Crafting Global Optimizers to Reasoning Tasks via Algebraic Objects in Neural Nets

Anonymous Author(s)

Abstract

We prove rich algebraic structures of the solution space for 2-layer neural networks with quadratic activation and L_2 loss, trained on reasoning tasks in Abelian group (e.g., modular addition). Such a rich structure enables us to *analytically* construct the global optimal solutions to the task from partial solutions that only satisfy part of the loss, despite its high nonlinearity. Specifically, we show that the union-ed solution space of different number of hidden nodes of the 2-layer network is endowed with a semi-ring algebraic structure, and the loss function to be optimized consists of *monomial potentials* which are ring homomorphism, allowing composition of partial solutions by ring addition and multiplication. While the constructed global optimizers only require small number of hidden nodes, we show that overparameterization asymptotically decouples the training dynamics and thus is beneficial. We further show that training dynamics move towards simpler solutions under regularization, by proving that global optimizers algebraically connected by ring multiplication are also topologically connected. Experiments verify our theoretical findings.

1 Introduction

Large Language Models (LLMs) have shown impressive results in various disciplines [18, 1, 22, 4, 5, 11], while they also make surprising mistakes in basic reasoning tasks [17, 2]. Therefore, it remains an open problem whether it can truly do reasoning tasks. On one hand, existing works demonstrate that the models can learn efficient algorithm (e.g., dynamic programming [27] for language structure modeling, gradient descent [24] for linear regressions, etc) and good representations [12]. Some reports emergent behaviors [25] when scaling up with data and model size. On the other hand, many works also show that LLMs cannot self-correct [9], and cannot generalize very well beyond the training set for simple tasks [6, 28, 19], let alone complicated planning [13, 26].

To understand how the model performs reasoning and further improve its reasoning power, people have been studying simple arithmetic reasoning problems in depth. Modular addition [16, 29], i.e., predicting $a + b \mod d$ given a and b, is a popular one due to its simple and intuitive structure yet surprising behaviors in learning dynamics (e.g., grokking [20]) and learned representations (e.g., Fourier bases [30]). Most works focus on various metrics to measure the behaviors and extracting interpretable circuits from trained models [16, 23, 10]. Analytic solutions can be constructed and/or reverse-engineered [8, 29, 16] but it is not clear how to generalize the results.

In this work, we systematically analyze 2-layer neural networks with quadratic activation and L_2 loss on predicting group multiplication in Abelian group G, which is an extension of modular addition. We find that global optimizers can be constructed *algebraically* from small partial solutions that are optimal only for parts of the loss. We achieve this by showing that (1) for the 2-layer network, there exists a ***semi-ring*** structure over the set of solutions *across different order* (i.e., number of hidden nodes or network width), with specifically defined addition and multiplication (Def. 3), and (2) the

Submitted to NeurIPS 2024 Workshop on Compositional Learning: Perspectives, Methods, and Paths Forward. Do not distribute.

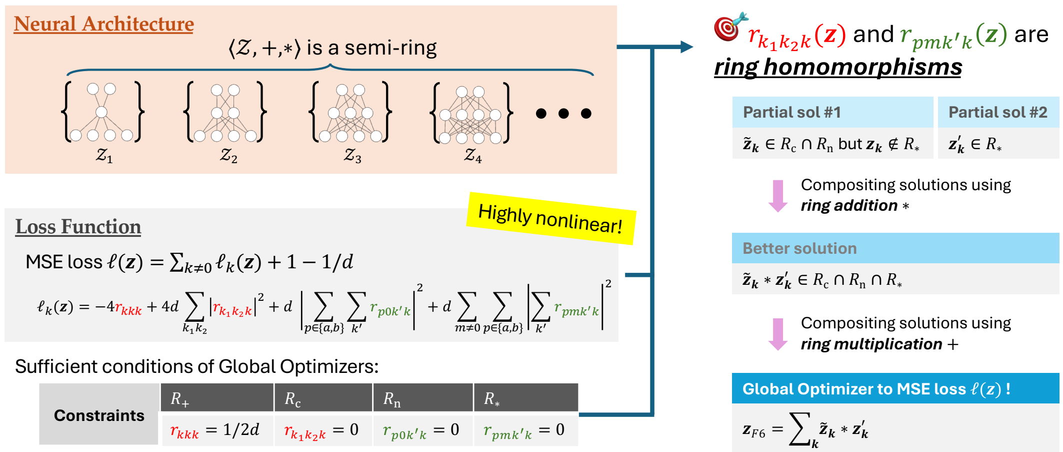

Figure 1: Overview of proposed theoretical framework CaGO. (1) The family of 1-hidden layer neural networks, $\mathcal{Z}$, form a *semi-ring* algebraic structure ring addition and multiplication (Theorem 2). $\mathcal{Z} = \bigcup_{q\geq 0} \mathcal{Z}_q$ where $\mathcal{Z}_q$ is a collection of all weights (solutions) with order-q (i.e., q hidden nodes). (2) For Abelian reasoning task, the MSE loss $\ell(\boldsymbol{z})$ is a function of *monomial potentials* (MPs) $r_{k_1k_2k}(\boldsymbol{z})$ and $r_{pmk'k}(\boldsymbol{z})$ (Theorem 1), which are ring homomorphism (Theorem 3). (3) Thanks to the property of ring homomorphism, global optimizers to MSE loss $\ell(\boldsymbol{z})$ with quadratic activation can be constructed *algebraically* from partial solutions that only satisfy a subset of constraints (Sec. A.1) using ring addition and multiplication, instead of running gradient descent. Examples include Fourier solution $\boldsymbol{z}_{F6}$ (Corollary 2) and perfect memorization solution $\boldsymbol{z}_M$ (Corollary 4). In Sec. B, we analyze the role played of MPs in gradient dynamics, showing that the dynamics favors low-order global optimizers (Theorem 5) under weight decay regularization, and the dynamics of MPs become decoupled with infinite width (Theorem 6).

L_2 loss is a function of ***monomial potentials*** (MPs), which are ring homomorphisms (Theorem 1) that allow compositions of partial solutions into global ones using ring addition and multiplication.

As a result, our theoretical framework, named CaGO (i.e., *Crafting Global Optimizers*), successfully constructs two distinct types of Fourier-based solutions of per-frequency order 4 ($= 2 \times 2$) and order 6 ($= 2 \times 3$) that is global optimal, which are verified in the experiments, and global optimal solutions of order d^2 that correspond to perfect memorization. To our best knowledge, we are the first to discover such algebraic structures inside network training, and apply it to analyze solutions to reasoning tasks such as modular additions in details.

In addition, we also analyze the training dynamics of MPs. We show that the dynamics favors low-order solutions and perfect memorization is unfavorable in the dynamics, and the MP dynamics becomes decoupled when the network width goes to infinite, demystifying why overparameterization improves the performance.

Most Related work. Existing theoretical work [15] also shows group-theoretical results on algebraic tasks related to finite groups, also for networks with one-hidden layers and quadratic activations. However, they use the max-margin framework with a special regularization ($L_{2,3}$ norm) rather than MSE loss, do not characterize and leverage algebraic structures to construct solutions, and do not analyze the training dynamics.

2 Decoupling L_2 Loss in reasoning tasks of Abelian group

Problem Setup. We consider the following 2-layer networks with one layer of hidden nodes, trained with (projected) ℓ_2 loss on prediction of group multiplication in Abelian group G with $|G| = d$:

$$\ell = \sum_i \|P_1^\perp(\boldsymbol{o}[i] - l[i])\|^2, \qquad \boldsymbol{o}[i] = V\sigma(W^\top \boldsymbol{f}[i]) = \sum_j \boldsymbol{v}_j \sigma(\mathbf{w}_j^\top \boldsymbol{f}[i]) \tag{1}$$

where $\sigma(x) = x^2$ is the quadratic activation function, $P_1^\perp = I - \frac{1}{d}\mathbf{1}\mathbf{1}^\top$ is the zero-mean projection matrix, $W = [\mathbf{w}_1, \ldots, \mathbf{w}_q] \in \mathbb{R}^{d\times q}$, $V = [\boldsymbol{v}_1, \ldots, \boldsymbol{v}_q]^\top \in \mathbb{R}^{d\times q}$ are learnable parameters. $\boldsymbol{f}[i] \in \mathbb{R}^d$ are input embeddings. i is the sample index.

Input and Output. The input contains the two group elements $g_1[i]$ and $g_2[i]$, encoded as $\boldsymbol{f}[i] = U_{G_1}\boldsymbol{e}_{g_1[i]} + U_{G_2}\boldsymbol{e}_{g_2[i]}$, where U_{G_1} and U_{G_2} are column orthogonal embedding matrices. The output is the result $g_1[i]g_2[i] \in G$, encoded as the label $l[i] = g_1[i]g_2[i]$ to be predicted.

Let $\boldsymbol{\phi}_k = [\phi_k(g)]_{g\in G} \in \mathbb{C}^d$ be the scaled Fourier bases (or more formally, *character function* of the finite Abelian group G, see Appendix D). Then weight vector $\mathbf{w}_j$ and $\boldsymbol{v}_j$ can be written as:

$$\mathbf{w}_j = U_{G_1}\sum_{k\neq 0} z_{akj}\boldsymbol{\phi}_k + U_{G_2}\sum_{k\neq 0} z_{bkj}\boldsymbol{\phi}_k, \qquad \boldsymbol{v}_j = \sum_{k\neq 0} z_{ckj}\bar{\boldsymbol{\phi}}_k \tag{2}$$

where $\boldsymbol{z} := \{z_{pkj}\}$ are the complex coefficients ($p \in \{a, b, c\}$, $0 \leq k < d$ and j runs through hidden nodes). Leveraging the property of quadratic activation functions, we can write down the loss function analytically (see Appendix D):

Theorem 1 (Analytic form of L_2 loss with quadratic activation)**.** *The objective of 2-layer MLP network with quadratic activation can be written as $\ell = \sum_{k\neq 0}\ell_k + (d-1)/d$, where*

$$\ell_k = -4r_{kkk} + 4d\sum_{k_1k_2}|r_{k_1k_2k}|^2 + d\Big|\sum_{p\in\{a,b\}}\sum_{k'} r_{p0k'k}\Big|^2 + d\sum_{m\neq 0}\sum_{p\in\{a,b\}}\Big|\sum_{k'} r_{pmk'k}\Big|^2 \tag{3}$$

Here $r_{k_1k_2k} := \sum_j z_{ak_1j}z_{bk_2j}z_{ckj}$ and $r_{pmk'k} := \sum_j z_{pk'j}z_{p,m-k',j}z_{ckj}$.

Note that for cyclic group G, the frequency k is a mod-d integer. For general Abelian group which can be decomposed into direct sum of cyclic groups according to Fundamental Theorem of Finite Abelian Groups, k is a multidimensional frequency index. For convenience, we define $\phi_{-k} := \overline{\phi_k}$ as the conjugate representation of ϕ_k. The reason why $\boldsymbol{\phi}_0 \equiv 1$ is excluded is that the constant bias term has been filtered out by the top-down gradient from the loss function. Since weights are all real, the Hermitian constraints holds, i.e., $\overline{z_{ckj}} = \overline{\boldsymbol{\phi}_k^*\boldsymbol{v}_j} = \boldsymbol{\phi}_{-k}^*\boldsymbol{v}_j = z_{c,-k,j}$ (and similar for z_{akj} and z_{bkj}). Therefore, $z_{p,-k,j} = \bar{z}_{pkj}$, $r_{-k,-k,-k} = \bar{r}_{kkk}$ and ℓ is real and can be minimized.

Lemma 1 (A Sufficient Conditions of Global optimizers of Eqn. 3)**.** *If a solution $\boldsymbol{z}$ to Eqn. 3 satisfies the following, then it is a global optimizer with zero loss $\ell(\boldsymbol{z}) = 0$.*

$$r_{kkk}(\boldsymbol{z}) = \mathbb{I}(k \neq 0)/2d, \quad r_{k_1k_2k}(\boldsymbol{z}) = 0, \quad r_{pmk'k}(\boldsymbol{z}) = 0 \tag{4}$$

Lemma 1 provides a *sufficient* condition since there may exist other solutions that achieve global optimum (e.g., $\sum_{k'} r_{pmk'k} = 0$). It turns out Eqn. 4 already leads to very rich algebraic structures and we will not discuss more broader cases in this work.

3 Beyond Fixed Parameter Space: The Semi-ring structure

We define the *solution space* $\mathcal{Z}_q = \{\boldsymbol{z}\}$ to include all the weight matrices with q hidden nodes ($\mathcal{Z}_0$ means an empty network). Let $\mathcal{Z} = \bigcup_{q\geq 0}\mathcal{Z}_q$ be the solution space of all different number of hidden nodes. For clarity, we use bold symbol $\boldsymbol{z}$ to represent the collection of all its components $\{z_{pkj}\}$, and $\boldsymbol{z}_1 := \{z^{(1)}_{pkj}\}$ and $\boldsymbol{z}_2 := \{z^{(2)}_{pkj}\}$ represent two solutions.

Directly finding the global optimizers to Eqn. 4 can be a bit complicated and highly non-intuitive. Interestingly, the $\mathcal{Z}$ naturally has an algebraic (semi-ring) structure, and global optimizers can be composited from non-optimal ones that only satisfies a subset of terms of the loss! Both the Fourier bases solution and the perfect memorization solution can be represented this way.

Definition 1 (Order of $\boldsymbol{z}$)**.** *The order* $\mathrm{ord}(\boldsymbol{z})$ *of $\boldsymbol{z} \in \mathcal{Z}$ is its number of hidden nodes.*

Definition 2 (Identification of $\mathcal{Z}$)**.** *In $\mathcal{Z}$, two solutions of the same order that differ only by a permutation along hidden dimension j are considered identical.*

Note that for any two solutions $\boldsymbol{z}_1, \boldsymbol{z}_2 \in \mathcal{Z}$, we can define their operations:

Definition 3 (Addition and Multiplication in $\mathcal{Z}$)**.** *Define $\boldsymbol{z} = \boldsymbol{z}_1 + \boldsymbol{z}_2$ in which $z_{pk\cdot} :=$ $\mathrm{concat}(z^{(1)}_{pk\cdot}, z^{(2)}_{pk\cdot})$ and $\boldsymbol{z} = \boldsymbol{z}_1 * \boldsymbol{z}_2$, in which $z_{pk\cdot} := z^{(1)}_{pk\cdot} \otimes z^{(2)}_{pk\cdot}$. The addition and multiplication respect Hermitian and the identity element $\mathbf{1}$ is the 1-order solutions with $\{z_{pk0} = 1\}$.*

Note that the multiplication definition is one special case of Khatri–Rao product [14]. Although the Kronocker product and concatenation are not commutative, thanks to the identification (Def. 2), $\boldsymbol{z}_1 + \boldsymbol{z}_2 = \boldsymbol{z}_2 + \boldsymbol{z}_1$ and $\boldsymbol{z}_1 * \boldsymbol{z}_2 = \boldsymbol{z}_2 * \boldsymbol{z}_1$ and thus both operations are commutative. Then:

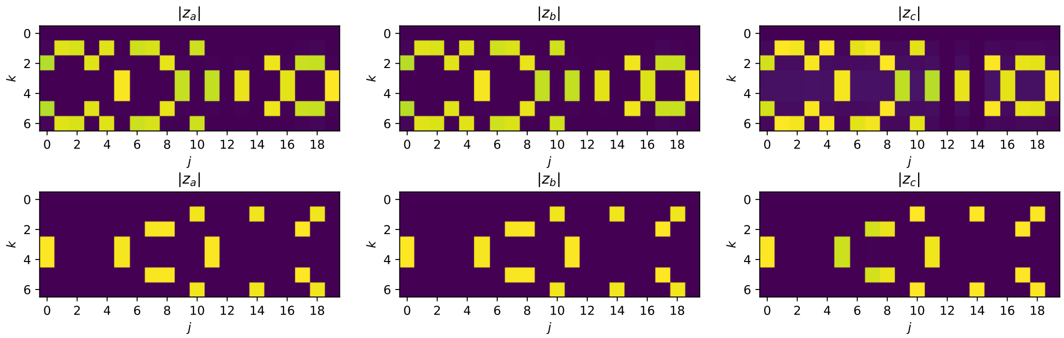

Figure 2: Solutions obtained by the Adam optimizers on ℓ_2 loss for modular addition task with $|G| = d = 7$ and $q = 20$ hidden nodes. **Top:** For each frequency $\pm k$, there are exactly 6 hidden nodes representing such a frequency, consistent with Corollary 2. **Bottom:** Optimizing Eqn. 3 without the last term $\sum_{m\neq 0}\sum_{p\in\{a,b\}}\left|\sum_{k'} r_{pmk'k}\right|^2$ (equivalently removing the constraint $R_{\circledast}$). Now each frequency has exactly 3 hidden nodes, which is also consistent with our analysis (Lemma 2).

Theorem 2 (Algebraic Structure of $\mathcal{Z}$)**.** *$\langle \mathcal{Z}, +, * \rangle$ is a commutative semi-ring.*

In the following sections, the semi-ring structure of $\mathcal{Z}$ paves the way to construct explicitly the global optimal solutions for our ℓ_2 objectives.

Now let us study the structure of the loss function Eqn. 3 and how they are related to the semi-ring structure of $\mathcal{Z}$. For this, we first define the concept of *monomial potentials*:

Definition 4 (Monomial potential (MP))**.** *Define the* monomial potential (MP) $r(\boldsymbol{z}) := \sum_j \prod_{(p,k)\in \mathrm{idx}(r)} z_{pkj}$ *where* $\mathrm{idx}(r)$ *specifies the indices involved in the monomial terms.*

Following this definition, terms in the loss function (Theorem 1) are examples of MPs.

