Unraveling and Overcoming Challenges in Machine Learning: Generalizability, Adaptability, and Multifacetedness

Hanseul Cho jhs4015@kaist.ac.kr
Kim Jaechul Graduate School of AI, KAIST

Contents

1. Introduction

Can deep learning (DL)—or machine learning (ML) in a broader term—be the ultimate methodology to solve every difficult problem?

It is undeniable that ML and DL have driven monumental success in the last decades, both academically and economically, in sequence modeling, natural language processing, image/video processing, material discovery, robotics, and many other fields. These triumphs originate from the capability of models trained to estimate the relationships between variables from data, enabling them to make plausible predictions for unseen data similar to the training set.

Nonetheless, ML is still far from perfect; the so-called "deep-learning magic" does not always happen in reality. One of the main challenges in ML is that the generalization capability (or **generalizability**) of ML models often falls short, especially when a significant deviation in data distribution occurs, although the generalization task may seem obvious for humans. Another critical challenge is the **adaptability** of ML models. After some training iterations, they often struggle to adjust their inferences in the face of constantly evolving environments. This difficulty restricts their real-world applications, where the informative pool of data may frequently change over time. In addition to the aspects of the model's (in)abilities, the **multifacetedness**—the characteristics of having multiple goals and purposes, some of which might be incompatible with others—of the problem setting adjoins more complexity to learning. Addressing all these challenges requires a careful, systematic, and mathematically rigorous analysis of their underlying mechanisms. This is because DL models are often hardly interpretable so they are often perceived as "black boxes."

In this research, we focus on three primary keywords—*generalizability, adaptability, and multifacetedness*—characterizing three vital challenges in ML. By doing so, we aim to rigorously investigate the root causes of these obstacles, develop an intuitive understanding of their mechanisms, and propose new methodologies to overcome the obstacles.

1.1 Organization

This research proposal consists of several sections, each aiming to motivate the research topics, summarize the intermediate achievements, and propose ongoing/future projects. In Section 2, we study the challenges in out-of-distribution (OOD) generalization of modern sequence-to-sequence models, mainly focusing on decoder-only Transformers. In Section 3, we study the challenges in maintaining the capability of an ML model to adapt to the circumstance shifts, being crucial in continual, incremental, and reinforcement learning. Lastly, in Section 4, we study the learning problems with multiple goals; it is divided into two parts: minimax optimization problems and a multi-constrained optimization setup.

2. Out-of-Distribution Generalization of Sequence-to-Sequence Models

2.1 Backgrounds and Related Works

It is a controversial subject whether an artificial intelligence (AI) model established on top of a modern large language model (LLM) can acquire reasoning capability from data (Kambhampati et al., 2024; Mirzadeh et al., 2024; Yang et al., 2024; Yax et al., 2024). Even if this statement could be true, it is obvious that the reasoning of LLMs is quite different from humans' reasoning. In particular, there are several tasks in which humans can succeed but LLMs usually fail, which indeed plays a crucial role in scrutinizing the properties of a pre-trained LLM's inference. In this section, we mostly focus on one such problem setting: length generalization. In addition, the architecture we are the most interested in is (decoder-only) Transformers (Vaswani et al., 2017), building blocks of cutting-edge LLMs (Dubey et al., 2024; Gemini et al., 2023; OpenAI, 2023).

Length generalization refers to a sort of OOD generalization capability of a sequence-to-sequence model to extrapolate its performance to longer sequences than those in training data. Unfortunately, it has recently been illuminated that Transformers often lack the ability of length generalization (Anil et al., 2022; Deletang et al., 2023; Wu et al., 2023; Zhang et al., 2023), although the underlying sequence generation rules seem to apply to any lengths. Understanding and mitigating the failures in length generalization is of great importance because of the following two aspects:

1. *Limited Generalizability*: It corroborates the fundamental limitation of LLMs that they do not genuinely understand the underlying task structure but may rely on short-cut learning which is only applicable to sequences of trained lengths;

2. *Efficiency*: Improving length generalization can automatically extend the applicability of the models in both memory- and computation-efficient ways.

Despite the huge revolution in ML due to LLMs, Transformers often struggle with length generalization even for simple arithmetic and algorithmic tasks. Thus, these tasks are widely regarded as reasonable but engaging test beds to study the capabilities of Transformers (Abbe et al., 2023; Jelassi et al., 2023; Kazemnejad et al., 2023; Kim et al., 2021; Lee et al., 2024b; McLeish et al., 2024; Nogueira et al., 2021; Shen et al., 2023; Zhou et al., 2024a,b). On the other hand, humans can length-generalize in arithmetic tasks by learning the essential task-solving rules. Hence, the failure in arithmetic tasks implies that the original Transformers and its typical variants are not capable of implementing the true task-solving algorithm.

It is worth mentioning that an increasing amount of recent intriguing works are trying to meticulously elaborate on the class of tasks that are length-generalizable by Transformers or other sequence models (Ahuja and Mansouri, 2024; Chen et al., 2024; Huang et al., 2024; Yang and Chiang, 2024).

2.2 Intermediate Results: Position Coupling

In Cho et al. (2024), we proposed *Position Coupling*, a simple method for improving length generalization of decoder-only Transformers by injecting the positional structure of a given arithmetic/algorithmic task into a learned absolute positional embedding (APE) module. With the proposed method, we achieved robust and significant length generalizations in several tasks such as long-integer addition, $N \times 2$ multiplication (expecting generalization in the multiplicand's length while the multiplier's length is fixed as 2), copying, reversing, and more. In particular, our models trained on up to 30-digit additions showcased near-perfect generalizations for up to 200-digit additions; our models also achieved 500-digit generalizations with up to 160-digit training. Basically, our proposed Position Coupling is a collection of position ID assignment rules established on top of a learned APE. Our method can work under two assumptions: we know a task where we want to achieve length generalization; we know the positional correspondence between tokens regardless of token lengths. Then, we assign the *same* position IDs to positionally relevant tokens, which we call a procedure of *coupling* the position IDs, unlike the usual method of assigning position IDs in an increasing order starting from 0. We theoretically explained and empirically verified that our method helps generate attention patterns beneficial to solving the given task, enabling the model to entirely solve integer addition and $N \times 2$ multiplication tasks with exponentially long operands in theory.

