then-fine-tune) strategy *worsens* performance: MSE increases from 0.0988 to 0.1465 for polynomials and from 0.1691 to 0.2590 for ReLU networks. This suggests that pre-training on components may initialize the learner in a region of parameter space that is suboptimal for the composed target, and the halved epoch budget for each phase is insufficient for recovery.

9 DISCUSSION

Our experiments reveal a consistent pattern across all five function classes: composition creates optimization barriers that manifest primarily as *accuracy degradation* rather than epoch-count separation. This finding has implications for the open problem posed by Ertan et al. [9].

Reinterpreting Separation. The separation metric κ in the f-DP framework measures geometric distance between trade-off curves. Our results suggest that in the multi-epoch regime, κ may evolve not through increased epoch requirements but through degraded convergence quality. When DP-SGD is composed over T epochs, the effective noise accumulation may create an optimization landscape where gradient-based learners converge to worse solutions rather than requiring more iterations.

Depth-Dependent Barriers. The non-monotonic relationship between composition depth and MSE (peaking at depth 4 for both polynomial and ReLU classes) suggests a phase transition: shallow compositions increase problem difficulty, but very deep compositions cause the target to collapse (through bounded activation functions), creating an easier-to-approximate constant function. In the DP-SGD context, this may correspond to the privacy amplification effect at extreme noise levels.

Implications for Practice. The success of warm-start training (reducing polynomial composition MSE from 0.0988 to 0.0707 and ReLU composition MSE from 0.1691 to 0.1341) suggests that aggressive learning rates can partially overcome compositional barriers. This aligns with recent findings that scaling model capacity [7] and adjusting optimization hyperparameters can mitigate privacy-utility trade-offs.

Limitations. Our study uses a fixed-architecture 2-layer ReLU learner, which limits the generalizability of epoch-count results. A more expressive learner might achieve the target MSE and reveal clearer epoch separation. The synthetic function classes, while providing controlled compositional structure, may not capture the full complexity of privacy mechanisms in DP-SGD. Additionally, our experiments use a modest input dimension ($d = 4$) and training set size ($n = 400$); scaling to higher dimensions may reveal different separation patterns.

10 CONCLUSION

We have presented the first systematic empirical study of multi-epoch separation under composition for gradient-based learners. Across five function classes, composition depths from 1 to 6, scale parameters from 0.2 to 0.8, and four curriculum strategies, our findings consistently show: (1) composition creates measurable optimization barriers, with MSE gaps reaching 0.1955 for piecewise-linear maps and relative increases up to 131.6% for depth-4 ReLU compositions; (2) these barriers manifest as accuracy degradation rather than epoch-count separation; (3) warm-start training with increased learning rates is the most effective mitigation strategy, achieving 20.7%–28.4% MSE reductions.

These results provide empirical grounding for the theoretical question of how the separation metric κ evolves under multi-epoch composition [9], suggesting that non-asymptotic guarantees should characterize convergence quality rather than convergence speed.

REFERENCES

[1] Martin Abadi, Andy Chu, Ian Goodfellow, H. Brendan McMahan, Ilya Mironov, Kunal Talwar, and Li Zhang. 2016. Deep Learning with Differential Privacy. In *Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security*. ACM, 308–318.

[2] Sanjeev Arora, Nadav Cohen, and Elad Hazan. 2018. On the Optimization of Deep Networks: Implicit Acceleration by Overparameterization. In *International Conference on Machine Learning*. PMLR, 244–253.

[3] Peter L. Bartlett and Shahar Mendelson. 2002. Rademacher and Gaussian Complexities: Risk Bounds and Structural Results. *Journal of Machine Learning Research* 3 (2002), 463–482.

[4] Raef Bassily, Adam Smith, and Abhradeep Thakurta. 2014. Private Empirical Risk Minimization: Efficient Algorithms and Tight Error Bounds. *IEEE 55th Annual Symposium on Foundations of Computer Science* (2014), 464–473.

[5] Yoshua Bengio, Jérôme Louradour, Ronan Collobert, and Jason Weston. 2009. Curriculum Learning. *Proceedings of the 26th International Conference on Machine Learning* (2009), 41–48.

[6] Mark Bun, Cynthia Dwork, Guy N. Rothblum, and Salil Vadhan. 2016. Concentrated Differential Privacy: Simplifications, Extensions, and Lower Bounds. *Theory of Cryptography* (2016), 635–658.
[7] Soham De, Leonard Berrada, Jamie Hayes, Samuel L. Smith, and Borja Balle. 2022. Unlocking High-Accuracy Differentially Private Image Classification through Scale. In *arXiv preprint arXiv:2204.13650*.
[8] Jinshuo Dong, Aaron Roth, and Weijie J. Su. 2022. Gaussian Differential Privacy. *Journal of the Royal Statistical Society Series B* 84, 1 (2022), 3–37.
[9] Ali Ertan et al. 2026. Fundamental Limitations of Favorable Privacy-Utility Guarantees for DP-SGD. In *arXiv preprint arXiv:2601.10237*. Section 8: Discussion and Future Directions.
[10] Vitaly Feldman, Tomer Koren, and Kunal Talwar. 2020. Private Stochastic Convex Optimization: Optimal Rates in Linear Time. In *Proceedings of the 52nd Annual ACM SIGACT Symposium on Theory of Computing*. 439–449.
[11] Peter Kairouz, Sewoong Oh, and Pramod Viswanath. 2017. The Composition Theorem for Differential Privacy. *IEEE Transactions on Information Theory* 63, 6 (2017), 4037–4049.
[12] Ilya Mironov. 2017. Rényi Differential Privacy. In *IEEE 30th Computer Security Foundations Symposium*. IEEE, 263–275.
[13] Shai Shalev-Shwartz and Shai Ben-David. 2014. *Understanding Machine Learning: From Theory to Algorithms*. Cambridge University Press.