Observation 1 (Specific MPs)**.** *$r_{k_1k_2k}(\boldsymbol{z})$ and $r_{pmk'k}(\boldsymbol{z})$ defined in Theorem 1 are MPs.*

So what is the relationship between MPs, which are parts of the loss function, and the semi-ring structure of $\mathcal{Z}$? The following theorem tells that, MPs are ring homomorphism.

Theorem 3. *For any monomial potential $r : \mathcal{Z} \mapsto \mathbb{C}$, $r(\mathbf{1}) = 1$, $r(\boldsymbol{z}_1 + \boldsymbol{z}_2) = r(\boldsymbol{z}_1) + r(\boldsymbol{z}_2)$ and $r(\boldsymbol{z}_1 * \boldsymbol{z}_2) = r(\boldsymbol{z}_1)r(\boldsymbol{z}_2)$ and thus r is a ring homomorphism.*

Observation 2. *The order function $\mathrm{ord} : \mathcal{Z} \mapsto \mathbb{N}$ is also a ring homomorphism.*

Due to the property of ring homomorphism, we immediatenly know that there exists infinitely many global minimizers, via ring multiplication (Def. 3):

Definition 5 (Unit)**.** *$\boldsymbol{z}$ is called a* unit *if $r_{kkk}(\boldsymbol{z}) = 1$ for all $k \neq 0$.*

Corollary 1. *If $\boldsymbol{z}$ is a global optimizer and $\boldsymbol{y}$ is a unit, then $\boldsymbol{z} * \boldsymbol{y}$ is also a global optimizer.*

More importantly, a global optimizer can be constructed from partial solutions that satisfy only some of the constraints. For example, if there exists $\boldsymbol{z}_1$ that satisfies constraint $r_1(\boldsymbol{z}_1) = 0$ and $\boldsymbol{z}_2$ that satisfies constraint $r_2(\boldsymbol{z}_2) = 0$, then their product $\boldsymbol{z}_1 * \boldsymbol{z}_2$ satisfy both constraints. In particular, we want such seed solutions to be small in order, so that the order of the final solutions is not too large.

4 Summary of the Appendix

In Appendix A, we show concrete solutions that are constructed following the semi-ring structure, including a per-frequency order-6 solution $\boldsymbol{z}_{F6}$ (Corollary 2), a order-4 solution $\boldsymbol{z}_{F4}$ (Corollary 3) and the perfect memorization solution $\boldsymbol{z}_M$ (Corollary 4). If we remove the last term in ℓ_2 loss, then there will be order-3 solution (Lemma 2), as shown in Fig. 2.

We also provide gradient dynamics analysis in Appendix B that shows that the inductive bias in gradient descent prefers simpler global optimizers (Theorem 5) and overparameterization decouples gradient dynamics for each MP, and thus makes the training easier (Theorem 6). We also provide experiments to verify the claim.

References

[1] Anthropic. The claude 3 model family: Opus, sonnet, haiku.

[2] Lukas Berglund, Meg Tong, Max Kaufmann, Mikita Balesni, Asa Cooper Stickland, Tomasz Korbak, and Owain Evans. The reversal curse: Llms trained on" a is b" fail to learn" b is a". *arXiv preprint arXiv:2309.12288*, 2023.

[3] Keith Conrad. Characters of finite abelian groups. *Lecture Notes*, 17, 2010.

[4] DeepSeek-AI, Aixin Liu, Bei Feng, Bin Wang, Bingxuan Wang, Bo Liu, Chenggang Zhao, Chengqi Dengr, Chong Ruan, Damai Dai, Daya Guo, Dejian Yang, Deli Chen, Dongjie Ji, Erhang Li, Fangyun Lin, Fuli Luo, Guangbo Hao, Guanting Chen, Guowei Li, H. Zhang, Hanwei Xu, Hao Yang, Haowei Zhang, Honghui Ding, Huajian Xin, Huazuo Gao, Hui Li, Hui Qu, J. L. Cai, Jian Liang, Jianzhong Guo, Jiaqi Ni, Jiashi Li, Jin Chen, Jingyang Yuan, Junjie Qiu, Junxiao Song, Kai Dong, Kaige Gao, Kang Guan, Lean Wang, Lecong Zhang, Lei Xu, Leyi Xia, Liang Zhao, Liyue Zhang, Meng Li, Miaojun Wang, Mingchuan Zhang, Minghua Zhang, Minghui Tang, Mingming Li, Ning Tian, Panpan Huang, Peiyi Wang, Peng Zhang, Qihao Zhu, Qinyu Chen, Qiushi Du, R. J. Chen, R. L. Jin, Ruiqi Ge, Ruizhe Pan, Runxin Xu, Ruyi Chen, S. S. Li, Shanghao Lu, Shangyan Zhou, Shanhuang Chen, Shaoqing Wu, Shengfeng Ye, Shirong Ma, Shiyu Wang, Shuang Zhou, Shuiping Yu, Shunfeng Zhou, Size Zheng, T. Wang, Tian Pei, Tian Yuan, Tianyu Sun, W. L. Xiao, Wangding Zeng, Wei An, Wen Liu, Wenfeng Liang, Wenjun Gao, Wentao Zhang, X. Q. Li, Xiangyue Jin, Xianzu Wang, Xiao Bi, Xiaodong Liu, Xiaohan Wang, Xiaojin Shen, Xiaokang Chen, Xiaosha Chen, Xiaotao Nie, Xiaowen Sun, Xiaoxiang Wang, Xin Liu, Xin Xie, Xingkai Yu, Xinnan Song, Xinyi Zhou, Xinyu Yang, Xuan Lu, Xuecheng Su, Y. Wu, Y. K. Li, Y. X. Wei, Y. X. Zhu, Yanhong Xu, Yanping Huang, Yao Li, Yao Zhao, Yaofeng Sun, Yaohui Li, Yaohui Wang, Yi Zheng, Yichao Zhang, Yiliang Xiong, Yilong Zhao, Ying He, Ying Tang, Yishi Piao, Yixin Dong, Yixuan Tan, Yiyuan Liu, Yongji Wang, Yongqiang Guo, Yuchen Zhu, Yuduan Wang, Yuheng Zou, Yukun Zha, Yunxian Ma, Yuting Yan, Yuxiang You, Yuxuan Liu, Z. Z. Ren, Zehui Ren, Zhangli Sha, Zhe Fu, Zhen Huang, Zhen Zhang, Zhenda Xie, Zhewen Hao, Zhihong Shao, Zhiniu Wen, Zhipeng Xu, Zhongyu Zhang, Zhuoshu Li, Zihan Wang, Zihui Gu, Zilin Li, and Ziwei Xie. Deepseek-v2: A strong, economical, and efficient mixture-of-experts language model, 2024.

[5] Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, Arun Rao, Aston Zhang, Aurelien Rodriguez, Austen Gregerson, Ava Spataru, Baptiste Roziere, Bethany Biron, Binh Tang, Bobbie Chern, Charlotte Caucheteux, Chaya Nayak, Chloe Bi, Chris Marra, Chris McConnell, Christian Keller, Christophe Touret, Chunyang Wu, Corinne Wong, Cristian Canton Ferrer, Cyrus Nikolaidis, Damien Allonsius, Daniel Song, Danielle Pintz, Danny Livshits, David Esiobu, Dhruv Choudhary, Dhruv Mahajan, Diego Garcia-Olano, Diego Perino, Dieuwke Hupkes, Egor Lakomkin, Ehab AlBadawy, Elina Lobanova, Emily Dinan, Eric Michael Smith, Filip Radenovic, Frank Zhang, Gabriel Synnaeve, Gabrielle Lee, Georgia Lewis Anderson, Graeme Nail, Gregoire Mialon, Guan Pang, Guillem Cucurell, Hailey Nguyen, Hannah Korevaar, Hu Xu, Hugo Touvron, Iliyan Zarov, Imanol Arrieta Ibarra, Isabel Kloumann, Ishan Misra, Ivan Evtimov, Jade Copet, Jaewon Lee, Jan Geffert, Jana Vranes, Jason Park, Jay Mahadeokar, Jeet Shah, Jelmer van der Linde, Jennifer Billock, Jenny Hong, Jenya Lee, Jeremy Fu, Jianfeng Chi, Jianyu Huang, Jiawen Liu, Jie Wang, Jiecao Yu, Joanna Bitton, Joe Spisak, Jongsoo Park, Joseph Rocca, Joshua Johnstun, Joshua Saxe, Junteng Jia, Kalyan Vasuden Alwala, Kartikeya Upasani, Kate Plawiak, Ke Li, Kenneth Heafield, Kevin Stone, Khalid El-Arini, Krithika Iyer, Kshitiz Malik, Kuenley Chiu, Kunal Bhalla, Lauren Rantala-Yeary, Laurens van der Maaten, Lawrence Chen, Liang Tan, Liz Jenkins, Louis Martin, Lovish Madaan, Lubo Malo, Lukas Blecher, Lukas Landzaat, Luke de Oliveira, Madeline Muzzi, Mahesh Pasupuleti, Mannat Singh, Manohar Paluri, Marcin Kardas, Mathew Oldham, Mathieu Rita, Maya Pavlova, Melanie Kambadur, Mike Lewis, Min Si, Mitesh Kumar Singh, Mona Hassan, Naman Goyal, Narjes Torabi, Nikolay Bashlykov, Nikolay Bogoychev, Niladri Chatterji, Olivier Duchenne, Onur Celebi, Patrick Alrassy, Pengchuan Zhang, Pengwei Li, Petar Vasic, Peter Weng, Prajjwal Bhargava, Pratik Dubal, Praveen Krishnan, Punit Singh Koura, Puxin Xu, Qing He, Qingxiao Dong, Ragavan Srinivasan, Raj

Ganapathy, Ramon Calderer, Ricardo Silveira Cabral, Robert Stojnic, Roberta Raileanu, Rohit Girdhar, Rohit Patel, Romain Sauvestre, Ronnie Polidoro, Roshan Sumbaly, Ross Taylor, Ruan Silva, Rui Hou, Rui Wang, Saghar Hosseini, Sahana Chennabasappa, Sanjay Singh, Sean Bell, Seohyun Sonia Kim, Sergey Edunov, Shaoliang Nie, Sharan Narang, Sharath Raparthy, Sheng Shen, Shengye Wan, Shruti Bhosale, Shun Zhang, Simon Vandenhende, Soumya Batra, Spencer Whitman, Sten Sootla, Stephane Collot, Suchin Gururangan, Sydney Borodinsky, Tamar Herman, Tara Fowler, Tarek Sheasha, Thomas Georgiou, Thomas Scialom, Tobias Speckbacher, Todor Mihaylov, Tong Xiao, Ujjwal Karn, Vedanuj Goswami, Vibhor Gupta, Vignesh Ramanathan, Viktor Kerkez, Vincent Gonguet, Virginie Do, Vish Vogeti, Vladan Petrovic, Weiwei Chu, Wenhan Xiong, Wenyin Fu, Whitney Meers, Xavier Martinet, Xiaodong Wang, Xiaoqing Ellen Tan, Xinfeng Xie, Xuchao Jia, Xuewei Wang, Yaelle Goldschlag, Yashesh Gaur, Yasmine Babaei, Yi Wen, Yiwen Song, Yuchen Zhang, Yue Li, Yuning Mao, Zacharie Delpierre Coudert, Zheng Yan, Zhengxing Chen, Zoe Papakipos, Aaditya Singh, Aaron Grattafiori, Abha Jain, Adam Kelsey, Adam Shajnfeld, Adithya Gangidi, Adolfo Victoria, Ahuva Goldstand, Ajay Menon, Ajay Sharma, Alex Boesenberg, Alex Vaughan, Alexei Baevski, Allie Feinstein, Amanda Kallet, Amit Sangani, Anam Yunus, Andrei Lupu, Andres Alvarado, Andrew Caples, Andrew Gu, Andrew Ho, Andrew Poulton, Andrew Ryan, Ankit Ramchandani, Annie Franco, Aparajita Saraf, Arkabandhu Chowdhury, Ashley Gabriel, Ashwin Bharambe, Assaf Eisenman, Azadeh Yazdan, Beau James, Ben Maurer, Benjamin Leonhardi, Bernie Huang, Beth Loyd, Beto De Paola, Bhargavi Paranjape, Bing Liu, Bo Wu, Boyu Ni, Braden Hancock, Bram Wasti, Brandon Spence, Brani Stojkovic, Brian Gamido, Britt Montalvo, Carl Parker, Carly Burton, Catalina Mejia, Changhan Wang, Changkyu Kim, Chao Zhou, Chester Hu, Ching-Hsiang Chu, Chris Cai, Chris Tindal, Christoph Feichtenhofer, Damon Civin, Dana Beaty, Daniel Kreymer, Daniel Li, Danny Wyatt, David Adkins, David Xu, Davide Testuggine, Delia David, Devi Parikh, Diana Liskovich, Didem Foss, Dingkang Wang, Duc Le, Dustin Holland, Edward Dowling, Eissa Jamil, Elaine Montgomery, Eleonora Presani, Emily Hahn, Emily Wood, Erik Brinkman, Esteban Arcaute, Evan Dunbar, Evan Smothers, Fei Sun, Felix Kreuk, Feng Tian, Firat Ozgenel, Francesco Caggioni, Francisco Guzmán, Frank Kanayet, Frank Seide, Gabriela Medina Florez, Gabriella Schwarz, Gada Badeer, Georgia Swee, Gil Halpern, Govind Thattai, Grant Herman, Grigory Sizov, Guangyi, Zhang, Guna Lakshminarayanan, Hamid Shojanazeri, Han Zou, Hannah Wang, Hanwen Zha, Haroun Habeeb, Harrison Rudolph, Helen Suk, Henry Aspegren, Hunter Goldman, Ibrahim Damlaj, Igor Molybog, Igor Tufanov, Irina-Elena Veliche, Itai Gat, Jake Weissman, James Geboski, James Kohli, Japhet Asher, Jean-Baptiste Gaya, Jeff Marcus, Jeff Tang, Jennifer Chan, Jenny Zhen, Jeremy Reizenstein, Jeremy Teboul, Jessica Zhong, Jian Jin, Jingyi Yang, Joe Cummings, Jon Carvill, Jon Shepard, Jonathan McPhie, Jonathan Torres, Josh Ginsburg, Junjie Wang, Kai Wu, Kam Hou U, Karan Saxena, Karthik Prasad, Kartikay Khandelwal, Katayoun Zand, Kathy Matosich, Kaushik Veeraraghavan, Kelly Michelena, Keqian Li, Kun Huang, Kunal Chawla, Kushal Lakhotia, Kyle Huang, Lailin Chen, Lakshya Garg, Lavender A, Leandro Silva, Lee Bell, Lei Zhang, Liangpeng Guo, Licheng Yu, Liron Moshkovich, Luca Wehrstedt, Madian Khabsa, Manav Avalani, Manish Bhatt, Maria Tsimpoukelli, Martynas Mankus, Matan Hasson, Matthew Lennie, Matthias Reso, Maxim Groshev, Maxim Naumov, Maya Lathi, Meghan Keneally, Michael L. Seltzer, Michal Valko, Michelle Restrepo, Mihir Patel, Mik Vyatskov, Mikayel Samvelyan, Mike Clark, Mike Macey, Mike Wang, Miquel Jubert Hermoso, Mo Metanat, Mohammad Rastegari, Munish Bansal, Nandhini Santhanam, Natascha Parks, Natasha White, Navyata Bawa, Nayan Singhal, Nick Egebo, Nicolas Usunier, Nikolay Pavlovich Laptev, Ning Dong, Ning Zhang, Norman Cheng, Oleg Chernoguz, Olivia Hart, Omkar Salpekar, Ozlem Kalinli, Parkin Kent, Parth Parekh, Paul Saab, Pavan Balaji, Pedro Rittner, Philip Bontrager, Pierre Roux, Piotr Dollar, Polina Zvyagina, Prashant Ratanchandani, Pritish Yuvraj, Qian Liang, Rachad Alao, Rachel Rodriguez, Rafi Ayub, Raghotham Murthy, Raghu Nayani, Rahul Mitra, Raymond Li, Rebekkah Hogan, Robin Battey, Rocky Wang, Rohan Maheswari, Russ Howes, Ruty Rinott, Sai Jayesh Bondu, Samyak Datta, Sara Chugh, Sara Hunt, Sargun Dhillon, Sasha Sidorov, Satadru Pan, Saurabh Verma, Seiji Yamamoto, Sharadh Ramaswamy, Shaun Lindsay, Shaun Lindsay, Sheng Feng, Shenghao Lin, Shengxin Cindy Zha, Shiva Shankar, Shuqiang Zhang, Shuqiang Zhang, Sinong Wang, Sneha Agarwal, Soji Sajuyigbe, Soumith Chintala, Stephanie Max, Stephen Chen, Steve Kehoe, Steve Satterfield, Sudarshan Govindaprasad, Sumit Gupta, Sungmin Cho, Sunny Virk, Suraj Subramanian, Sy Choudhury, Sydney Goldman, Tal Remez, Tamar Glaser, Tamara Best, Thilo Kohler, Thomas Robinson, Tianhe Li, Tianjun Zhang, Tim Matthews, Timothy Chou, Tzook

Shaked, Varun Vontimitta, Victoria Ajayi, Victoria Montanez, Vijai Mohan, Vinay Satish Kumar, Vishal Mangla, Vitor Albiero, Vlad Ionescu, Vlad Poenaru, Vlad Tiberiu Mihailescu, Vladimir Ivanov, Wei Li, Wenchen Wang, Wenwen Jiang, Wes Bouaziz, Will Constable, Xiaocheng Tang, Xiaofang Wang, Xiaojian Wu, Xiaolan Wang, Xide Xia, Xilun Wu, Xinbo Gao, Yanjun Chen, Ye Hu, Ye Jia, Ye Qi, Yenda Li, Yilin Zhang, Ying Zhang, Yossi Adi, Youngjin Nam, Yu, Wang, Yuchen Hao, Yundi Qian, Yuzi He, Zach Rait, Zachary DeVito, Zef Rosnbrick, Zhaoduo Wen, Zhenyu Yang, and Zhiwei Zhao. The llama 3 herd of models, 2024.

