# Adapting to High Dimensional Concepts with Metalearning

Author Names Withheld

Under Review for the Workshop on High-dimensional Learning Dynamics, ICML 2025

**Abstract**

Rapidly learning abstract rules from limited examples is a hallmark of human intelligence. This work investigates whether gradient-based meta-learning can equip neural networks with inductive biases for efficient few-shot acquisition of compositional Boolean concepts. We compare meta-learning strategies against a supervised learning baseline on Boolean tasks generated by a probabilistic context-free grammar, varying concept complexity and input dimensionality. Using a consistent multilayer perceptron (MLP) architecture, we evaluate performance based on final validation accuracy and learning efficiency. Our findings indicate that meta-learning, particularly when allowed more adaptation steps, offers significant advantages in data efficiency and final performance on lower-dimensional tasks. However, all methods face challenges as input dimensionality and concept complexity increase, highlighting the intricate interplay between learning strategies, task structure, and data representation in high-dimensional settings.

## 1 Introduction

Humans excel at few-shot concept learning, inferring abstract logical rules from minimal evidence [13]. Meta-learning algorithms, such as Model-Agnostic Meta-Learning (MAML; 5) and its derivatives, aim to instill similar capabilities in neural networks by learning an initialization optimized for rapid adaptation. While effective in perceptual domains, their efficacy on structured, rule-based problems, especially those involving symbolic composition and high-dimensional inputs, remains an active research area. This study addresses this gap by asking: *How do meta-learning strategies and standard supervised learning compare in acquiring PCFG-generated Boolean concepts as a function of concept complexity?* We analyze learning dynamics through classification performance and data efficiency, contributing to the understanding of meta-learning dynamics in high-dimensional concept spaces.

## 2 Related Work

**Gradient-based meta-learning.** MAML [5] introduced a framework for learning model initializations that adapt quickly via gradient descent. Meta-SGD [14] extends this by learning per-parameter step sizes, enabling one-step adaptation. First-order approximations such as FOMAML and Reptile [15] omit Hessian terms to reduce cost, yet their performance often matches full MAML on vision tasks. Theoretical analyses highlight that second-order updates embed an implicit contrastive objective, which can improve generalization on harder tasks [10]. Recent work has explored scaling meta-learning beyond classification, including reinforcement learning [16] and program induction [4].

**Compositional generalization and concept learning.** Symbolic rule induction methods, such as Bayesian Program Learning (BPL) [13] and the Rational Rules model [6], achieve human-level one-shot learning by leveraging explicit grammars. However, they require handcrafted generative models and search. Neural sequence-to-sequence models struggle with systematic generalization on tasks like SCAN [12], and neural meta-learners underperform on benchmarks like CURI [21]. Meta-learning has recently been used to improve compositional generalization in NLP [7] and neuro-symbolic reasoning systems [23], but its role in Boolean concept induction remains unclear. A theoretical framework for compositional generalization

in neural networks was recently proposed [1], and surveys highlight the challenges and opportunities for compositional AI [19]. We study this in a controlled Boolean PCFG setting to isolate logical structure.

# 3 Experimental Setup

Our experimental setup starts with a concept-generating Probabilistic Context-Free Grammar (PCFG) from Goodman et al. 2008 [6]. We modify the PCFG to explicitly control concept complexity via recursion depth ($D \in \{3, 5, 7\}$) and feature dimensionality (the number of literals $F \in \{8, 16, 32\}$). The grammar production rule is defined recursively : $C \to L \,|\, \neg C \,|\, (C \wedge C) \,|\, (C \vee C)$; where $L \to x_i$, over $F$ binary features $\mathcal{X} = \{x_1, \ldots, x_F\}$. For each task $C$, a $K_{shot}$ support set $S_C$ ($K_{shot} = 5$ positive and $K_{shot} = 5$ negative examples $(\mathbf{x}, C(\mathbf{x}))$) and a query set $Q_C$ are generated from uniformly sampled $\mathbf{x} \in \{0, 1\}^F$.