[6] Nouha Dziri, Ximing Lu, Melanie Sclar, Xiang Lorraine Li, Liwei Jiang, Bill Yuchen Lin, Peter West, Chandra Bhagavatula, Ronan Le Bras, Jena D Hwang, et al. Faith and fate: Limits of transformers on compositionality (2023). *arXiv preprint arXiv:2305.18654*, 2023.

[7] William Fulton and Joe Harris. *Representation theory: a first course*, volume 129. Springer Science & Business Media, 2013.

[8] Andrey Gromov. Grokking modular arithmetic. *arXiv preprint arXiv:2301.02679*, 2023.

[9] Jie Huang, Xinyun Chen, Swaroop Mishra, Huaixiu Steven Zheng, Adams Wei Yu, Xinying Song, and Denny Zhou. Large language models cannot self-correct reasoning yet. *arXiv preprint arXiv:2310.01798*, 2023.

[10] Yufei Huang, Shengding Hu, Xu Han, Zhiyuan Liu, and Maosong Sun. Unified view of grokking, double descent and emergent abilities: A perspective from circuits competition. *arXiv preprint arXiv:2402.15175*, 2024.

[11] Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. Mistral 7b, 2023.

[12] Charles Jin and Martin Rinard. Emergent representations of program semantics in language models trained on programs, 2024.

[13] Subbarao Kambhampati, Karthik Valmeekam, Lin Guan, Mudit Verma, Kaya Stechly, Siddhant Bhambri, Lucas Saldyt, and Anil Murthy. Llms can't plan, but can help planning in llm-modulo frameworks, 2024.

[14] CG Khatri and C Radhakrishna Rao. Solutions to some functional equations and their applications to characterization of probability distributions. *Sankhyā: the Indian journal of statistics, series A*, pages 167–180, 1968.

[15] Depen Morwani, Benjamin L Edelman, Costin-Andrei Oncescu, Rosie Zhao, and Sham Kakade. Feature emergence via margin maximization: case studies in algebraic tasks. *arXiv preprint arXiv:2311.07568*, 2023.

[16] Neel Nanda, Lawrence Chan, Tom Lieberum, Jess Smith, and Jacob Steinhardt. Progress measures for grokking via mechanistic interpretability. In *The Eleventh International Conference on Learning Representations*, 2023.

[17] Marianna Nezhurina, Lucia Cipolina-Kun, Mehdi Cherti, and Jenia Jitsev. Alice in wonderland: Simple tasks showing complete reasoning breakdown in state-of-the-art large language models, 2024.

[18] OpenAI, Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, Red Avila, Igor Babuschkin, Suchir Balaji, Valerie Balcom, Paul Baltescu, Haiming Bao, Mohammad Bavarian, Jeff Belgum, Irwan Bello, Jake Berdine, Gabriel Bernadett-Shapiro, Christopher Berner, Lenny Bogdonoff, Oleg Boiko, Madelaine Boyd, Anna-Luisa Brakman, Greg Brockman, Tim Brooks, Miles Brundage, Kevin Button, Trevor Cai, Rosie Campbell, Andrew Cann, Brittany Carey, Chelsea Carlson, Rory Carmichael, Brooke Chan, Che Chang, Fotis Chantzis, Derek Chen, Sully Chen, Ruby Chen, Jason Chen, Mark Chen, Ben Chess, Chester Cho, Casey Chu, Hyung Won Chung, Dave Cummings, Jeremiah Currier, Yunxing Dai, Cory Decareaux, Thomas Degry, Noah Deutsch, Damien Deville, Arka Dhar, David Dohan, Steve

Dowling, Sheila Dunning, Adrien Ecoffet, Atty Eleti, Tyna Eloundou, David Farhi, Liam Fedus, Niko Felix, Simón Posada Fishman, Juston Forte, Isabella Fulford, Leo Gao, Elie Georges, Christian Gibson, Vik Goel, Tarun Gogineni, Gabriel Goh, Rapha Gontijo-Lopes, Jonathan Gordon, Morgan Grafstein, Scott Gray, Ryan Greene, Joshua Gross, Shixiang Shane Gu, Yufei Guo, Chris Hallacy, Jesse Han, Jeff Harris, Yuchen He, Mike Heaton, Johannes Heidecke, Chris Hesse, Alan Hickey, Wade Hickey, Peter Hoeschele, Brandon Houghton, Kenny Hsu, Shengli Hu, Xin Hu, Joost Huizinga, Shantanu Jain, Shawn Jain, Joanne Jang, Angela Jiang, Roger Jiang, Haozhun Jin, Denny Jin, Shino Jomoto, Billie Jonn, Heewoo Jun, Tomer Kaftan, Łukasz Kaiser, Ali Kamali, Ingmar Kanitscheider, Nitish Shirish Keskar, Tabarak Khan, Logan Kilpatrick, Jong Wook Kim, Christina Kim, Yongjik Kim, Jan Hendrik Kirchner, Jamie Kiros, Matt Knight, Daniel Kokotajlo, Łukasz Kondraciuk, Andrew Kondrich, Aris Konstantinidis, Kyle Kosic, Gretchen Krueger, Vishal Kuo, Michael Lampe, Ikai Lan, Teddy Lee, Jan Leike, Jade Leung, Daniel Levy, Chak Ming Li, Rachel Lim, Molly Lin, Stephanie Lin, Mateusz Litwin, Theresa Lopez, Ryan Lowe, Patricia Lue, Anna Makanju, Kim Malfacini, Sam Manning, Todor Markov, Yaniv Markovski, Bianca Martin, Katie Mayer, Andrew Mayne, Bob McGrew, Scott Mayer McKinney, Christine McLeavey, Paul McMillan, Jake McNeil, David Medina, Aalok Mehta, Jacob Menick, Luke Metz, Andrey Mishchenko, Pamela Mishkin, Vinnie Monaco, Evan Morikawa, Daniel Mossing, Tong Mu, Mira Murati, Oleg Murk, David Mély, Ashvin Nair, Reiichiro Nakano, Rajeev Nayak, Arvind Neelakantan, Richard Ngo, Hyeonwoo Noh, Long Ouyang, Cullen O'Keefe, Jakub Pachocki, Alex Paino, Joe Palermo, Ashley Pantuliano, Giambattista Parascandolo, Joel Parish, Emy Parparita, Alex Passos, Mikhail Pavlov, Andrew Peng, Adam Perelman, Filipe de Avila Belbute Peres, Michael Petrov, Henrique Ponde de Oliveira Pinto, Michael, Pokorny, Michelle Pokrass, Vitchyr H. Pong, Tolly Powell, Alethea Power, Boris Power, Elizabeth Proehl, Raul Puri, Alec Radford, Jack Rae, Aditya Ramesh, Cameron Raymond, Francis Real, Kendra Rimbach, Carl Ross, Bob Rotsted, Henri Roussez, Nick Ryder, Mario Saltarelli, Ted Sanders, Shibani Santurkar, Girish Sastry, Heather Schmidt, David Schnurr, John Schulman, Daniel Selsam, Kyla Sheppard, Toki Sherbakov, Jessica Shieh, Sarah Shoker, Pranav Shyam, Szymon Sidor, Eric Sigler, Maddie Simens, Jordan Sitkin, Katarina Slama, Ian Sohl, Benjamin Sokolowsky, Yang Song, Natalie Staudacher, Felipe Petroski Such, Natalie Summers, Ilya Sutskever, Jie Tang, Nikolas Tezak, Madeleine B. Thompson, Phil Tillet, Amin Tootoonchian, Elizabeth Tseng, Preston Tuggle, Nick Turley, Jerry Tworek, Juan Felipe Cerón Uribe, Andrea Vallone, Arun Vijayvergiya, Chelsea Voss, Carroll Wainwright, Justin Jay Wang, Alvin Wang, Ben Wang, Jonathan Ward, Jason Wei, CJ Weinmann, Akila Welihinda, Peter Welinder, Jiayi Weng, Lilian Weng, Matt Wiethoff, Dave Willner, Clemens Winter, Samuel Wolrich, Hannah Wong, Lauren Workman, Sherwin Wu, Jeff Wu, Michael Wu, Kai Xiao, Tao Xu, Sarah Yoo, Kevin Yu, Qiming Yuan, Wojciech Zaremba, Rowan Zellers, Chong Zhang, Marvin Zhang, Shengjia Zhao, Tianhao Zheng, Juntang Zhuang, William Zhuk, and Barret Zoph. Gpt-4 technical report, 2024.

[19] Simon Ouellette, Rolf Pfister, and Hansueli Jud. Counting and algorithmic generalization with transformers. *arXiv preprint arXiv:2310.08661*, 2023.

[20] Alethea Power, Yuri Burda, Harri Edwards, Igor Babuschkin, and Vedant Misra. Grokking: Generalization beyond overfitting on small algorithmic datasets. *arXiv preprint arXiv:2201.02177*, 2022.

[21] Benjamin Steinberg. Representation theory of finite groups. *Carleton University*, 2009.

[22] Gemini Team, Petko Georgiev, Ving Ian Lei, Ryan Burnell, Libin Bai, Anmol Gulati, Garrett Tanzer, Damien Vincent, Zhufeng Pan, Shibo Wang, Soroosh Mariooryad, Yifan Ding, Xinyang Geng, Fred Alcober, Roy Frostig, Mark Omernick, Lexi Walker, Cosmin Paduraru, Christina Sorokin, Andrea Tacchetti, Colin Gaffney, Samira Daruki, Olcan Sercinoglu, Zach Gleicher, Juliette Love, Paul Voigtlaender, Rohan Jain, Gabriela Surita, Kareem Mohamed, Rory Blevins, Junwhan Ahn, Tao Zhu, Kornraphop Kawintiranon, Orhan Firat, Yiming Gu, Yujing Zhang, Matthew Rahtz, Manaal Faruqui, Natalie Clay, Justin Gilmer, JD Co-Reyes, Ivo Penchev, Rui Zhu, Nobuyuki Morioka, Kevin Hui, Krishna Haridasan, Victor Campos, Mahdis Mahdieh, Mandy Guo, Samer Hassan, Kevin Kilgour, Arpi Vezer, Heng-Tze Cheng, Raoul de Liedekerke, Siddharth Goyal, Paul Barham, DJ Strouse, Seb Noury, Jonas Adler, Mukund Sundararajan, Sharad Vikram, Dmitry Lepikhin, Michela Paganini, Xavier Garcia, Fan Yang, Dasha Valter, Maja Trebacz, Kiran Vodrahalli, Chulayuth Asawaroengchai, Roman Ring, Norbert Kalb, Livio Baldini Soares, Siddhartha Brahma, David Steiner, Tianhe

Yu, Fabian Mentzer, Antoine He, Lucas Gonzalez, Bibo Xu, Raphael Lopez Kaufman, Laurent El Shafey, Junhyuk Oh, Tom Hennigan, George van den Driessche, Seth Odoom, Mario Lucic, Becca Roelofs, Sid Lall, Amit Marathe, Betty Chan, Santiago Ontanon, Luheng He, Denis Teplyashin, Jonathan Lai, Phil Crone, Bogdan Damoc, Lewis Ho, Sebastian Riedel, Karel Lenc, Chih-Kuan Yeh, Aakanksha Chowdhery, Yang Xu, Mehran Kazemi, Ehsan Amid, Anastasia Petrushkina, Kevin Swersky, Ali Khodaei, Gowoon Chen, Chris Larkin, Mario Pinto, Geng Yan, Adria Puigdomenech Badia, Piyush Patil, Steven Hansen, Dave Orr, Sebastien M. R. Arnold, Jordan Grimstad, Andrew Dai, Sholto Douglas, Rishika Sinha, Vikas Yadav, Xi Chen, Elena Gribovskaya, Jacob Austin, Jeffrey Zhao, Kaushal Patel, Paul Komarek, Sophia Austin, Sebastian Borgeaud, Linda Friso, Abhimanyu Goyal, Ben Caine, Kris Cao, Da-Woon Chung, Matthew Lamm, Gabe Barth-Maron, Thais Kagohara, Kate Olszewska, Mia Chen, Kaushik Shivakumar, Rishabh Agarwal, Harshal Godhia, Ravi Rajwar, Javier Snaider, Xerxes Dotiwalla, Yuan Liu, Aditya Barua, Victor Ungureanu, Yuan Zhang, Bat-Orgil Batsaikhan, Mateo Wirth, James Qin, Ivo Danihelka, Tulsee Doshi, Martin Chadwick, Jilin Chen, Sanil Jain, Quoc Le, Arjun Kar, Madhu Gurumurthy, Cheng Li, Ruoxin Sang, Fangyu Liu, Lampros Lamprou, Rich Munoz, Nathan Lintz, Harsh Mehta, Heidi Howard, Malcolm Reynolds, Lora Aroyo, Quan Wang, Lorenzo Blanco, Albin Cassirer, Jordan Griffith, Dipanjan Das, Stephan Lee, Jakub Sygnowski, Zach Fisher, James Besley, Richard Powell, Zafarali Ahmed, Dominik Paulus, David Reitter, Zalan Borsos, Rishabh Joshi, Aedan Pope, Steven Hand, Vittorio Selo, Vihan Jain, Nikhil Sethi, Megha Goel, Takaki Makino, Rhys May, Zhen Yang, Johan Schalkwyk, Christina Butterfield, Anja Hauth, Alex Goldin, Will Hawkins, Evan Senter, Sergey Brin, Oliver Woodman, Marvin Ritter, Eric Noland, Minh Giang, Vijay Bolina, Lisa Lee, Tim Blyth, Ian Mackinnon, Machel Reid, Obaid Sarvana, David Silver, Alexander Chen, Lily Wang, Loren Maggiore, Oscar Chang, Nithya Attaluri, Gregory Thornton, Chung-Cheng Chiu, Oskar Bunyan, Nir Levine, Timothy Chung, Evgenii Eltyshev, Xiance Si, Timothy Lillicrap, Demetra Brady, Vaibhav Aggarwal, Boxi Wu, Yuanzhong Xu, Ross McIlroy, Kartikeya Badola, Paramjit Sandhu, Erica Moreira, Wojciech Stokowiec, Ross Hemsley, Dong Li, Alex Tudor, Pranav Shyam, Elahe Rahimtoroghi, Salem Haykal, Pablo Sprechmann, Xiang Zhou, Diana Mincu, Yujia Li, Ravi Addanki, Kalpesh Krishna, Xiao Wu, Alexandre Frechette, Matan Eyal, Allan Dafoe, Dave Lacey, Jay Whang, Thi Avrahami, Ye Zhang, Emanuel Taropa, Hanzhao Lin, Daniel Toyama, Eliza Rutherford, Motoki Sano, HyunJeong Choe, Alex Tomala, Chalence Safranek-Shrader, Nora Kassner, Mantas Pajarskas, Matt Harvey, Sean Sechrist, Meire Fortunato, Christina Lyu, Gamaleldin Elsayed, Chenkai Kuang, James Lottes, Eric Chu, Chao Jia, Chih-Wei Chen, Peter Humphreys, Kate Baumli, Connie Tao, Rajkumar Samuel, Cicero Nogueira dos Santos, Anders Andreassen, Nemanja Rakićević, Dominik Grewe, Aviral Kumar, Stephanie Winkler, Jonathan Caton, Andrew Brock, Sid Dalmia, Hannah Sheahan, Iain Barr, Yingjie Miao, Paul Natsev, Jacob Devlin, Feryal Behbahani, Flavien Prost, Yanhua Sun, Artiom Myaskovsky, Thanumalayan Sankaranarayana Pillai, Dan Hurt, Angeliki Lazaridou, Xi Xiong, Ce Zheng, Fabio Pardo, Xiaowei Li, Dan Horgan, Joe Stanton, Moran Ambar, Fei Xia, Alejandro Lince, Mingqiu Wang, Basil Mustafa, Albert Webson, Hyo Lee, Rohan Anil, Martin Wicke, Timothy Dozat, Abhishek Sinha, Enrique Piqueras, Elahe Dabir, Shyam Upadhyay, Anudhyan Boral, Lisa Anne Hendricks, Corey Fry, Josip Djolonga, Yi Su, Jake Walker, Jane Labanowski, Ronny Huang, Vedant Misra, Jeremy Chen, RJ Skerry-Ryan, Avi Singh, Shruti Rijhwani, Dian Yu, Alex Castro-Ros, Beer Changpinyo, Romina Datta, Sumit Bagri, Arnar Mar Hrafnkelsson, Marcello Maggioni, Daniel Zheng, Yury Sulsky, Shaobo Hou, Tom Le Paine, Antoine Yang, Jason Riesa, Dominika Rogozinska, Dror Marcus, Dalia El Badawy, Qiao Zhang, Luyu Wang, Helen Miller, Jeremy Greer, Lars Lowe Sjos, Azade Nova, Heiga Zen, Rahma Chaabouni, Mihaela Rosca, Jiepu Jiang, Charlie Chen, Ruibo Liu, Tara Sainath, Maxim Krikun, Alex Polozov, Jean-Baptiste Lespiau, Josh Newlan, Zeyncep Cankara, Soo Kwak, Yunhan Xu, Phil Chen, Andy Coenen, Clemens Meyer, Katerina Tsihlas, Ada Ma, Juraj Gottweis, Jinwei Xing, Chenjie Gu, Jin Miao, Christian Frank, Zeynep Cankara, Sanjay Ganapathy, Ishita Dasgupta, Steph Hughes-Fitt, Heng Chen, David Reid, Keran Rong, Hongmin Fan, Joost van Amersfoort, Vincent Zhuang, Aaron Cohen, Shixiang Shane Gu, Anhad Mohananey, Anastasija Ilic, Taylor Tobin, John Wieting, Anna Bortsova, Phoebe Thacker, Emma Wang, Emily Caveness, Justin Chiu, Eren Sezener, Alex Kaskasoli, Steven Baker, Katie Millican, Mohamed Elhawaty, Kostas Aisopos, Carl Lebsack, Nathan Byrd, Hanjun Dai, Wenhao Jia, Matthew Wiethoff, Elnaz Davoodi, Albert Weston, Lakshman Yagati, Arun Ahuja, Isabel Gao, Golan Pundak, Susan Zhang, Michael Azzam, Khe Chai Sim, Sergi Caelles, James Keeling, Abhanshu Sharma, Andy Swing, YaGuang Li, Chenxi Liu, Carrie Grimes

Bostock, Yamini Bansal, Zachary Nado, Ankesh Anand, Josh Lipschultz, Abhijit Karmarkar, Lev Proleev, Abe Ittycheriah, Soheil Hassas Yeganeh, George Polovets, Aleksandra Faust, Jiao Sun, Alban Rrustemi, Pen Li, Rakesh Shivanna, Jeremiah Liu, Chris Welty, Federico Lebron, Anirudh Baddepudi, Sebastian Krause, Emilio Parisotto, Radu Soricut, Zheng Xu, Dawn Bloxwich, Melvin Johnson, Behnam Neyshabur, Justin Mao-Jones, Renshen Wang, Vinay Ramasesh, Zaheer Abbas, Arthur Guez, Constant Segal, Duc Dung Nguyen, James Svensson, Le Hou, Sarah York, Kieran Milan, Sophie Bridgers, Wiktor Gworek, Marco Tagliasacchi, James Lee-Thorp, Michael Chang, Alexey Guseynov, Ale Jakse Hartman, Michael Kwong, Ruizhe Zhao, Sheleem Kashem, Elizabeth Cole, Antoine Miech, Richard Tanburn, Mary Phuong, Filip Pavetic, Sebastien Cevey, Ramona Comanescu, Richard Ives, Sherry Yang, Cosmo Du, Bo Li, Zizhao Zhang, Mariko Iinuma, Clara Huiyi Hu, Aurko Roy, Shaan Bijwadia, Zhenkai Zhu, Danilo Martins, Rachel Saputro, Anita Gergely, Steven Zheng, Dawei Jia, Ioannis Antonoglou, Adam Sadovsky, Shane Gu, Yingying Bi, Alek Andreev, Sina Samangooei, Mina Khan, Tomas Kocisky, Angelos Filos, Chintu Kumar, Colton Bishop, Adams Yu, Sarah Hodkinson, Sid Mittal, Premal Shah, Alexandre Moufarek, Yong Cheng, Adam Bloniarz, Jaehoon Lee, Pedram Pejman, Paul Michel, Stephen Spencer, Vladimir Feinberg, Xuehan Xiong, Nikolay Savinov, Charlotte Smith, Siamak Shakeri, Dustin Tran, Mary Chesus, Bernd Bohnet, George Tucker, Tamara von Glehn, Carrie Muir, Yiran Mao, Hideto Kazawa, Ambrose Slone, Kedar Soparkar, Disha Shrivastava, James Cobon-Kerr, Michael Sharman, Jay Pavagadhi, Carlos Araya, Karolis Misiunas, Nimesh Ghelani, Michael Laskin, David Barker, Qiujia Li, Anton Briukhov, Neil Houlsby, Mia Glaese, Balaji Lakshminarayanan, Nathan Schucher, Yunhao Tang, Eli Collins, Hyeontaek Lim, Fangxiaoyu Feng, Adria Recasens, Guangda Lai, Alberto Magni, Nicola De Cao, Aditya Siddhant, Zoe Ashwood, Jordi Orbay, Mostafa Dehghani, Jenny Brennan, Yifan He, Kelvin Xu, Yang Gao, Carl Saroufim, James Molloy, Xinyi Wu, Seb Arnold, Solomon Chang, Julian Schrittwieser, Elena Buchatskaya, Soroush Radpour, Martin Polacek, Skye Giordano, Ankur Bapna, Simon Tokumine, Vincent Hellendoorn, Thibault Sottiaux, Sarah Cogan, Aliaksei Severyn, Mohammad Saleh, Shantanu Thakoor, Laurent Shefey, Siyuan Qiao, Meenu Gaba, Shuo yiin Chang, Craig Swanson, Biao Zhang, Benjamin Lee, Paul Kishan Rubenstein, Gan Song, Tom Kwiatkowski, Anna Koop, Ajay Kannan, David Kao, Parker Schuh, Axel Stjerngren, Golnaz Ghiasi, Gena Gibson, Luke Vilnis, Ye Yuan, Felipe Tiengo Ferreira, Aishwarya Kamath, Ted Klimenko, Ken Franko, Kefan Xiao, Indro Bhattacharya, Miteyan Patel, Rui Wang, Alex Morris, Robin Strudel, Vivek Sharma, Peter Choy, Sayed Hadi Hashemi, Jessica Landon, Mara Finkelstein, Priya Jhakra, Justin Frye, Megan Barnes, Matthew Mauger, Dennis Daun, Khuslen Baatarsukh, Matthew Tung, Wael Farhan, Henryk Michalewski, Fabio Viola, Felix de Chaumont Quitry, Charline Le Lan, Tom Hudson, Qingze Wang, Felix Fischer, Ivy Zheng, Elspeth White, Anca Dragan, Jean baptiste Alayrac, Eric Ni, Alexander Pritzel, Adam Iwanicki, Michael Isard, Anna Bulanova, Lukas Zilka, Ethan Dyer, Devendra Sachan, Srivatsan Srinivasan, Hannah Muckenhirn, Honglong Cai, Amol Mandhane, Mukarram Tariq, Jack W. Rae, Gary Wang, Kareem Ayoub, Nicholas FitzGerald, Yao Zhao, Woohyun Han, Chris Alberti, Dan Garrette, Kashyap Krishnakumar, Mai Gimenez, Anselm Levskaya, Daniel Sohn, Josip Matak, Inaki Iturrate, Michael B. Chang, Jackie Xiang, Yuan Cao, Nishant Ranka, Geoff Brown, Adrian Hutter, Vahab Mirrokni, Nanxin Chen, Kaisheng Yao, Zoltan Egyed, Francois Galilee, Tyler Liechty, Praveen Kallakuri, Evan Palmer, Sanjay Ghemawat, Jasmine Liu, David Tao, Chloe Thornton, Tim Green, Mimi Jasarevic, Sharon Lin, Victor Cotruta, Yi-Xuan Tan, Noah Fiedel, Hongkun Yu, Ed Chi, Alexander Neitz, Jens Heitkaemper, Anu Sinha, Denny Zhou, Yi Sun, Charbel Kaed, Brice Hulse, Swaroop Mishra, Maria Georgaki, Sneha Kudugunta, Clement Farabet, Izhak Shafran, Daniel Vlasic, Anton Tsitsulin, Rajagopal Ananthanarayanan, Alen Carin, Guolong Su, Pei Sun, Shashank V, Gabriel Carvajal, Josef Broder, Iulia Comsa, Alena Repina, William Wong, Warren Weilun Chen, Peter Hawkins, Egor Filonov, Lucia Loher, Christoph Hirnschall, Weiyi Wang, Jingchen Ye, Andrea Burns, Hardie Cate, Diana Gage Wright, Federico Piccinini, Lei Zhang, Chu-Cheng Lin, Ionel Gog, Yana Kulizhskaya, Ashwin Sreevatsa, Shuang Song, Luis C. Cobo, Anand Iyer, Chetan Tekur, Guillermo Garrido, Zhuyun Xiao, Rupert Kemp, Huaixiu Steven Zheng, Hui Li, Ananth Agarwal, Christel Ngani, Kati Goshvadi, Rebeca Santamaria-Fernandez, Wojciech Fica, Xinyun Chen, Chris Gorgolewski, Sean Sun, Roopal Garg, Xinyu Ye, S. M. Ali Eslami, Nan Hua, Jon Simon, Pratik Joshi, Yelin Kim, Ian Tenney, Sahitya Potluri, Lam Nguyen Thiet, Quan Yuan, Florian Luisier, Alexandra Chronopoulou, Salvatore Scellato, Praveen Srinivasan, Minmin Chen, Vinod Koverkathu, Valentin Dalibard, Yaming Xu, Brennan Saeta, Keith Anderson, Thibault Sellam, Nick Fernando, Fantine Huot,

Junehyuk Jung, Mani Varadarajan, Michael Quinn, Amit Raul, Maigo Le, Ruslan Habalov, Jon Clark, Komal Jalan, Kalesha Bullard, Achintya Singhal, Thang Luong, Boyu Wang, Sujeevan Rajayogam, Julian Eisenschlos, Johnson Jia, Daniel Finchelstein, Alex Yakubovich, Daniel Balle, Michael Fink, Sameer Agarwal, Jing Li, Dj Dvijotham, Shalini Pal, Kai Kang, Jaclyn Konzelmann, Jennifer Beattie, Olivier Dousse, Diane Wu, Remi Crocker, Chen Elkind, Siddhartha Reddy Jonnalagadda, Jong Lee, Dan Holtmann-Rice, Krystal Kallarackal, Rosanne Liu, Denis Vnukov, Neera Vats, Luca Invernizzi, Mohsen Jafari, Huanjie Zhou, Lilly Taylor, Jennifer Prendki, Marcus Wu, Tom Eccles, Tianqi Liu, Kavya Kopparapu, Francoise Beaufays, Christof Angermueller, Andreea Marzoca, Shourya Sarcar, Hilal Dib, Jeff Stanway, Frank Perbet, Nejc Trdin, Rachel Sterneck, Andrey Khorlin, Dinghua Li, Xihui Wu, Sonam Goenka, David Madras, Sasha Goldshtein, Willi Gierke, Tong Zhou, Yaxin Liu, Yannie Liang, Anais White, Yunjie Li, Shreya Singh, Sanaz Bahargam, Mark Epstein, Sujoy Basu, Li Lao, Adnan Ozturel, Carl Crous, Alex Zhai, Han Lu, Zora Tung, Neeraj Gaur, Alanna Walton, Lucas Dixon, Ming Zhang, Amir Globerson, Grant Uy, Andrew Bolt, Olivia Wiles, Milad Nasr, Ilia Shumailov, Marco Selvi, Francesco Piccinno, Ricardo Aguilar, Sara McCarthy, Misha Khalman, Mrinal Shukla, Vlado Galic, John Carpenter, Kevin Villela, Haibin Zhang, Harry Richardson, James Martens, Matko Bosnjak, Shreyas Rammohan Belle, Jeff Seibert, Mahmoud Alnahlawi, Brian McWilliams, Sankalp Singh, Annie Louis, Wen Ding, Dan Popovici, Lenin Simicich, Laura Knight, Pulkit Mehta, Nishesh Gupta, Chongyang Shi, Saaber Fatehi, Jovana Mitrovic, Alex Grills, Joseph Pagadora, Dessie Petrova, Danielle Eisenbud, Zhishuai Zhang, Damion Yates, Bhavishya Mittal, Nilesh Tripuraneni, Yannis Assael, Thomas Brovelli, Prateek Jain, Mihajlo Velimirovic, Canfer Akbulut, Jiaqi Mu, Wolfgang Macherey, Ravin Kumar, Jun Xu, Haroon Qureshi, Gheorghe Comanici, Jeremy Wiesner, Zhitao Gong, Anton Ruddock, Matthias Bauer, Nick Felt, Anirudh GP, Anurag Arnab, Dustin Zelle, Jonas Rothfuss, Bill Rosgen, Ashish Shenoy, Bryan Seybold, Xinjian Li, Jayaram Mudigonda, Goker Erdogan, Jiawei Xia, Jiri Simsa, Andrea Michi, Yi Yao, Christopher Yew, Steven Kan, Isaac Caswell, Carey Radebaugh, Andre Elisseeff, Pedro Valenzuela, Kay McKinney, Kim Paterson, Albert Cui, Eri Latorre-Chimoto, Solomon Kim, William Zeng, Ken Durden, Priya Ponnapalli, Tiberiu Sosea, Christopher A. Choquette-Choo, James Manyika, Brona Robenek, Harsha Vashisht, Sebastien Pereira, Hoi Lam, Marko Velic, Denese Owusu-Afriyie, Katherine Lee, Tolga Bolukbasi, Alicia Parrish, Shawn Lu, Jane Park, Balaji Venkatraman, Alice Talbert, Lambert Rosique, Yuchung Cheng, Andrei Sozanschi, Adam Paszke, Praveen Kumar, Jessica Austin, Lu Li, Khalid Salama, Wooyeol Kim, Nandita Dukkipati, Anthony Baryshnikov, Christos Kaplanis, XiangHai Sheng, Yuri Chervonyi, Caglar Unlu, Diego de Las Casas, Harry Askham, Kathryn Tunyasuvunakool, Felix Gimeno, Siim Poder, Chester Kwak, Matt Miecnikowski, Vahab Mirrokni, Alek Dimitriev, Aaron Parisi, Dangyi Liu, Tomy Tsai, Toby Shevlane, Christina Kouridi, Drew Garmon, Adrian Goedeckemeyer, Adam R. Brown, Anitha Vijayakumar, Ali Elqursh, Sadegh Jazayeri, Jin Huang, Sara Mc Carthy, Jay Hoover, Lucy Kim, Sandeep Kumar, Wei Chen, Courtney Biles, Garrett Bingham, Evan Rosen, Lisa Wang, Qijun Tan, David Engel, Francesco Pongetti, Dario de Cesare, Dongseong Hwang, Lily Yu, Jennifer Pullman, Srini Narayanan, Kyle Levin, Siddharth Gopal, Megan Li, Asaf Aharoni, Trieu Trinh, Jessica Lo, Norman Casagrande, Roopali Vij, Loic Matthey, Bramandia Ramadhana, Austin Matthews, CJ Carey, Matthew Johnson, Kremena Goranova, Rohin Shah, Shereen Ashraf, Kingshuk Dasgupta, Rasmus Larsen, Yicheng Wang, Manish Reddy Vuyyuru, Chong Jiang, Joana Ijazi, Kazuki Osawa, Celine Smith, Ramya Sree Boppana, Taylan Bilal, Yuma Koizumi, Ying Xu, Yasemin Altun, Nir Shabat, Ben Bariach, Alex Korchemniy, Kiam Choo, Olaf Ronneberger, Chimezie Iwuanyanwu, Shubin Zhao, David Soergel, Cho-Jui Hsieh, Irene Cai, Shariq Iqbal, Martin Sundermeyer, Zhe Chen, Elie Bursztein, Chaitanya Malaviya, Fadi Biadsy, Prakash Shroff, Inderjit Dhillon, Tejasi Latkar, Chris Dyer, Hannah Forbes, Massimo Nicosia, Vitaly Nikolaev, Somer Greene, Marin Georgiev, Pidong Wang, Nina Martin, Hanie Sedghi, John Zhang, Praseem Banzal, Doug Fritz, Vikram Rao, Xuezhi Wang, Jiageng Zhang, Viorica Patraucean, Dayou Du, Igor Mordatch, Ivan Jurin, Lewis Liu, Ayush Dubey, Abhi Mohan, Janek Nowakowski, Vlad-Doru Ion, Nan Wei, Reiko Tojo, Maria Abi Raad, Drew A. Hudson, Vaishakh Keshava, Shubham Agrawal, Kevin Ramirez, Zhichun Wu, Hoang Nguyen, Ji Liu, Madhavi Sewak, Bryce Petrini, DongHyun Choi, Ivan Philips, Ziyue Wang, Ioana Bica, Ankush Garg, Jarek Wilkiewicz, Priyanka Agrawal, Xiaowei Li, Danhao Guo, Emily Xue, Naseer Shaik, Andrew Leach, Sadh MNM Khan, Julia Wiesinger, Sammy Jerome, Abhishek Chakladar, Alek Wenjiao Wang, Tina Ornduff, Folake Abu, Alireza Ghaffarkhah, Marcus Wainwright, Mario Cortes, Frederick Liu, Joshua Maynez, Andreas Terzis, Pouya Saman-

gouei, Riham Mansour, Tomasz Kepa, Francois-Xavier Aubet, Anton Algymr, Dan Banica, Agoston Weisz, Andras Orban, Alexandre Senges, Ewa Andrejczuk, Mark Geller, Niccolo Dal Santo, Valentin Anklin, Majd Al Merey, Martin Baeuml, Trevor Strohman, Junwen Bai, Slav Petrov, Yonghui Wu, Demis Hassabis, Koray Kavukcuoglu, Jeffrey Dean, and Oriol Vinyals. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context, 2024.

[23] Vikrant Varma, Rohin Shah, Zachary Kenton, János Kramár, and Ramana Kumar. Explaining grokking through circuit efficiency. *arXiv preprint arXiv:2309.02390*, 2023.

[24] Johannes Von Oswald, Eyvind Niklasson, Ettore Randazzo, João Sacramento, Alexander Mordvintsev, Andrey Zhmoginov, and Max Vladymyrov. Transformers learn in-context by gradient descent. In *International Conference on Machine Learning*, pages 35151–35174. PMLR, 2023.

[25] Jason Wei, Yi Tay, Rishi Bommasani, Colin Raffel, Barret Zoph, Sebastian Borgeaud, Dani Yogatama, Maarten Bosma, Denny Zhou, Donald Metzler, et al. Emergent abilities of large language models. *TMLR*, 2022.