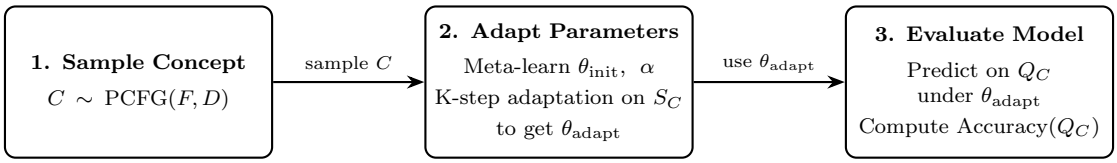


Figure 1: The three-stage meta-learning pipeline. 1. A Boolean concept $C$ is sampled from a PCFG (features $F$, depth $D$). 2. Meta-learned initial parameters $\theta_{init}$ and per-parameter learning rates $\alpha$ are adapted on a support set $S_C$ for $K_{adapt}$ gradient steps to yield $\theta_{adapt}$. 3. The adapted model $\theta_{adapt}$ is evaluated on a query set $Q_C$.

All methods use a 5-layer MLP (128 hidden units/layer, ReLU, sigmoid output). We compare models trained with four stochastic gradient descent (SGD) learning algorithms, varying the order of the gradients and adaptation steps: 1st-Order and 2nd-Order Meta-SGD with 1 adaptation (gradient) step, 1st-Order Meta-SGD with 10 adaptation steps, and regular SGD: training from scratch per task using Adam [11] (LR 0.001) on $S_C$.

During adaptation for Meta-SGD variants, the parameters $\theta$ are updated for $K_{adapt}$ steps. Starting with $\theta^{(0)} = \theta_{init}$ (the meta-learned initialization), for each step $k \in \{0, \ldots, K_{adapt} - 1\}$, the update is:

$$\theta^{(k+1)} = \theta^{(k)} - \alpha \odot \nabla_{\theta^{(k)}} \mathcal{L}_{S_C}(\theta^{(k)})$$

where $\alpha$ are the meta-learned per-parameter learning rates and $\mathcal{L}_{S_C}$ is the loss (binary cross-entropy) on the support set $S_C$. The final adapted parameters are $\theta_{adapt} = \theta^{(K_{adapt})}$. Increasing $K_{adapt}$ allows the model to perform a more extensive search in the task-specific loss landscape, potentially finding a solution that better minimizes the empirical risk on $S_C$. For Boolean concepts, this translates to a more refined adjustment of the MLP's decision boundaries to correctly classify the examples in the support set.

Meta-SGD models were meta-trained for 10,000 episodes. All evaluations were averaged over 5 random seeds on 1,000 unseen tasks. For trajectory comparisons, SGD is trained for steps equivalent to processing a fixed total number of samples.

Performance is assessed using: **1. Final Mean Accuracy** (see Appendix A.1 for summary bar chart, Figure 4); and **2. Data Efficiency**: The number of training samples (episodes for MetaSGD, scaled appropriately for SGD) required to reach a threshold accuracy of 60% (Figure 3).

# 4 Results

We present findings across varying feature dimensionalities ($F$) and concept depths ($D$), averaged over 5 seeds. Figure 2 shows the learning trajectories. meta-SGD methods demonstrate a clear advantage over SGD from scratch, learning faster and converging to higher accuracies, particularly for $F = 8$ and $F = 16$. Surprisingly, increasing the adaptation steps in the first-order meta-SGD variant matches or exceeds the performance of 2nd order variant, especially in terms of learning speed. Final noise-averaged accuracies are summarized in Appendix A.1 (Figure 4), confirming these trends.

The benefit of meta-learning and increased adaptation steps is further highlighted by the data efficiency plot (Figure 3). MetaSGD methods require substantially fewer samples to reach the 60% accuracy threshold