[26] Jian Xie, Kai Zhang, Jiangjie Chen, Tinghui Zhu, Renze Lou, Yuandong Tian, Yanghua Xiao, and Yu Su. Travelplanner: A benchmark for real-world planning with language agents, 2024.

[27] Tian Ye, Zicheng Xu, Yuanzhi Li, and Zeyuan Allen-Zhu. Physics of language models: Part 2.1, grade-school math and the hidden reasoning process. *arXiv preprint arXiv:2407.20311*, 2024.

[28] Gilad Yehudai, Haim Kaplan, Asma Ghandeharioun, Mor Geva, and Amir Globerson. When can transformers count to n? *arXiv preprint arXiv:2407.15160*, 2024.

[29] Ziqian Zhong, Ziming Liu, Max Tegmark, and Jacob Andreas. The clock and the pizza: Two stories in mechanistic explanation of neural networks. *Advances in Neural Information Processing Systems*, 36, 2024.

[30] Tianyi Zhou, Deqing Fu, Vatsal Sharan, and Robin Jia. Pre-trained large language models use fourier features to compute addition. *arXiv preprint arXiv:2406.03445*, 2024.

A Constructing global optimizers

As mentioned in the main text, we find a mechanism to construct global optimizers from partial solutions that only make a subset of terms vanish in the loss function. This motivates us to find the "seed" solutions that satisfy individual constraints (MPs) in the loss, and then combine them. For this, we group MPs from the loss (Eqn. 3) into three types of constraints. Next, we discuss the partial solutions that satisfy a subset of them, which can be combined to obtain global optimizers.

Definition 6 (Sets of Constraints)**.** *Four sets of constraints exist in MSE loss (Eqn. 3):*

- *The main term constraints* $R_+ := \{\boldsymbol{z} | r_{kkk}(\boldsymbol{z}) = 1/2d\}$*;*
- *The cross term constraints* $R_\mathrm{c} := \{\boldsymbol{z} | r_{k_1 k_2 k}(\boldsymbol{z}) = 0 \text{ except for } k_1 = k_2 = k\}$*;*
- *The norm constrains* $R_\mathrm{n} := \{\boldsymbol{z} | r_{p0k'k}(\boldsymbol{z}) = \sum_j |z_{pk'j}|^2 z_{ckj} = 0\}$*;*
- *The circular convolution constraints* $R_\circledast = \{\boldsymbol{z} | r_{pmk'k}(\boldsymbol{z}) = 0 \text{ for } m \neq 0\}$.

A.1 Global Optimizers leveraging Fourier Bases

We first consider the case that the solution is only nonzero at frequency k_0 but not others, i.e., $z_{pkj} = 0$ for $k \neq \pm k_0$. Such solution corresponds to Fourier bases in the original domain.

Lemma 2 (Solutions satisfying R_c)**.** *All order-1 or order-2 solutions satisfying R_c must have $r_{kkk} = 0$ for all k. With small L_2 regularization, all order-3 solutions can be decomposed into $\boldsymbol{z} = \tilde{\boldsymbol{z}}_{k_0} * \boldsymbol{y}$ for certain frequency k_0, where $\tilde{\boldsymbol{z}}_{k_0} = \{\tilde{z}_{pkj}\}$ has order 3 and corresponds to Fourier bases in the original domain:*

$$\tilde{z}_{pk_0\cdot} = [1, \omega_3, \omega_3^2]/\sqrt[3]{6d} \tag{5}$$

where $\omega_3 := e^{-2\pi\mathrm{i}/3}$ and $\boldsymbol{y}$ is a order-1 unit.

Note that by simple calculation, $\tilde{\boldsymbol{z}}_{k_0} \in R_\mathrm{n}$ but $\tilde{\boldsymbol{z}}_{k_0} \notin R_\circledast$. Fortunately, leveraging the property of ring homomorphism, we can construct another solution $\boldsymbol{z}'_{k_0} \in R_\circledast$ of order-2, and they combined to form global optimizers.

Corollary 2 (Order-6 global optimizers of Eqn. 3)**.** *The following* "3×2" *Fourier solutions satisfies the global optimality condition (Eqn. 4):*

$$\boldsymbol{z}_{F6} = \sum_{k=1}^{(d-1)/2} \tilde{\boldsymbol{z}}_k * \boldsymbol{z}'_k * \boldsymbol{y}_k \tag{6}$$

where $\boldsymbol{z}'_k$ is order-2 (see Proof). As a result, $\mathrm{ord}(\boldsymbol{z}_{F6}) = 3 \cdot 2 \cdot 1 \cdot (d-1)/2 = 3(d-1)$ *and each frequency is affiliated with 6 hidden nodes (order-6).*

Fig. 2 shows a case with $d = 7$. In this case, each frequency, out of $(d-1)/2 = 3$ total number of frequencies, is associated with 6 hidden nodes. If we remove the last term in the loss that corresponds to constraints $R_\circledast$, then an order-3 solution suffices.

Interestingly, there also exists a lower-order solution, 2×2, which involves $\omega_8 := e^{-\pi\mathrm{i}/4}$, that meets R_c and $R_\circledast$ but not R_n:

Corollary 3 (Order-4 "almost" global optimizers)**.** *The following order-2 solution satisfies R_c except for $r_{k_0,k_0,-k_0} = 0$, $R_\circledast$ and $r_{k_0k_0k_0} = 1/\sqrt{2d}$:*

$$z_{ak_0\cdot} = [1, \bar{\omega}_8^2]/\sqrt{2}, \quad z_{bk_0\cdot} = [\bar{\omega}_8, \omega_8]/\sqrt{2}, \quad z_{ck_0\cdot} = [\omega_8, \omega_8]/\sqrt{2d} \tag{7}$$

and the following order-2 solution satisfies $r_{k_0,k_0,-k_0} = 0$ and $r_{k_0k_0k_0} = 1/\sqrt{2d}$:

$$z_{ak_0\cdot} = [1, \omega_8]/\sqrt{2}, \quad z_{bk_0\cdot} = [\omega_8, \bar{\omega}_8^2]/\sqrt{2}, \quad z_{ck_0\cdot} = [\bar{\omega}_8, \omega_8]/\sqrt{2d} \tag{8}$$

Therefore, their product $\boldsymbol{z}_{F4}$, an "2×2" *order-4 solution satisfies both R_c and $R_\circledast$.*

Note that this solution is perceived in the experiments, in particular for larger scale problems, showing a strong preference of gradient descent towards lower order solutions.

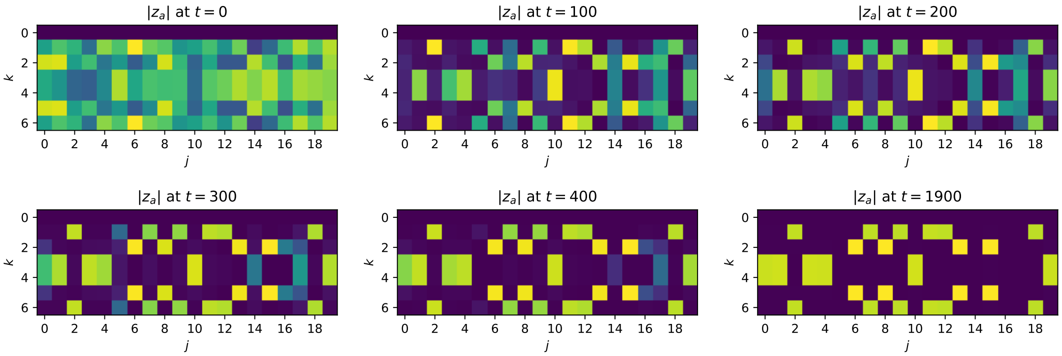

Figure 3: The convergence path of $z_{a\cdot\cdot}$ when training modular addition using Adam optimizer (learning rate 0.05, weight decay 0.005). The final solution contains 2 order-6 ($\boldsymbol{z}_{F6}$) and 1 order-4 ($\boldsymbol{z}_{F4}$) solutions. For each hidden node j, once a dominant frequency emerges, others fade away.

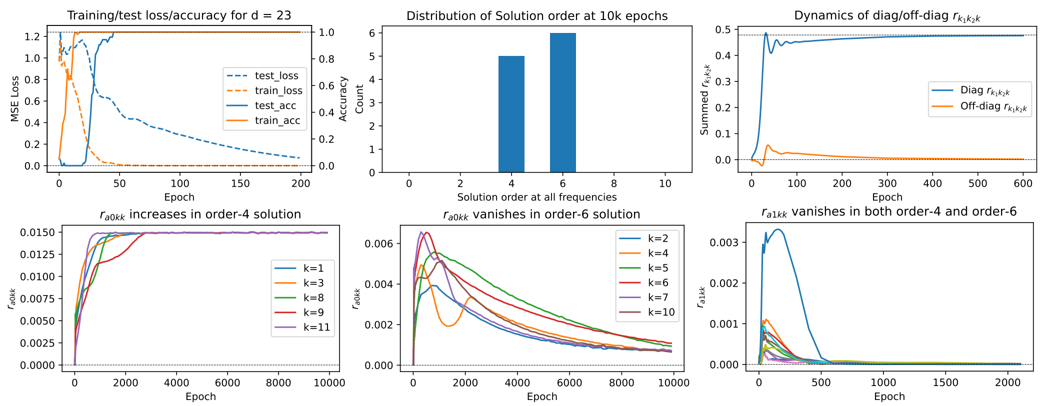

Figure 4: Dynamics of monomial potentials (MPs) over the training process for modular addition with $d = 23$ and $q = 1024$ hidden nodes. **Top Row.** *Left*: Training/test accuracy reaches 100% and loss close to 0. Test accuracy jumps after training reaches 100% (grokking). *Mid*: After 5k epochs, the distribution of solution orders are concentrated at 4 and 6 (Corollary 2,3). *Right*: Dynamics of $r_{k_1k_2k}$. Summation of diagonal r_{kkk} converges towards $(d-1)/2d$ (dotted line) with ripple effects, while off-diagonal $r_{k_1k_2k}$ converges towards 0. **Bottom Row.** Dynamics of different MPs. Note that order-4 and order-6 solutions have very different behaviors on r_{a0kk} (similar for r_{b0kk}).

A.2 Global Optimizers using Pure Memorization

We can also construct perfect memorization solutions as follows.

Corollary 4 (Perfect Memorization)**.** *Construct the following two d-order weights $\boldsymbol{z}_a$ and $\boldsymbol{z}_b$. Specifically, for $0 \le j < d$ and $k \neq 0$:*

$$z^{(a)}_{akj} = \omega_d^{kj}/\sqrt{d}, \qquad z^{(a)}_{bkj} = 1/\sqrt{d}, \qquad z^{(a)}_{ckj} = \omega_d^{-kj}/\sqrt{2d} \tag{9}$$

$$z^{(b)}_{bkj} = 1/\sqrt{d}, \qquad z^{(b)}_{akj} = \omega_d^{kj}/\sqrt{d}, \qquad z^{(b)}_{ckj} = \omega_d^{-kj}/\sqrt{2d} \tag{10}$$

where $\omega_d := e^{-2\pi \mathrm{i}/d}$ is the d-th root of unity. Here $\boldsymbol{z}_a \in R_{\mathrm{c}}(k_1 \neq k) \cap R_{\mathrm{n}} \cap R_{\circledast}(p = b \text{ or } m \neq k)$, $\boldsymbol{z}_b \in R_{\mathrm{c}}(k_2 \neq k) \cap R_{\mathrm{n}} \cap R_{\circledast}(p = a \text{ or } m \neq k)$. Then $\boldsymbol{z}_M = \boldsymbol{z}_a * \boldsymbol{z}_b$ satisfies the global optimality condition (Eqn. 4) and is the perfect memorization solution with $\mathrm{ord}(\boldsymbol{z}_M) = d^2$:

$$z^{(M)}_{akj_1j_2} = \omega^{kj_1}/d, \qquad z^{(M)}_{bkj_1j_2} = \omega^{kj_2}/d, \qquad z^{(M)}_{ckj_1j_2} = \omega^{-k(j_1+j_2)}/2d \tag{11}$$

where each hidden node is indexed by $j = (j_1, j_2)$, $0 \le j_1, j_2 < d$, $k \neq 0$.

To see why this corresponds to perfect memorization, simply apply an inverse Fourier transform for each hidden node (j_1, j_2), and the original weights are (zero-mean) delta function located at j_1, j_2 and $j_1 + j_2$ accordingly.

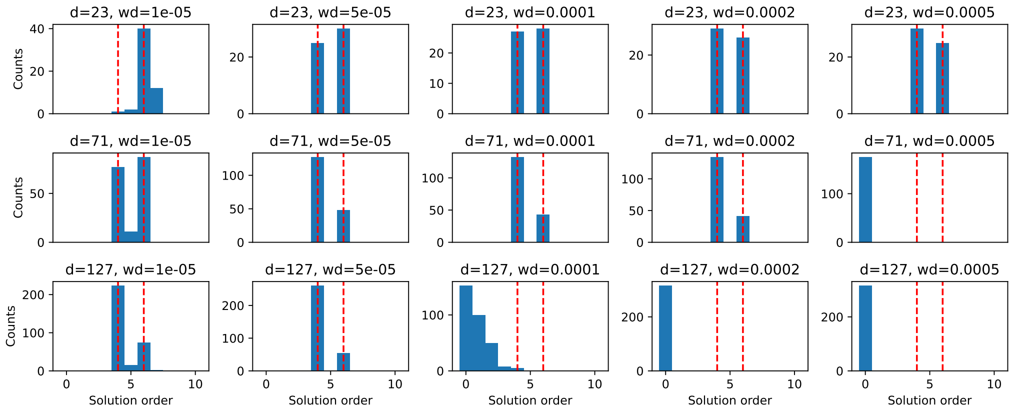

Figure 5: Solution distribution over different weight decay regularization for $q = 512$, trained with 10k epochs with Adams with learning rate 0.01 on modular addition (i.e., predicting $a+b \mod d$) with $d \in \{23, 71, 127\}$. The two red dashed lines correspond to order-4/6 solutions. The histogram is accumulated over 5 random seeds. While heavily over-parameterized (in particular for small d), final solution order remains constant, consistent with Corollary 1. Heavy weight decay shifts the distribution to the left (i.e., low-order solutions) until model collapsing, consistent with Theorem 5.

B Gradient dynamics

Now we have characterized the structures of global optimizers. One natural question arises: why the optimization procedure does not converge to the perfect memorization solution $\boldsymbol{z}_M$, but to the Fourier solutions $\boldsymbol{z}_{F6}$ and $\boldsymbol{z}_{F4}$? The answer is given by gradient dynamics.

Let $\boldsymbol{r} = [r_{k_1k_2k}, r_{pmk'k}] \in \mathbb{C}^{4d^3}$ be a vector of all MPs, and $J := \frac{\partial \boldsymbol{r}}{\partial \boldsymbol{z}}\frac{\partial \boldsymbol{z}}{\partial \mathcal{W}}$ be the Jacobian matrix of the mapping $\boldsymbol{r} = \boldsymbol{r}(\boldsymbol{z}(\mathcal{W}))$ in which $\mathcal{W}$ is the collection of original weights. Note that when we take derivatives with respect to r and apply chain rules, we treat r and its complex conjugate (e.g., r_{kkk} and $r_{-k,-k,-k} = \bar{r}_{kkk}$) as independent variables.

Since we run the gradient descent on $\mathcal{W}$, will such (indirect) optimization leads to a descent of $\boldsymbol{r}$ towards the desired targets (Eqn. 4)? This is confirmed by the following theorem:

Theorem 4 (Dynamics of MPs). *The dynamics of MPs satisfies $\dot{\boldsymbol{r}} = -JJ^*\overline{\nabla_{\boldsymbol{r}}\ell}$, which has positive inner product with the negative gradient direction $-\overline{\nabla_{\boldsymbol{r}}\ell}$.*

Corollary 1 shows that by ring multiplication, we could create infinitely many global optima from a base one. The following theorem answers which solution gradient dynamics picks.

Theorem 5 (The Occam's Razer: Preference of low-order solutions). *If $\boldsymbol{z} = \boldsymbol{y} * \boldsymbol{z}'$ and both $\boldsymbol{z}$ (of order q) and $\boldsymbol{z}'$ are global optimal solutions, then there exists a path of zero loss connecting $\boldsymbol{z}$ and $\boldsymbol{z}'$ in the space of $\mathcal{Z}_q$. As a result, lower-order solutions are preferred if trained with L_2 regularization.*

This shows that gradient dynamics with weight decay will pick a lower-order (i.e., simpler) solution. Fig. 5 verifies it with experiments.

The following theorem shows that the dynamics also enjoys *asymptotic freedom*:

Theorem 6 (Infinite Width Limits at Initialization). *Considering the modified loss of Eqn. 3 with only the first two terms: $\tilde{\ell}_k := r_{kkk} + d\sum_{k_1k_2}|r_{k_1k_2k}|^2$, if the weights are i.i.d Gaussian and network width $q \to +\infty$, then JJ^* converge to diagonal and the dynamics of MPs is decoupled.*

Intuitively, this means that a large enough network width ($q \to +\infty$) makes the dynamics much easier to analyze, while the final solution may not require that large M. As analyzed in Corollary 2, for each frequency, to achieve global optimality, only 6 hidden nodes are needed.