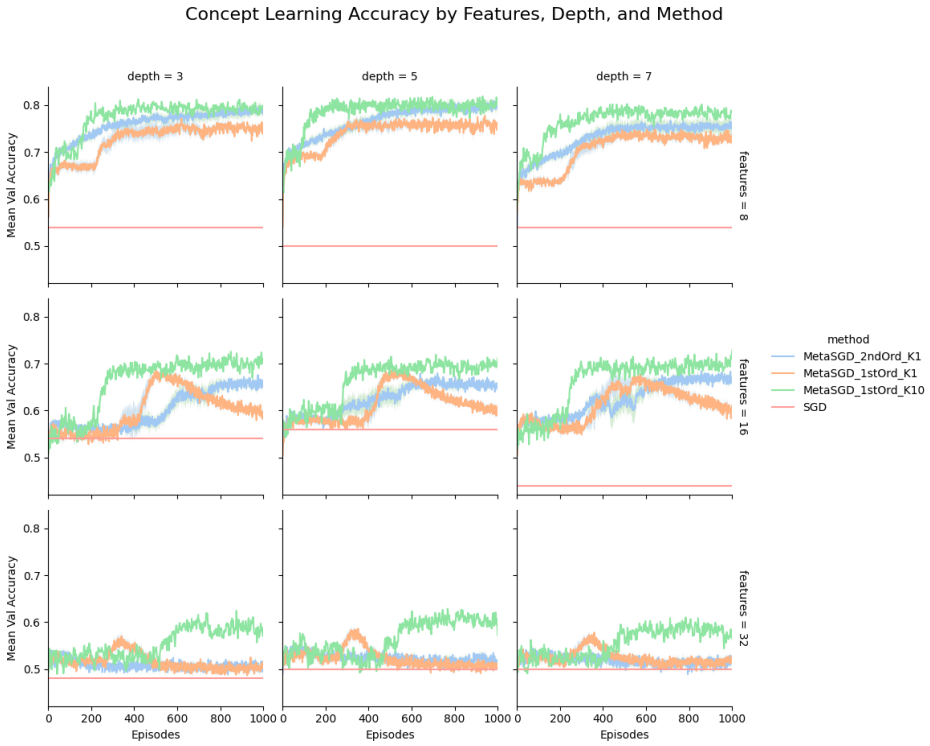


Figure 2: Mean Validation Accuracy Trajectories. Comparison of Meta-SGD variants (K1 and K10 for 1st-Order, K1 for 2nd-Order) and SGD across features (rows) and concept depths (columns) over normalized training episodes. `MetaSGD_1stOrd_K10` often learns fastest and achieves competitive or superior accuracy to `MetaSGD_2ndOrd_K1`.

compared to SGD. 1st order meta-SGD with increased adaption is often the most data-efficient, demonstrating that more adaptation steps enable the model to more rapidly specialize to the task at hand using the limited support set. For instance, at $F = 8, D = 3$, `MetaSGD_1stOrd_K10` reaches the threshold with orders of magnitude fewer samples than SGD.

As input dimensionality increases to $F = 32$, all methods exhibit a significant drop in performance across all metrics. While Meta-SGD variants still maintain an edge over SGD, the absolute accuracies are much lower, and the data efficiency gains are less pronounced relative to the sheer number of samples still required. Notably, even in this high-dimensional regime ($F = 32$), the increased adaptation with first-order strategy tends to yield the largest relative performance improvements over its 1-step counterparts (both 1st and 2nd order Meta-SGD), suggesting that the benefit of more extensive adaptation becomes even more crucial when the effective concept space is larger.

# 5 Discussion

Our experiments demonstrate that meta-learning confers both accuracy and efficiency benefits for few-shot Boolean concept learning when feature dimensionality is moderate ($F \leq 16$). Meta-learning, via Meta-SGD, generally offers substantial advantages over standard supervised training (SGD from scratch) in both test accuracy (Figure 2 and Appendix A.1) and data efficiency (Figure 3). This suggests that providing the meta-learned initialization $(\theta_{init}, \alpha)$ with more opportunity to fine-tune to the specific task $S_C$ (via $K_{adapt} = 10$ steps) is highly effective for Boolean concept learning. In particular we hint at a tradeoff in metalearning dynamics between multi-step adaptation and curvature information. First-order Meta-SGD with $K_{adapt} = 10$ matches or exceeds the performance of second-order Meta-SGD with a single step, indicating that additional adaptation iterations can often substitute for curvature corrections in navigating complex loss landscapes, in line with [10, 15]. We observe that even with meta-learned priors, neural models struggle to generalize compositionally at scale. This aligns with findings in natural language SCAN [12] and visual concept

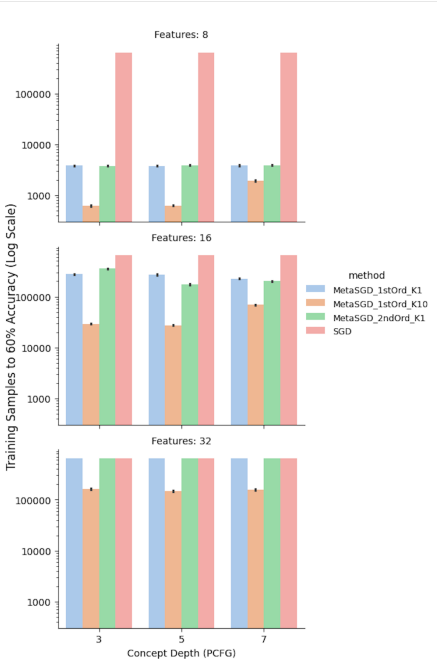


Figure 3: Data Efficiency: Training Samples to Reach 60% Validation Accuracy (Log Scale). Lower values indicate higher efficiency. MetaSGD methods, especially with increased adaptation steps, ie `MetaSGD_1stOrd_K10`, are significantly more data-efficient than SGD.