Ripple effects. While Theorem 6 only holds at initialization, the resulting decoupled MP dynamics, e.g., $\mathrm{d}r_{kkk}/\mathrm{d}t = 1 - 2dr_{kkk}$ that leads to $r_{kkk}(t) = (1 - e^{-t})/2d$, already captures the rough shape of the curve (Fig. 4 top right). To capture its fine structures (e.g., ripples before stabilization), we can also model the dynamics of the diagonal element in JJ^*. Consider a symmetric 1D case on a fixed frequency k, where all diagonal $r_{kkk} = r_0 - r$ (where $r_0 = 1/2d$) and all off-diagonal $r_{k_1k_2k} = r$, then

$$\dot{r} = -\dot{r}_{kkk} = \kappa(r_{kkk} - r_0) = -\kappa r, \quad \dot{\kappa} = \alpha(r_0 - r_{kkk}) - (1-\alpha) r_{k_1 k_2 k} - c_0 = (2\alpha - 1) r - c_0 \quad (12)$$

where $\kappa > 0$ is the diagonal element of JJ^* and α is a coefficient that characterizes the relative strength of two negative gradient $-\overline{\nabla_{r_{kkk}} \ell} = r_0 - r_{kkk}$ and $-\overline{\nabla_{r_{k_1 k_2 k}} \ell} = -r_{k_1 k_2 k}$, and c_0 is the gradient terms caused by asymmetry and/or other frequencies. This yields a second-order ODE that has complex roots in the characteristic function when $c_0 > 0$.

C Conclusion and future work

In this work, we propose CaGO (*Crafting Global Optimizers*), a theoretical framework that models the algebraic structure of global optimizers when training a 2-layer network on reasoning tasks of Abelian group with L_2 loss. We find that the global optimizers can be algebraically composited (i.e., "crafted") by non-optimal partial solutions that only fit to parts of the loss, using ring operations defined in the solution space of the 2-layer neural networks across different network widths. Our constructed solutions (i.e., $\boldsymbol{z}_{F4}$ and $\boldsymbol{z}_{F6}$, see Corollary 3 and Corollary 2) are verified in modular addition tasks. Under CaGO, we also analyze the training dynamics, show the benefit of over-parameterization, and the inductive bias towards simpler solutions due to topological connectivity between algebraically linked high-order (i.e., involving more hidden nodes) and low-order global optimizers.

Develop novel training algorithms. Our analysis suggests that instead of applying (stochastic) gradient descent to a greatly overparameterized network, we may be able to decompose the loss, construct low-order solutions and combine them to achieve the final solutions on the fly using algebraic operations. Such an approach may be more efficient (it takes a long time to get model training converged), and more scalable than a holistic end2end approach using gradient descent, due to its factorizable nature. Also our framework works for any loss function that is a combination of monomial potentials (L_2 loss is just one example), which opens a new dimension for loss function design.

Putting different widths into the same framework. Many existing theoretical works often assume that the network has a fixed width. However, our study demonstrates that nice mathematical structures can emerge when we consider networks of different widths together, which can be an interesting direction to consider in the future work.

Grokking. When learning modular addition, there exists a phase transition from *memorization* to *generalization* during training, known as *grokking* [23, 20], long after the training performance becomes (almost) perfect. While our work focuses more on what representation is learned on a uniform training data distribution, by applying it to different data distribution, grokking can be studied.

Extension to other activation functions. One key assumption of our approach is that the activation function is quadratic. For other activation functions (e.g., SiLU) with $\sigma(0) = 0$, we can do a Taylor expansion around the origin and the same framework can still apply (with higher rank MPs).

D Decoupling L_2 Loss (Proof)

We use the *character function* $\phi : G \to \mathbb{C}$, which maps a group element g into a complex number.

Lemma 3. *For finite Abelian group, the character function ϕ has the following properties [7, 21]:*

- *It is a 1-dimensional (irreducible) representation of the group G, i.e., $|\phi(g)| = 1$ for $g \in G$ and for any $g_1, g_2 \in G$, $\phi(g_1 g_2) = \phi(g_1)\phi(g_2)$.*
- *There exists d character functions $\{\phi_k\}$ that satisfy the orthonormal condition $\frac{1}{d}\sum_{g\in G}\phi_k(g)\overline{\phi_{k'}}(g) = \mathbb{I}(k = k')$. Here $\overline{\phi}$ is the complex conjugate of ϕ and is also a character function.*
- *The set of character functions $\{\phi_k\}$ forms a* character group $\hat{G}$ *under pairwise multiplication: $\phi_{k_1+k_2} = \phi_{k_1} \circ \phi_{k_2}$.*

Note that the *frequency* k goes from 0 to $d-1$, where $\phi_0 \equiv 1$ is the trivial representation (i.e., all $g \in G$ maps to 1). According to the Fundamental Theorem of Finite Abelian Groups, each finite Abelian group can be decomposed into a direct sum of cyclic groups, and the character function of each cyclic group is exactly (scaled) Fourier bases. Therefore, in Abelian group, k is a multi-dimensional frequency index. [3] shows that $\hat{G} \cong G$ (Theorem 3.13) so each character function $\phi \in \hat{G}$ can also be indexed by g itself. Right now we keep the index k.

For convenience, we define $\phi_{-k} := \overline{\phi}_k$ as the conjugate representation of ϕ_k.

Let $\boldsymbol{\phi}_k = [\phi_k(g)]_{g\in G} \in \mathbb{C}^d$ be the vector that contains the value of the character function ϕ_k. Then $\{\boldsymbol{\phi}_k\}$ form an orthogonal base in $\mathbb{C}^d$ and we can represent the weight vector $\mathbf{w}_j$ and $\boldsymbol{v}_j$ as the following:

$$\mathbf{w}_j = U_{G_1}\sum_{k\neq 0} z_{akj}\boldsymbol{\phi}_k + U_{G_2}\sum_{k\neq 0} z_{bkj}\boldsymbol{\phi}_k, \qquad \boldsymbol{v}_j = \sum_{k\neq 0} z_{ckj}\bar{\boldsymbol{\phi}}_k \tag{13}$$

where $\boldsymbol{z} := \{z_{pkj}\}$ are the complex coefficients ($p \in \{a, b, c\}$, $0 \le k < d$ and j runs through hidden nodes). Then it is clear that $\mathbf{w}_j^\top \boldsymbol{f}[i] = \sum_{k\neq 0} h_{akj}\phi_k(\iota_0(g[i])) + \sum_{k\neq 0} h_{bkj}\phi_k(x[i])$.

Theorem 1 (Analytic form of L_2 loss with quadratic activation)**.** *The objective of 2-layer MLP network with quadratic activation can be written as $\ell = \sum_{k\neq 0}\ell_k + (d-1)/d$, where*

$$\ell_k = -4r_{kkk} + 4d\sum_{k_1k_2}|r_{k_1k_2k}|^2 + d\Big|\sum_{p\in\{a,b\}}\sum_{k'} r_{p0k'k}\Big|^2 + d\sum_{m\neq 0}\sum_{p\in\{a,b\}}\Big|\sum_{k'} r_{pmk'k}\Big|^2 \tag{3}$$

Here $r_{k_1k_2k} := \sum_j z_{ak_1j}z_{bk_2j}z_{ckj}$ and $r_{pmk'k} := \sum_j z_{pk'j}z_{p,m-k',j}z_{ckj}$.

Proof. Note that the objective ℓ can be written down as

$$\begin{aligned}\ell &= \mathbb{E}_{g,x}\left[\|P_1^\perp(\boldsymbol{o}(g,x) - \boldsymbol{e}_{gx})\|^2\right] &(14)\\ &= \mathbb{E}_{g,x}\left[\boldsymbol{o}^\top P_1^\perp\boldsymbol{o} - 2\boldsymbol{o}^\top P_1^\perp\boldsymbol{e}_{gx} + \boldsymbol{e}_{gx}^\top P_1^\perp\boldsymbol{e}_{gx}\right] &(15)\end{aligned}$$

For $\mathbb{E}\left[\boldsymbol{o}^\top P_1^\perp\boldsymbol{e}_{gx}\right]$, since

$$\begin{aligned}\boldsymbol{e}_{gx}^\top P_1^\perp\boldsymbol{o} &= \sum_j \boldsymbol{e}_{gx}^\top P_1^\perp\boldsymbol{v}_j\sigma(\mathbf{w}_j^\top\boldsymbol{f}(g,x)) &(16)\\ &= \sum_j\left(\sum_{k'\neq 0} c_{k'j}\bar{\phi}_{k'}(gx)\right)\left(\sum_k a_{kj}\phi_k(\iota_0(g)) + b_{kj}\phi_k(x) + \boldsymbol{e}_g^\top\mathbf{w}_j^\perp\right)^2 &(17)\end{aligned}$$

Note that by our previous analysis, there exists $y_1 := \iota_0(g)$ so that $gy = x_1y$. Let $x_2 := x$. For notation brevity, let $z_{akj} := a_{kj}$, $z_{bkj} := b_{kj}$ and $z_{ckj} := c_{kj}$,, then we have:

$$\boldsymbol{e}_{gx}^\top P_1^\perp\boldsymbol{o} = \sum_j\left(\sum_{k'\neq 0} c_{k'j}\bar{\phi}_{k'}(x_1x_2)\right)\left(\sum_k\sum_p z_{pkj}\phi_k(x_p) + \boldsymbol{e}_{x_1}^\top\mathbf{w}_j^\perp\right)^2 \tag{18}$$

Therefore, we have:

$$\mathbb{E}_{g,x}\left[\boldsymbol{e}_{gx}^\top P_1^\perp \boldsymbol{o}\right] = \sum_{k_1,k_2,k'\neq 0,p_1,p_2,j} c_{k'j} z_{p_1k_1j} z_{p_2k_2j} \mathbb{E}\left[\bar\phi_{k'}(x_1)\bar\phi_{k'}(x_2)\phi_{k_1}(x_{p_1})\phi_{k_2}(x_{p_2})\right] \tag{19}$$

Note that due to the fact that $\mathbb{E}_{g\in\iota_0^{-1}(x_1)}\left[\boldsymbol{e}_g^\top \mathbf{w}_j^\perp\right] = 0$ and $\mathbb{E}_{g\in\iota_0^{-1}(x_1)}\left[\boldsymbol{e}_g\boldsymbol{e}_g^\top\right]$ is only a function of x_1 and becomes 0 if multiplied with $\sum_{k'\neq 0} c_{k'j}\bar\phi_{k'}(x_1x_2)$ and taking expectation w.r.t x_2, in the final expression, all terms involving $\mathbf{w}_j^\perp$ vanish.

Since $\mathbb{E}_x\left[\phi_k(x)\bar\phi_{k'}(x)\right] = \mathbb{I}(k=k')$, there are only a few cases that the summand is nonzero:

- $p_1 = 1, p_2 = 2, k' = k_1 = k_2 \neq 0$.
- $p_1 = 2, p_2 = 1, k' = k_1 = k_2 \neq 0$.

In both cases, the summation reduces to $\sum_{k\neq 0,j} c_{kj}z_{1kj}z_{2kj} = \sum_{k\neq 0,j} c_{kj}a_{kj}b_{kj}$. Let $r_{k_1k_2k'} := \sum_j a_{k_1j}b_{k_2j}c_{k'j}$, then we have

$$\mathbb{E}\left[\boldsymbol{o}^\top P_1^\perp \boldsymbol{e}_{gy}\right] = 2\sum_{k\neq 0,j} a_{kj}b_{kj}c_{kj} = 2\sum_{k\neq 0} x_{kkk} \tag{20}$$

For $\mathbb{E}\left[\boldsymbol{o}^\top P_1^\perp \boldsymbol{o}\right]$, if $\mathbf{w}_j^\perp = 0$, then we have:

$$\boldsymbol{o}^\top P_1^\perp \boldsymbol{o} = \sum_{j,j'} \boldsymbol{v}_j^\top P_1^\perp \boldsymbol{v}_{j'}\sigma(\mathbf{w}_j^\top \boldsymbol{f}(g,y))\sigma(\mathbf{w}_{j'}^\top \boldsymbol{f}(g,y)) \tag{21}$$

here

$$\boldsymbol{v}_j^\top P_1^\perp \boldsymbol{v}_{j'} = \left(\sum_{k'\neq 0} c_{k'j}\bar{\boldsymbol{\phi}}_{k'}\right)^\top \left(\sum_{k''\neq 0} \bar c_{k''j'}\boldsymbol{\phi}_{k''}\right) = d\sum_{k'\neq 0} c_{k'j}\bar c_{k'j'} \tag{22}$$

due to the fact that $\bar{\boldsymbol{\phi}}_k^\top \boldsymbol{\phi}_{k'} = \sum_y \bar\phi_k(y)\phi_{k'}(y) = d\mathbb{I}(k=k')$.

Then the key part is to compute the following terms:

$$\mathbb{E}_{y_1,y_2}\left[z_{p_1k_1j_1}z_{p_2k_2j_1}z_{p_3k_3j_2}z_{p_4k_4j_2}c_{k'j_1}\bar c_{k'j_2}\phi_{k_1}(y_{p_1})\phi_{k_2}(y_{p_2})\phi_{k_3}(y_{p_3})\phi_{k_4}(y_{p_3})\right] \tag{23}$$

summing over $\{p_1,p_2,p_3,p_4,k_1,k_2,k_3,k_4,k'\neq 0, j_1, j_2\}$. Note that since each $p\in\{a,b\}$, there are $2^4 = 16$ choices of (p_1,p_2,p_3,p_4). For notation brevity, we use $(1,3)$ to represent the subset of p that takes the value of a (e.g., $(1,3)$ means that $p_1 = p_3 = a$ and $p_2 = p_4 = b$). It is clear that for odd assignments such as $(1,2,3)$, since $z_{p0j} = 0$, the summation is zero. Then, we only discuss the even cases as follows:

Case 1: $(1,3)$**,** $(2,4)$**,** $(1,4)$**,** $(2,3)$. The 4 cases are identical so we only need to analyze one. We take $(1,3)$ as an example. For $(1,3)$, $p_1 = p_3 = a$, $p_2 = p_4 = b$ and the only nonzero terms is when $k_1 + k_3 = 0 \mod d$, $k_2 + k_4 = 0 \mod d$, since $\mathbb{E}_{y_1}\left[\phi_{k_1}(y_1)\phi_{k_3}(y_1)\right] = \mathbb{I}(k_1+k_3 = 0 \mod d)$ (and similar in other cases). Then Eqn. 23 becomes:

$$\sum_{k_1,k_2,k'\neq 0}\sum_{j_1j_2} z_{ak_1j_1}z_{bk_2j_1}z_{a,-k_1,j_2}z_{b,-k_2,j_2}c_{k'j_1}\bar c_{k'j_2} \tag{24}$$

$$= \sum_{k_1,k_2,k'\neq 0}\sum_{j_1} z_{ak_1j_1}z_{bk_2j_1}c_{k'j_1}\overline{\sum_{j_2} z_{ak_1j_2}z_{bk_2j_2}c_{k'j_2}} \tag{25}$$

$$= \sum_{k_1,k_2,k'\neq 0}\sum_{j_1} a_{k_1j_1}b_{k_2j_1}c_{k'j_1}\overline{\sum_{j_2} a_{k_1j_2}b_{k_2j_2}c_{k'j_2}} \tag{26}$$

$$= \sum_{k_1,k_2,k'\neq 0} r_{k_1k_2k'}\overline{r_{k_1k_2k'}} = \sum_{k_1,k_2,k'\neq 0} |r_{k_1k_2k'}|^2 \tag{27}$$

Since there are 4 such cases, we have:

$$\epsilon_1 = 4\sum_{k'\neq 0}\sum_{k_1k_2} |r_{k_1k_2k'}|^2 \tag{28}$$

Case 2: $(1,2)$ **and** $(3,4)$. The two cases are identical. Take $(1,2)$ as an example. In this case, $p_1 = p_2 = a$ and $p_3 = p_4 = b$. The only non-zero terms are when $k_1 + k_2 = 0$, $k_3 + k_4 = 0$. Then Eqn. 23 becomes:

$$\sum_{k_1,k_3,k'\neq 0} \sum_{j_1 j_2} z_{ak_1j_1} \bar{z}_{ak_1j_1} z_{bk_3j_2} \bar{z}_{bk_3j_2} c_{k'j_1} \bar{c}_{k'j_2} \tag{29}$$

$$= \sum_{k_1,k_3,k'\neq 0} \sum_{j_1} |a_{k_1j_1}|^2 c_{k'j_1} \sum_{j_2} |b_{k_3j_2}|^2 \bar{c}_{k'j_2} \tag{30}$$

$$= \sum_{k'\neq 0} \left[\sum_{j_1} \left(\sum_{k_1} |a_{k_1j_1}|^2 \right) c_{k'j_1} \right] \left[\sum_{j_2} \left(\sum_{k_3} |b_{k_3j_2}|^2 \right) \bar{c}_{k'j_2} \right] \tag{31}$$

Let $r^{\circledast}_{amk'} := \sum_j \left(\sum_{k_1+k_2=m} a_{k_1j} a_{k_2j} \right) c_{k'j}$ (similar for $r^{\circledast}_{bmk'}$), then the above becomes $\sum_{k'\neq 0} r^{\circledast}_{a0k'} \bar{r}^{\circledast}_{b0k'}$.

Similarly, for $(3,4)$, the above equation becomes $\sum_{k'\neq 0} \bar{r}^{\circledast}_{a0k'} r^{\circledast}_{b0k'}$. Therefore, we have:

$$\epsilon_2 = \sum_{k'\neq 0} r^{\circledast}_{a0k'} \bar{r}^{\circledast}_{b0k'} + \bar{r}^{\circledast}_{a0k'} r^{\circledast}_{b0k'} \tag{32}$$

Note that this term can be negative. However, we will see that when it is combined with the following terms, all terms will be non-negative.

Case 3: $(1,2,3,4)$ **and** $()$. In this case we have:

$$\sum_{k'\neq 0} \sum_{j_1 j_2} \sum_{p\in\{1,2\}} \sum_{k_1+k_2+k_3+k_4=0} z_{pk_1j_1} z_{pk_2j_1} z_{pk_3j_2} z_{pk_4j_2} c_{k'j_1} \bar{c}_{k'j_2} \tag{33}$$

$$= \sum_{k'\neq 0} \sum_{j_1 j_2} \sum_{p\in\{1,2\}} \sum_{k_1+k_2=k_3+k_4} z_{pk_1j_1} z_{pk_2j_1} \bar{z}_{pk_3j_2} \bar{z}_{pk_4j_2} c_{k'j_1} \bar{c}_{k'j_2} \tag{34}$$

$$= \sum_{k'\neq 0} \sum_{m} \sum_{p\in\{1,2\}} \sum_{j_1 j_2} \sum_{p\in\{1,2\}} \sum_{k_1+k_2=m} \sum_{k_3+k_4=m} z_{pk_1j_1} z_{pk_2j_1} \bar{z}_{pk_3j_2} \bar{z}_{pk_4j_2} c_{k'j_1} \bar{c}_{k'j_2} \tag{35}$$

$$= \sum_{k'\neq 0} \sum_{m} \sum_{p\in\{1,2\}} \left[\sum_{j_1} \left(\sum_{k_1+k_2=m} z_{pk_1j_1} z_{pk_2j_1} \right) c_{k'j_1} \right] \left[\sum_{j_2} \left(\sum_{k_3+k_4=m} \overline{z_{pk_3j_2} z_{pk_4j_2}} \right) \bar{c}_{k'j_2} \right]$$

$$= \sum_{k'\neq 0} \sum_{m} |r^{\circledast}_{amk'}|^2 + |r^{\circledast}_{bmk'}|^2 \tag{36}$$

In particular, when $m = 0$, we have $\sum_{k'\neq 0} |r^{\circledast}_{a0k'}|^2 + |r^{\circledast}_{b0k'}|^2$. Therefore, we have

$$\epsilon_2 + \epsilon_{3,m=0} = \sum_{k'\neq 0} |r^{\circledast}_{a0k'} + r^{\circledast}_{b0k'}|^2 \tag{37}$$

Finally, putting them together, we have:

$$\mathbb{E}\left[\boldsymbol{o}^\top P_1^\perp \boldsymbol{o}\right] = d(\epsilon_1 + \epsilon_2 + \epsilon_3) = d(\epsilon_1 + (\epsilon_2 + \epsilon_{3,m=0}) + \epsilon_{3,m\neq 0}) \tag{38}$$

$$= d \sum_{k'\neq 0} \left(4 \sum_{k_1k_2} |r_{k_1k_2k'}|^2 + |r^{\circledast}_{a0k'} + r^{\circledast}_{b0k'}|^2 + \sum_{m\neq 0} |r^{\circledast}_{amk'}|^2 + |r^{\circledast}_{bmk'}|^2 \right)$$

$$\geq 0 \tag{39}$$

□

Lemma 1 (A Sufficient Conditions of Global optimizers of Eqn. 3)**.** *If a solution $\boldsymbol{z}$ to Eqn. 3 satisfies the following, then it is a global optimizer with zero loss $\ell(\boldsymbol{z}) = 0$.*

$$r_{kkk}(\boldsymbol{z}) = \mathbb{I}(k \neq 0)/2d, \quad r_{k_1k_2k}(\boldsymbol{z}) = 0, \quad r_{pmk'k}(\boldsymbol{z}) = 0 \tag{4}$$

Proof. Note that $d^{-1}\sum_k r_{kkk} - \sum_k |r_{kkk}|^2$ has a minimizer $r_{kkk} = 1/2d$. Therefore, the best loss value any assignment of weights is able to achieve is the following:

$$r_{k_1k_2k'} = \sum_j a_{k_1j}b_{k_2j}c_{k'j} = \frac{1}{2d}\mathbb{I}(k_1 = k_2 = k') \qquad k' \neq 0 \tag{40}$$

$$r^{\circledast}_{a0k'} + r^{\circledast}_{b0k'} := \sum_j \left(\sum_k |a_{kj}|^2 + |b_{kj}|^2\right) c_{k'j} = 0 \qquad k' \neq 0 \tag{41}$$

$$r^{\circledast}_{amk'} := \sum_j \left(\sum_{k_1+k_2=m} a_{k_1j}a_{k_2j}\right) c_{k'j} = 0 \qquad k' \neq 0, m \neq 0 \tag{42}$$

$$r^{\circledast}_{bmk'} := \sum_j \left(\sum_{k_1+k_2=m} b_{k_1j}b_{k_2j}\right) c_{k'j} = 0 \qquad k' \neq 0, m \neq 0 \tag{43}$$

Therefore the sufficient conditions (Eqn. 4) will make all above come true. □

E Semi-ring structure of $\mathcal{Z}$ (Proof)

Theorem 2 (Algebraic Structure of $\mathcal{Z}$)**.** *$\langle \mathcal{Z}, +, * \rangle$ is a commutative semi-ring.*

Proof. Straightforward from the definition of addition and multiplication (Def. 3) and identification of hidden nodes under permutation (Def. 2). Note that ring addition (i.e., concatenation) does not have inverse and thus it is a semi-ring. □

Theorem 3. *For any monomial potential $r : \mathcal{Z} \mapsto \mathbb{C}$, $r(\mathbf{1}) = 1$, $r(\boldsymbol{z}_1 + \boldsymbol{z}_2) = r(\boldsymbol{z}_1) + r(\boldsymbol{z}_2)$ and $r(\boldsymbol{z}_1 * \boldsymbol{z}_2) = r(\boldsymbol{z}_1)r(\boldsymbol{z}_2)$ and thus r is a ring homomorphism.*

Proof. Let $r(\boldsymbol{z}) = \sum_j \prod_{(p,k)\in \mathrm{idx}(r)} z_{pkj}$. Since the ring identity $\mathbf{1}$ is order-1 and all $z_{pkj} = 1$, it is obvious that $r(\mathbf{1}) = 1$.

Let $\mathrm{supp}(\boldsymbol{z}_1)$ be the subset of the hidden nodes that corresponds to $\boldsymbol{z}_1$ in the concatenated solution $\boldsymbol{z}_1 + \boldsymbol{z}_2$, similar for $\mathrm{supp}(\boldsymbol{z}_2)$. Note that

$$r(\boldsymbol{z}_1 + \boldsymbol{z}_2) = \sum_{j\in \mathrm{supp}(\boldsymbol{z}_1)} \prod_{(p,k)\in \mathrm{idx}(r)} z^{(1)}_{pkj} + \sum_{j\in \mathrm{supp}(\boldsymbol{z}_2)} \prod_{(p,k)\in \mathrm{idx}(r)} z^{(2)}_{pkj} = r(\boldsymbol{z}_1) + r(\boldsymbol{z}_2) \tag{44}$$

On the other hand, we have

$$\begin{aligned}
r(\boldsymbol{z}_1 * \boldsymbol{z}_2) &= \sum_{j_1j_2} \prod_{(p,k)\in \mathrm{idx}(r)} \left(z^{(1)}_{pkj_1} z^{(2)}_{pkj_2}\right) && (45)\\
&= \sum_{j_1j_2} \left(\prod_{(p,k)\in \mathrm{idx}(r)} z^{(1)}_{pkj_1}\right) \left(\prod_{(p,k)\in \mathrm{idx}(r)} z^{(2)}_{pkj_2}\right) && (46)\\
&= \left(\sum_{j_1} \prod_{(p,k)\in \mathrm{idx}(r)} z^{(1)}_{pkj_1}\right) \left(\sum_{j_2} \prod_{(p,k)\in \mathrm{idx}(r)} z^{(1)}_{pkj_2}\right) && (47)\\
&= r(\boldsymbol{z}_1)r(\boldsymbol{z}_2) && (48)
\end{aligned}$$

□

Corollary 1. *If $\boldsymbol{z}$ is a global optimizer and $\boldsymbol{y}$ is a unit, then $\boldsymbol{z} * \boldsymbol{y}$ is also a global optimizer.*

Proof. Straightforward by leveraging the property of ring homomorphism. E.g.,

$$r_{kkk}(\boldsymbol{z} * \boldsymbol{y}) = r_{kkk}(\boldsymbol{z})r_{kkk}(\boldsymbol{y}) = r_{kkk}(\boldsymbol{z}) \tag{49}$$

and the proof is complete. □

F Solution Construction (Proof)

Lemma 2 (Solutions satisfying R_c)**.** *All order-1 or order-2 solutions satisfying R_c must have $r_{kkk} = 0$ for all k. With small L_2 regularization, all order-3 solutions can be decomposed into $\boldsymbol{z} = \tilde{\boldsymbol{z}}_{k_0} * \boldsymbol{y}$ for certain frequency k_0, where $\tilde{\boldsymbol{z}}_{k_0} = \{\tilde{z}_{pkj}\}$ has order 3 and corresponds to Fourier bases in the original domain:*

$$\tilde{z}_{pk_0\cdot} = [1, \omega_3, \omega_3^2]/\sqrt[3]{6d} \tag{5}$$

where $\omega_3 := e^{-2\pi \mathrm{i}/3}$ and $\boldsymbol{y}$ is a order-1 unit.

Proof. We first prove that $\tilde{\boldsymbol{z}}_{k_0}$ satisfies R_c. To see this, we have

$$r_{k_1k_2k} = \sum_j \mathbb{I}(k_1 = k_2 = k = k_0)\omega_3^{3j} + \sum_j \mathbb{I}(-k_1 = k_2 = k = k_0)\omega_3^{j} \tag{50}$$

$$+ \ldots + \sum_j \mathbb{I}(-k_1 = -k_2 = -k = k_0)\bar{\omega}_3^{3j} \tag{51}$$

$$= 3\mathbb{I}(k_1 = k_2 = k = k_0) + 3\mathbb{I}(k_1 = k_2 = k = -k_0) \tag{52}$$

Note that all cross terms are gone since $\sum_j \omega_3^j = 0$. It is clear that $r_{k_1k_2k} \neq 0$ unless $k_1 = k_2 = k$ so $\boldsymbol{z}_0$ satisfies R_c.

To show the reverse direction, first notice that for any order-1 solution, for any k, in order to make $r_{k,-k,k} = z_{ak0}z_{b,-k,0}z_{ck0} = z_{ak0}\bar{z}_{bk0}z_{ck0} = 0$, either z_{ak0}, z_{bk0} or z_{ck0} has to be zero, which means that $r_{kkk} = 0$.

For order-2, first of all if any $z_{pk0} = 0$ for any $p \in \{a, b, c\}$, then a constraint like $r_{k,k,-k} = z_{ak0}z_{bk0}\bar{z}_{ck0} + z_{ak1}z_{bk1}\bar{z}_{ck1} = 0$ yields $z_{ak1}z_{bk1}z_{ck1} = 0$ and thus $r_{kkk} = 0$. If not, then for any two complex numbers z_{pk0} and z_{pk1}, there always exist four real numbers $\theta_p \in (-\pi, \pi]$, $\theta'_p \in (-\pi, \pi]$, $m_{p0} > 0$ and $m_{p1} > 0$ so that

$$z_{pk0} = m_{p0}e^{\mathrm{i}\theta'_p}e^{\mathrm{i}\theta_p}, \qquad z_{pk1} = m_{p1}e^{\mathrm{i}\theta'_p}e^{-\mathrm{i}\theta_p} \tag{53}$$

Then a constraint like $r_{k,k,-k} = z_{ak0}z_{bk0}\bar{z}_{ck0} + z_{ak1}z_{bk1}\bar{z}_{ck1} = 0$ can be written as $z_{ak0}z_{bk0}\bar{z}_{ck0} = -z_{ak1}z_{bk1}\bar{z}_{ck1}$, or equivalently:

$$m_{a0}m_{b0}m_{c0}e^{\mathrm{i}(\theta'_a+\theta'_b+\theta'_c)}e^{\mathrm{i}(\theta_a+\theta_b-\theta_c)} = -m_{a1}m_{b1}m_{c1}e^{\mathrm{i}(\theta'_a+\theta'_b+\theta'_c)}e^{-\mathrm{i}(\theta_a+\theta_b-\theta_c)} \tag{54}$$

$$m_{a0}m_{b0}m_{c0}e^{\mathrm{i}\theta_a}e^{\mathrm{i}\theta_b}e^{-\mathrm{i}\theta_c} = -m_{a1}m_{b1}m_{c1}e^{-\mathrm{i}\theta_a}e^{-\mathrm{i}\theta_b}e^{\mathrm{i}\theta_c} \tag{55}$$

Comparing their magnitude and phase, we have $m_{a0}m_{b0}m_{c0} = m_{a1}m_{b1}m_{c1}$ and

$$\theta_a + \theta_b - \theta_c = \pm\pi/2 \mod 2\pi \tag{56}$$

Similarly, we have:

$$\theta_a + \theta_c - \theta_b = \pm\pi/2 \mod 2\pi, \qquad \theta_b + \theta_c - \theta_a = \pm\pi/2 \mod 2\pi \tag{57}$$

Solving the three equations and we have 6 solutions:

$$(\theta_a, \theta_b, \theta_c) = (0, 0, \pm\pi/2) \mod 2\pi \tag{58}$$

$$(\theta_a, \theta_b, \theta_c) = (0, \pm\pi/2, 0) \mod 2\pi \tag{59}$$

$$(\theta_a, \theta_b, \theta_c) = (\pm\pi/2, 0, 0) \mod 2\pi \tag{60}$$

For all such solutions, we have $r_{kkk} = 0$.

For order-3 solutions, for each k, let $a_j := z_{akj}$, $b_j := z_{bkj}$ and $c_j := z_{ckj}$. Let $\boldsymbol{a} = [a_j] \in \mathbb{C}^3$, $\boldsymbol{b} = [b_j] \in \mathbb{C}^3$ and $\boldsymbol{c} = [c_j] \in \mathbb{C}^3$. Then the conditions yield that

$$(\boldsymbol{a} \circ \bar{\boldsymbol{b}})^\top \boldsymbol{c} = 0, \quad (\boldsymbol{a} \circ \bar{\boldsymbol{b}})^\top \bar{\boldsymbol{c}} = 0, \quad (\bar{\boldsymbol{a}} \circ \boldsymbol{b})^\top \boldsymbol{c} = 0, \quad (\bar{\boldsymbol{a}} \circ \boldsymbol{b})^\top \bar{\boldsymbol{c}} = 0 \tag{61}$$

which means that in $\mathbb{R}^3$ space, the following condition holds:

$$\mathrm{span}(\Re(\boldsymbol{a} \circ \bar{\boldsymbol{b}}), \Im(\boldsymbol{a} \circ \bar{\boldsymbol{b}})) \perp \mathrm{span}(\Re(\boldsymbol{c}), \Im(\boldsymbol{c})) \tag{62}$$

where $\Re(\cdot)$ and $\Im(\cdot)$ are real and imaginary parts of a complex vector. Since Eqn. 62 holds in $\mathbb{R}^3$, it must be the case that either $\Re(\boldsymbol{a} \circ \bar{\boldsymbol{b}})$ is co-linear with $\Im(\boldsymbol{a} \circ \bar{\boldsymbol{b}})$, or $\Re(\boldsymbol{c})$ is co-linear with $\Im(\boldsymbol{c})$.

If the former is true (i.e., there exists β so that $\Re(\boldsymbol{c}) = \beta\Im(\boldsymbol{c})$), then there exists a scalar θ so that $\boldsymbol{c}e^{-\mathrm{i}\theta} = \boldsymbol{c}_R \in \mathbb{R}^3$, since all angles in the components of $\boldsymbol{c}$ are the same. Then we have:

$$r_{kkk} = (\boldsymbol{a} \circ \boldsymbol{b})^\top \boldsymbol{c} = (\boldsymbol{a} \circ \boldsymbol{b})^\top \bar{\boldsymbol{c}} e^{2\mathrm{i}\theta} = 0 \tag{63}$$

If the latter is true, then there exists $\theta_{a\bar{b}}$ so that

$$(\boldsymbol{a} \circ \bar{\boldsymbol{b}})e^{-\mathrm{i}\theta_{a\bar{b}}} \in \mathbb{R}^3_+ \tag{64}$$

Applying the same reasoning symmetrically, in order to find cases such that $r_{kkk} \neq 0$, a necessary condition is that

$$(\boldsymbol{a} \circ \bar{\boldsymbol{b}})e^{-\mathrm{i}\theta_{a\bar{b}}} \in \mathbb{R}^3_+, \quad (\boldsymbol{b} \circ \bar{\boldsymbol{c}})e^{-\mathrm{i}\theta_{b\bar{c}}} \in \mathbb{R}^3_+, \quad (\boldsymbol{c} \circ \bar{\boldsymbol{a}})e^{-\mathrm{i}\theta_{c\bar{a}}} \in \mathbb{R}^3_+ \tag{65}$$

with the condition that $\theta_{a\bar{b}} + \theta_{b\bar{c}} + \theta_{c\bar{a}} = 0 \mod 2\pi$. To determine these angles, we look at a_0, b_0 and c_0 and their angles θ_{a0}, θ_{b0}, and θ_{c0}, it is clear that

$$\theta_{a\bar{b}} = \theta_{a0} - \theta_{b0} \mod 2\pi \tag{66}$$
$$\theta_{b\bar{c}} = \theta_{b0} - \theta_{c0} \mod 2\pi \tag{67}$$
$$\theta_{c\bar{a}} = \theta_{c0} - \theta_{a0} \mod 2\pi \tag{68}$$

Therefore, if we multiple $\boldsymbol{a}$, $\boldsymbol{b}$ and $\boldsymbol{c}$ with $e^{-\mathrm{i}\theta_{a0}}$, $e^{-\mathrm{i}\theta_{b0}}$ and $e^{-\mathrm{i}\theta_{c0}}$, and still note the resulting vectors to be $\boldsymbol{a}$, $\boldsymbol{b}$ and $\boldsymbol{c}$, then we have:

$$\boldsymbol{a} \circ \bar{\boldsymbol{b}} \in \mathbb{R}^3_+, \quad \boldsymbol{b} \circ \bar{\boldsymbol{c}} \in \mathbb{R}^3_+, \quad \boldsymbol{c} \circ \bar{\boldsymbol{a}} \in \mathbb{R}^3_+ \tag{69}$$

Note that is equivalent to a decomposition of $\boldsymbol{z}$ into a multiplication of 1-order term and another 3-order term. Then we have $\theta_{a0} = \theta_{b0} = \theta_{c0} = \theta_0 = 0$, $\theta_{a1} = \theta_{b1} = \theta_{c1} = \theta_1$, $\theta_{a2} = \theta_{b2} = \theta_{c2} = \theta_2$.