benchmarks [21]. The recent theoretical analysis suggests that achieving systematic generalization requires aligning internal representations with the combinatorial structure of tasks [1]. Furthermore, as we scale featural complexity, performance collapses at $F = 32$, reflecting the curse of dimensionality under extreme data sparsity (the search space explodes to $2^{32}$ possible inputs). This echoes interesting recent results in large-scale in-context concept learning for LLMs, where performance correlates strongly with Boolean complexity [22].

Pronounced difficulties of bias-free neural networks such as MLPs in generalizing within high-dimensional concept spaces underscore the imperative for additional inductive biases. While weight-space initialization via meta-learning provides a powerful prior even for small MLPs, future work could explore relational architectures—such as Relation Networks [18] and Graph Neural Networks that embed relational inductive biases [2]—or attention-based models like Transformers [20] to capture feature interactions compositionally. Likewise, contrastive learning objectives have proven effective at structuring representation spaces for fewer-shot separability [3, 8], and energy-based or regularized losses (e.g. 9) may further scaffold abstraction. Finally, neuro-symbolic hybrids and differentiable logic modules [4, 17] offer a route to inject explicit symbolic priors, potentially combining the strengths of meta-learned initialization with symbolic compositionality.

# 6 Conclusion

Building on our empirical findings, we conclude that gradient-based meta-learning—especially when endowed with multiple adaptation steps—dramatically accelerates and improves few-shot Boolean concept induction in moderate-dimensional settings. Yet, the persistent collapse of performance as feature dimensionality and combinatorial complexity grow exposes the limits of weight-initialization priors alone. To bridge this gap, future work must weave in richer inductive biases—ranging from relational and graph-structured architectures to attention mechanisms, contrastive objectives, and neurosymbolic modules—that explicitly mirror the compositional structure of concepts. Such synergies between meta-learned initialization and structured priors promise to push neural systems closer to the flexibility and efficiency of human concept learning in

high-dimensional, sparse regimes.

# References


[1] Sanjeev Arora, Ying Li, and Abhishek Panigrahi. Toward a theoretical understanding of compositional generalization in neural networks. *Journal of Machine Learning Research*, 25(1):1–39, 2024.

[2] Peter W. Battaglia, Jessica B. Hamrick, Victor Bapst, Alvaro Sanchez-Gonzalez, Vinicius Zambaldi, Mateusz Malinowski, Andrea Tacchetti, David Raposo, Adam Santoro, Ryan Faulkner, Caglar Gulcehre, Fang Song, Andrew Ballard, Justin Gilmer, George E. Dahl, Vijay Vasudevan, Adrian Plill, Razvan Pascanu, Charles Blundell, Koray Kavukcuoglu, Demis Hassabis, Pushmeet Kohli, and Matthew Botvinick. Relational inductive biases, deep learning, and graph networks. *arXiv preprint arXiv:1806.01261*, 2018.

[3] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *International Conference on Machine Learning (ICML)*, 2020.

[4] Kevin Ellis, Simon Nax, Daniel Tarlow, Tim Plotz, Ethan Perot, and Pushmeet Kohli. Learning to compose programs for image recognition. In *International Conference on Learning Representations (ICLR)*, 2018.

[5] Chelsea Finn, Pieter Abbeel, and Sergey Levine. Model-agnostic meta-learning for fast adaptation of deep networks. In *Proceedings of the 34th International Conference on Machine Learning (ICML)*, pages 1126–1135, 2017.

[6] Noah D. Goodman, Joshua B. Tenenbaum, Jacob Feldman, and Thomas L. Griffiths. A rational analysis of rule-based concept learning. In *Cognitive Science Society Conference (CogSci)*, pages 121–126, 2008.

[7] Xinyun Guo and David Chiang. Meta-learning to generalize to compositional tasks. In *Proceedings of ACL*, 2021.

[8] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020.

[9] Geoffrey E. Hinton. Training products of experts by minimizing contrastive divergence. *Neural Computation*, 14(8):1771–1800, 2002.

[10] Chia-Hsiang Kao, Wei-Chen Chiu, and Pin-Yu Chen. Maml is a noisy contrastive learner in classification. In *International Conference on Learning Representations (ICLR)*, 2022.

[11] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

[12] Brenden M. Lake and Marco Baroni. Generalization without systematicity: On the compositional skills of sequence-to-sequence recurrent networks. In *Proceedings of the 35th International Conference on Machine Learning (ICML)*, pages 2873–2882, 2018.

[13] Brenden M. Lake, Ruslan Salakhutdinov, and Joshua B. Tenenbaum. Human-level concept learning through probabilistic program induction. *Science*, 350(6266):1332–1338, 2015.

[14] Zhenguo Li, Fengwei Zhou, Fei Chen, and Hang Li. Meta-sgd: Learning to learn quickly for few-shot learning. *arXiv preprint arXiv:1707.09835*, 2017.

[15] Alex Nichol, Joshua Achiam, and John Schulman. On first-order meta-learning algorithms. *arXiv preprint arXiv:1803.02999*, 2018.

[16] Arun Rajeswaran, Chelsea Finn, Sham Kakade, and Sergey Levine. Meta-learning with implicit gradients. *NeurIPS*, 2019.

[17] Tim Rocktäschel and Sebastian Riedel. End-to-end differentiable proving. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2017.

[18] Adam Santoro, David Raposo, David G. T. Barrett, Mateusz Malinowski, Razvan Pascanu, Peter Battaglia, and Timothy Lillicrap. A simple neural network module for relational reasoning. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2017.

[19] Yiding Shen, Keith Bonawitz, and Michael Lewis. Compositionality in artificial intelligence: A survey. *ACM Computing Surveys*, 56(2):1–35, 2024.

[20] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2017.

[21] Ramakrishna Vedantam, Ari Morcos, and et al. Curi: A benchmark for productive concept learning under uncertainty. *arXiv preprint arXiv:2301.XXXXX*, 2023.

[22] Leroy Z. Wang, R. Thomas McCoy, and Shane Steinert-Threlkeld. Minimization of boolean complexity in in-context concept learning. *arXiv preprint arXiv:2412.02823*, 2024. URL `https://arxiv.org/abs/2412.02823`.

[23] Dingding Ye, Bin Zhou, Shih-Chieh Chang, and Jian Chen. Nemesys: Neural meta–symbolic synergy for compositional reasoning. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 35, pages 10000–10012, 2022.

# A Appendix

## A.1 Final Mean Validation Accuracy (Bar Chart)

This plot (Figure 4) complements the trajectory data in Figure 2 by providing a direct comparison of final performance levels across the different learning methods and task configurations.