Letting $m_j := |a_j||b_j||c_j|$, then the corresponding r_{kkk} can be written as:

$$r_{kkk} = \sum_{j=0}^{2} m_j e^{3\mathrm{i}\theta_j} \tag{70}$$

with the constraints that $\sum_{j=0}^2 m_j e^{\mathrm{i}\theta_j} = 0$ imposed by R_A. One interesting question is that what is the minimal norm representation that achieves the highest objective? For this we can solve the following optimization problem:

$$\max_{\{m_j,\theta_j\}} \sum_j m_j(e^{3\mathrm{i}\theta_j} + e^{-3\mathrm{i}\theta_j}) - \epsilon \sum_j m_j^2 \quad \text{s.t.} \sum_j m_j e^{\mathrm{i}\theta_j} = 0 \tag{71}$$

which achieves the maximal when $m_j = 1/\epsilon$, $\theta_1 = 2\pi j/3$ and $\theta_2 = 4\pi j/3$ (or vise versa). Note that θ_j is fixed no matter how small the regularization ϵ is.

To see that, let $u_j := e^{\mathrm{i}\theta_j}$. Then we have:

$$\sum_j m_j(u_j + \bar{u}_j)^3 = \sum_j m_j[u_j^3 + 3u_j\bar{u}_j(u_j + \bar{u}_j) + \bar{u}_j^3] = \sum_j m_j(u_j^3 + \bar{u}_j^3) \tag{72}$$

Therefore, we can instead solve the following optimization in $\mathbb{R}$:

$$\max_{\{m_j, -2\leq x_j\leq 2, x_0=2\}} \sum_j m_j x_j^3 - \epsilon \sum_j m_j^2 \quad \text{s.t.} \sum_j m_j x_j = 0 \tag{73}$$

whose solutions give a sufficient condition. Using Lagrangian multiplier, we have:

$$\frac{\partial L}{\partial x_j} = m_j(3x_j^2 - \lambda) = 0, \qquad \frac{\partial L}{\partial m_j} = x_j^3 - 2\epsilon m_j - \lambda x_j = 0 \tag{74}$$

which leads to $\lambda = 3$, $m_j = 1/\epsilon$ and $x_1 = x_2 = -1$. Therefore, $u_1 = \omega_3$ and $u_2 = \omega_3^2$ for 3-th root of unity $\omega_3 = e^{2\pi/3}$ (or vise versa).

Constructing $\boldsymbol{z}' \in R_\circledast$. It is clear that $r_{pmk_0k_0}(\tilde{\boldsymbol{z}}_{k_0}) \neq 0$ for $m = \pm 2k_0$ so $\tilde{\boldsymbol{z}}_{k_0} \notin R_\circledast$. We construct $\boldsymbol{z}'$ of order-2 so that $r_{pmk_0k_0}(\boldsymbol{z}'_{k_0}) = 0$:

$$z'_{pk1} = \mathbb{I}(k = k_0)\xi_p + \mathbb{I}(k = -k_0)\bar{\xi}_p, \qquad z'_{pk2} = \mathbb{I}(k = k_0)\bar{\xi}_p + \mathbb{I}(k = -k_0)\xi_p \tag{75}$$

with the constraint that $\Re(\xi_p^2\xi_c) = 0$ (i.e., pure imaginary) for $p \in \{a, b\}$ so that $r_{pmk_0k_0}(\boldsymbol{z}') = \xi_p^2\xi_c + \overline{\xi_p^2\xi_c} = 0$, but $\Re(\xi_a\xi_b\xi_c) > 0$ so that $r_{k_0k_0k_0} = \xi_a\xi_b\xi_c + \overline{\xi_a\xi_b\xi_c} > 0$. This is possible, e.g., by setting $\xi_b = \bar{\xi}_a = e^{\pm\pi\mathrm{i}/4}$ (i.e., ω_8 or $\bar{\omega}_8$), $\xi_c = 1$. $\square$

Corollary 4 (Perfect Memorization). *Construct the following two d-order weights $\boldsymbol{z}_a$ and $\boldsymbol{z}_b$. Specifically, for $0 \le j < d$ and $k \ne 0$:*

$$z_{akj}^{(a)} = \omega_d^{kj}/\sqrt{d}, \qquad z_{bkj}^{(a)} = 1/\sqrt{d}, \qquad z_{ckj}^{(a)} = \omega_d^{-kj}/\sqrt{2d} \tag{9}$$

$$z_{bkj}^{(b)} = 1/\sqrt{d}, \qquad z_{akj}^{(b)} = \omega_d^{kj}/\sqrt{d}, \qquad z_{ckj}^{(b)} = \omega_d^{-kj}/\sqrt{2d} \tag{10}$$

where $\omega_d := e^{-2\pi \mathrm{i}/d}$ is the d-th root of unity. Here $\boldsymbol{z}_a \in R_{\mathrm{c}}(k_1 \ne k) \cap R_{\mathrm{n}} \cap R_{\circledast}(p = b \text{ or } m \ne k)$, $\boldsymbol{z}_b \in R_{\mathrm{c}}(k_2 \ne k) \cap R_{\mathrm{n}} \cap R_{\circledast}(p = a \text{ or } m \ne k)$. Then $\boldsymbol{z}_M = \boldsymbol{z}_a * \boldsymbol{z}_b$ satisfies the global optimality condition (Eqn. 4) and is the perfect memorization solution with $\mathrm{ord}(\boldsymbol{z}_M) = d^2$:

$$z_{akj_1j_2}^{(M)} = \omega^{kj_1}/d, \qquad z_{bkj_1j_2}^{(M)} = \omega^{kj_2}/d, \qquad z_{ckj_1j_2}^{(M)} = \omega^{-k(j_1+j_2)}/2d \tag{11}$$

where each hidden node is indexed by $j = (j_1, j_2)$, $0 \le j_1, j_2 < d$, $k \ne 0$.

Proof. Simply plugging in the solution and check whether the equations specified the equations. For $\boldsymbol{z}_a$, for $k = 0$ everything is zero; for $k \ne 0$, we have:

$$r_{k_1k_2k}(\boldsymbol{z}_a) = \sum_j a_{k_1j} b_{k_2j} c_{kj} = \frac{1}{d\sqrt{2d}} \sum_j \omega^{j(k_1-k)} = \frac{1}{\sqrt{2d}} \mathbb{I}(k_1 = k \ne 0) \tag{76}$$

$$r_{amk'k}(\boldsymbol{z}_a) = \sum_j a_{k'j} a_{m-k',j} c_{kj} = \frac{1}{d\sqrt{2d}} \sum_j \omega^{j(m-k)} = \frac{1}{\sqrt{2d}} \mathbb{I}(m = k \ne 0) \tag{77}$$

$$r_{bmk'k}(\boldsymbol{z}_a) = \sum_j b_{k'j} b_{m-k',j} c_{kj} = \frac{1}{d\sqrt{2d}} \sum_j \omega^{-jk} = \frac{1}{\sqrt{2d}} \mathbb{I}(k = 0) = 0 \tag{78}$$

$$\tag{79}$$

Therefore, $\boldsymbol{z}_a \in R_{\mathrm{c}}(k_1 \ne k) \cap R_{\mathrm{n}} \cap R_{\circledast}(p = b \text{ or } m \ne k)$. Similar for $\boldsymbol{z}_b$. For $\boldsymbol{z}_M := \boldsymbol{z}_a * \boldsymbol{z}_b$, it satisfies all constraints (i.e., for any r, either $\boldsymbol{z}_a$ satisfies with $r(\boldsymbol{z}_a) = 0$, or $\boldsymbol{z}_b$ satisfies with $r(\boldsymbol{z}_b) = 0$) and we have:

$$r_{kkk}(\boldsymbol{z}_a * \boldsymbol{z}_b) = r_{kkk}(\boldsymbol{z}_a) r_{kkk}(\boldsymbol{z}_b) = 1/2d \tag{80}$$

So $\boldsymbol{z}_M$ satisfies the sufficient conditions (Eqn. 4). □

G Gradient Dynamics (Proof)

Theorem 4 (Dynamics of MPs). *The dynamics of MPs satisfies $\dot{\boldsymbol{r}} = -JJ^* \overline{\nabla_{\boldsymbol{r}} \ell}$, which has positive inner product with the negative gradient direction $-\overline{\nabla_{\boldsymbol{r}} \ell}$.*

Proof. By gradient descent of $\mathcal{W}$, we have $\dot{\mathcal{W}} = -\overline{\nabla_{\mathcal{W}} \ell}$. By chain rule, we have:

$$\dot{\mathcal{W}} = -\overline{\nabla_{\mathcal{W}} \ell} = -\overline{J^\top \nabla_{\boldsymbol{r}} \ell} = -J^* \overline{\nabla_{\boldsymbol{r}} \ell} \tag{81}$$

Then the dynamics of $\boldsymbol{r} = \boldsymbol{r}(\boldsymbol{z}(\mathcal{W}))$, as driven by the dynamics of $\mathcal{W}$, is given by

$$\dot{\boldsymbol{r}} = J\dot{\mathcal{W}} = -JJ^* \overline{\nabla_{\boldsymbol{r}} \ell} \tag{82}$$

To show positive inner product, we have:

$$-\overline{\nabla_{\boldsymbol{r}} \ell}^* \dot{\boldsymbol{r}} = \overline{\nabla_{\boldsymbol{r}} \ell}^* JJ^* \overline{\nabla_{\boldsymbol{r}} \ell} = \|J^* \overline{\nabla_{\boldsymbol{r}} \ell}\|_2^2 \ge 0 \tag{83}$$

□

Theorem 5 (The Occam's Razer: Preference of low-order solutions). *If $\boldsymbol{z} = \boldsymbol{y} * \boldsymbol{z}'$ and both $\boldsymbol{z}$ (of order q) and $\boldsymbol{z}'$ are global optimal solutions, then there exists a path of zero loss connecting $\boldsymbol{z}$ and $\boldsymbol{z}'$ in the space of $\mathcal{Z}_q$. As a result, lower-order solutions are preferred if trained with L_2 regularization.*

Proof. Let $\mathrm{ord}(\boldsymbol{z}) = q$ and $\mathrm{ord}(\boldsymbol{z}') = q'$. Then $q'|q$. Since both $\boldsymbol{z}$ and $\boldsymbol{z}'$ are global optimal. Since r_{kkk} is ring homomorphism, we know that $r_{kkk}(\boldsymbol{z}) = r_{kkk}(\boldsymbol{z}') r_{kkk}(\boldsymbol{y}) = 1/2d = r_{kkk}(\boldsymbol{z}')$ and thus $r_{kkk}(\boldsymbol{y}) = 1$ for all $k \ne 0$.

Let the augmented identity $\boldsymbol{e} \in \mathcal{Z}_q$ be $e_{pmj} = \mathbb{I}(j=0)$. Then $r_{kkk}(\boldsymbol{e}) = 1$ for all $k \neq 0$.

We want to construct a path in $\mathcal{Z}_q$, the space of order-q solutions as follows:

$$\tilde{\boldsymbol{z}}(t) = \tilde{\boldsymbol{y}}(t) * \boldsymbol{z}', \qquad 0 \le t \le 1 \tag{84}$$

in which $\tilde{\boldsymbol{y}}(0) = \boldsymbol{e}$, $\tilde{\boldsymbol{y}}(1) = \boldsymbol{y}$, and $r_{kkk}(\tilde{\boldsymbol{y}}(t)) = 1$ for any t. To see why this is possible, pick a continuous family of trajectories $\hat{\boldsymbol{y}}(t;\lambda)$ with $\lambda \in [0,1]$ so that they satisfies

$$\hat{\boldsymbol{y}}(0;\lambda) = \boldsymbol{e}, \qquad \hat{\boldsymbol{y}}(1;\lambda) = \boldsymbol{y}, \qquad r_{kkk}(\hat{\boldsymbol{y}}(t;0)) \le 1, \qquad r_{kkk}(\hat{\boldsymbol{y}}(t;1)) \le 1 \tag{85}$$

which can always be achieved by scaling some trajectory with a factor that depends on λ. Then by intermediate theorem, there exists $\lambda(t)$ so that $r_{kkk}(\hat{\boldsymbol{y}}(t;\lambda(t))) = 1$ for some k. Note that for different frequency k and k', r_{kkk} and $r_{k'k'k'}$ involves disjoint components of $\boldsymbol{z}$ so we could find such a path for all $k \neq 0$.

Therefore, for any monomial potential r included in MSE loss (Eqn. 3), we have

$$r(\tilde{\boldsymbol{z}}(t)) = r(\tilde{\boldsymbol{y}}(t)) r(\boldsymbol{z}') = \begin{cases} \text{finite} \cdot 0 = 0 & r \neq r_{kkk} \\ 1 \cdot 1/2d = 1/2d & r = r_{kkk} \end{cases} \tag{86}$$

and thus the entire trajectory $\tilde{z}(t) = \tilde{\boldsymbol{y}}(t) * \boldsymbol{z}' \in \mathcal{Z}_q$ connecting $\boldsymbol{z}$ and $\boldsymbol{e} * \boldsymbol{z}'$, which is $\boldsymbol{z}'$ in the space of $\mathcal{Z}_q$, is also globally optimal.

To see why weight decay regularization leads to lower-order solution, we could simply compare the ℓ_2 norm of $\boldsymbol{z} = \boldsymbol{y} * \boldsymbol{z}'$ and $\boldsymbol{e} * \boldsymbol{z}'$. At each frequency k, this reduces to the following optimization problem:

$$\min \sum_j |a_j|^2 + |b_j|^2 + |c_j|^2, \qquad \text{s.t.} \sum_j a_j b_j c_j = 1 \tag{87}$$

where $a_j := y_{akj}$, $b_j := y_{bkj}$ and $c_j := y_{ckj}$. Since we know that arithmetic mean is no less than geometric mean:

$$\frac{|a_j|^2 + |b_j|^2 + |c_j|^2}{3} \ge \sqrt[3]{|a_j b_j c_j|^2} \tag{88}$$

We have:

$$\sum_j |a_j|^2 + |b_j|^2 + |c_j|^2 \ge 3 \sum_j |a_j b_j c_j|^{2/3} \ge 3 \tag{89}$$

The last inequality holds because (1) if any $|a_j b_j c_j| \ge 1$, then it holds, (2) if all $|a_j b_j c_j| < 1$, then since a^x is a decreasing function for $a < 1$, $\sum_j |a_j b_j c_j|^{2/3} \ge \sum_j |a_j b_j c_j| \ge |\sum_j a_j b_j c_j| = 1$.

The minimizer is reached when $|a_j| = |b_j| = |c_j|$. Note that if $a_j b_j c_j$ has any complex phase or negative, then in order to satisfy $\sum_j a_j b_j c_j = 1$, objective function needs to be larger. So without loss of generality, we could study $a_j = b_j = c_j = x_j \ge 0$ and the optimization problem becomes

$$\min \sum_j x_j^2, \qquad \text{s.t.} \sum_j x_j^3 = 1, \quad x_j \ge 0 \tag{90}$$

which has a minimizer at the corners $(1, 0, \ldots)$. This corresponds to $a_j = b_j = c_j = \mathbb{I}(j=0)$, which is the augmented identity $\boldsymbol{e} \in \mathcal{Z}_q$. $\square$

Theorem 6 (Infinite Width Limits at Initialization)**.** *Considering the modified loss of Eqn. 3 with only the first two terms:* $\tilde{\ell}_k := r_{kkk} + d \sum_{k_1 k_2} |r_{k_1 k_2 k}|^2$, *if the weights are i.i.d Gaussian and network width* $q \to +\infty$, *then* JJ^* *converge to diagonal and the dynamics of MPs is decoupled.*

Proof. For each component of $H = JJ^*$, after computation, they can be written as the following:

$$h_{k_1 k_2 k_3, k_1' k_2' k_3'} = \sum_{pmj} \frac{\partial r_{k_1 k_2 k_3}}{\partial z_{pmj}} \overline{\frac{\partial r_{k_1' k_2' k_3'}}{\partial z_{pmj}}} \tag{91}$$

$$= \mathbb{I}(k_1 = k_1') \sum_j b_{k_2 j} \bar{b}_{k_2' j} c_{k_3 j} \bar{c}_{k_3' j} \tag{92}$$

$$+ \mathbb{I}(k_2 = k_2') \sum_j a_{k_1 j} \bar{a}_{k_1' j} c_{k_3 j} \bar{c}_{k_3' j} \tag{93}$$

$$+ \mathbb{I}(k_3 = k_3') \sum_j a_{k_1 j} \bar{a}_{k_1' j} b_{k_2 j} \bar{b}_{k_2' j} \tag{94}$$

where $a_{kj} := z_{akj}$, $b_{kj} := z_{bkj}$ and $c_{kj} := z_{ckj}$. Then for component $(k_1k_2k_3, k'_1, k'_2, k'_3)$, if any $k_p \neq k'_p$ for some $p \in \{a, b, c\}$, then the corresponding $z_{pk_pj}\bar{z}_{pk'_pj}$ has random phase for hidden node j, and $h_{k_1k_2k_3,k'_1k'_2k'_3} \to 0$ when $q \to +\infty$. $\square$