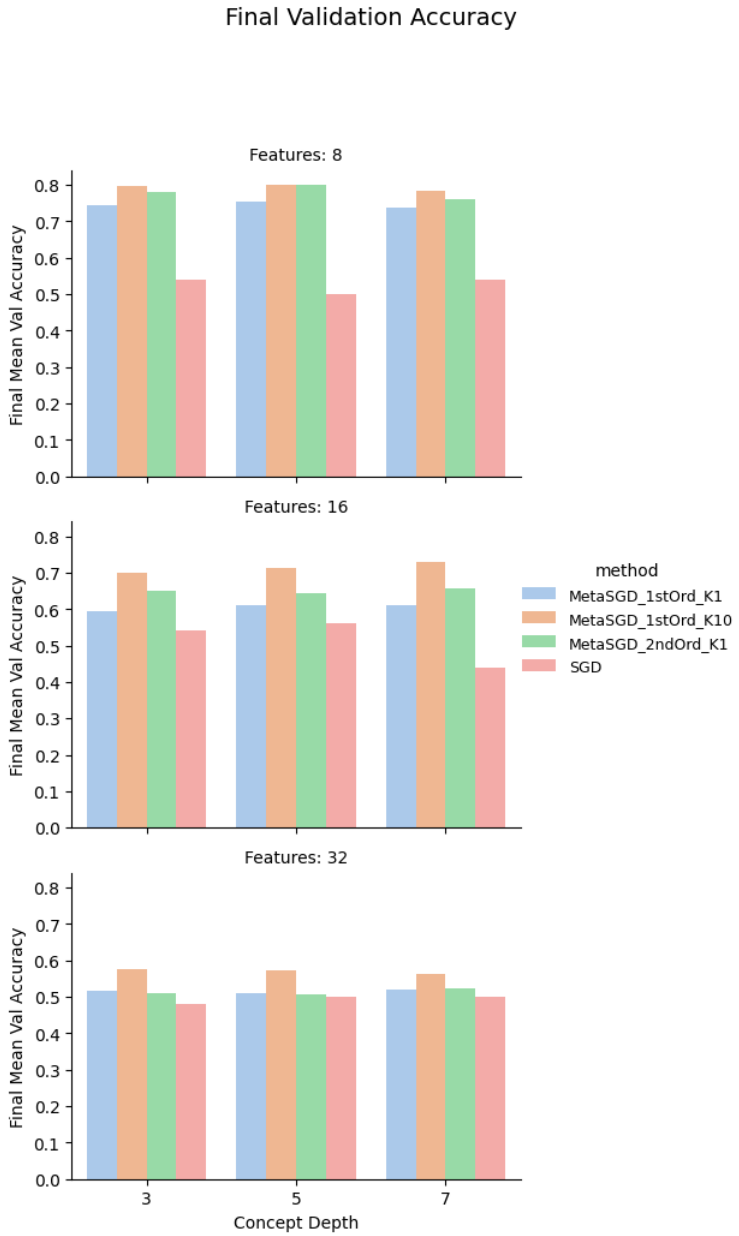


Figure 4: Final Mean Validation Accuracy (Bar Chart). Comparison of Meta-SGD variants and supervised SGD across different feature dimensionalities (rows/facets) and concept depths (x-axis).

## A.2 Layer-wise L2 Norm Comparison of Model Weights

Figure 5 presents a visual comparison of the average L2 norms of weights for each layer in the MLP architecture. The comparison is made between 1st-Order and 2nd-Order Meta-SGD methods. The plots are faceted by the number of input features (rows: 8, 16, 32) and the concept depth used in the filename for model

generation (columns: 3, 5, 7). This visualization allows for an examination of how weight magnitudes differ across layers, learning methods, and task configurations (feature dimensionality and concept complexity).

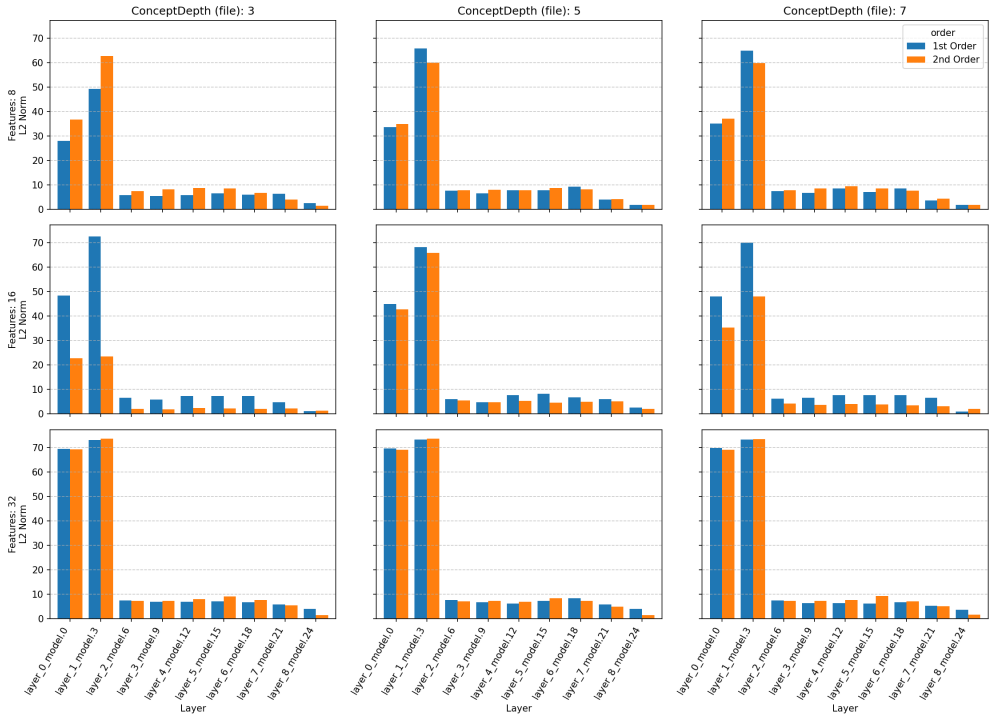


Figure 5: Average L2 Norm of MLP Layer Weights. Comparison of 1st-Order Meta-SGD and 2nd-Order Meta-SGD, faceted by input features (rows) and concept depth from filename (columns). Each bar group represents a layer within the MLP. Norms are averaged over available seeds for each configuration. The x-axis labels indicate the layer index.