We further extended the scope of the problem settings for which Position Coupling is applicable and effective, by introducing appropriate scratchpad methods (Cho et al., 2025). We first observe that the scratchpad recording intermediate solving steps (Anil et al., 2022), together with Position Coupling, enables a remarkable length generalization in the Parity task. Motivated by this observation, we proposed a couple of scratchpad methods, each of which is tailor-made for the 'multi-operand integer addition task (expecting generalization in both the number and the lengths of summands)' and the 'general integer multiplication task (expecting generalization in the lengths of both multiplicand and multiplier)'. We also designed a couple of multi-level position ID coupling methods for these two tasks equipped with scratchpads. With a non-trivial combination of Position Coupling and scratchpads, we eventually obtained significant length generalization results for both tasks, which is the first and the only outcome in the literature of arithmetic length generalization as far as we know. It is empirically shown to be impossible if we solely apply a single-level position ID coupling without any scratchpad. We strongly believe that this is because the targeted tasks require a linearly increasing number of important tokens to perform every step of the next-token prediction as the sequence gets longer. Furthermore, we mathematically proved that a decoder-only Transformer equipped with Position Coupling can entirely solve the scratch-padded version of the multi-operand integer addition task, where we require the embedding dimension that only scales logarithmically with the sequence length.

2.3 Ongoing Researches & Future Directions

Tons of questions remain unsolved in the field of understanding and improving the OOD generalization capability of state-of-the-art sequence-to-sequence architectures, without being limited to the length-generalization of Transformers.

Deep Dive into Position Coupling. A more rigorous understanding of Position Coupling's mechanism would be an immediate next goal of research. To this end, we would like to characterize the class of tasks that are length-generalizable thanks to Position Coupling under appropriate assumptions on sequence-to-sequence model architectures. We conjecture that not only there is a strict subclass of algorithmic tasks having a length-equivariant algorithm of coupling position IDs that enables length generalization (i.e., length-generalizable tasks), but also there exists a transformation of non-length-generalizable tasks into a length-generalizable task. This bold conjecture is built upon our previous works: two-operand addition is length-generalizable by Position Coupling, but not two-operand multiplication; nonetheless, a scratchpad transforms the latter one into length-generalizable. We also strongly believe that a class of tasks length-generalizable with Position Coupling is a strict super-class of tasks that are length-generalizable with simple absolute position IDs. This is because the two-operand integer addition task turns out to be non-length-generalizable in the sense of the latter class of tasks (Huang et al., 2024).

Exploiting Structure of Language Data for Length Extrapolation. Some researchers have reported that exploiting the hierarchical structures in natural or programming language datasets (e.g., sentences containing words, functions containing keywords or variables, etc.) enhances the Transformer's context length extrapolation (He et al., 2024; Zhang et al., 2024). In particular, Zhang et al. (2024) propose the hierarchical RoPE (HiRoPE), a simple two-dimensional extension of the rotary position embedding (RoPE) (Su et al., 2024), and reported the benefits in length extrapolation. We notice that HiRoPE is not the only way to implement the multi-level positional information. Our goal is to propose a better method of assigning the multi-level position IDs and a better extension of RoPE to properly reflect and exploit the structure of language data. We expect this will facilitate the further length extrapolation of Transformer-baed language models. We would like to mention that bi-level variants of RoPE have already been widely studied in the literature of Transformers for vision/tabular data Heo et al. (2025); Li et al. (2024); Ravi et al. (2024), although many of their problem settings are far from the context of extrapolation in sequence length.

Compositional Generalizaition of Non-recursive Sequence Models. *Compositional generalization* is another popular OOD generalization problem in the research of sequence modeling, natural language processing, and even computer vision. It refers to the problem of recognizing entirely new combinations of atomic concepts observed during training. However, non-recursive parallel architectures like Transformers usually fail in this problem setting, but not entirely. Then, in what condition the Transformer-based language model can combine its knowledge to make plausible reasoning?

3. Towards an Adaptable Learner under Circumstance Shifts

3.1 Backgrounds and Related Works

The fittest survives (Darwin, 1859; Spencer, 1864), and so does every intelligent learner. Data in the real world evolves, either abruptly or gradually, rather than staying still. To be constantly intelligent and useful, AI systems need to continually obtain, extend, and utilize knowledge from the evolving data. It is the key motivation for the research of continual/incremental/lifelong learning. Let us refer to the ability to adapt to new incoming information as *adaptability*, while many researchers also use the term *plasticity*, coined from the field of neuroscience (Fuchs and Flügge, 2014; Ramón y Cajal, 1907, 1913; Stahnisch and Nitsch, 2002), to indicate the same concept.

Unfortunately, the majority of researchers agree that ML models often have trouble adapting to the changing environment, thereby failing to acquire new knowledge from fresh data. This problem is often called *loss of plasticity* and has drawn the attention of several research communities on reinforcement learning (RL) as well as continual learning (CL) (Abbas et al., 2023; De Lange et al., 2021; Dohare et al., 2024; Hadsell et al., 2020; Klein et al., 2024; Lyle et al., 2023; Shi et al., 2024; Wang et al., 2024a,b).

Another major challenge especially in CL happens for models that exceedingly focus on adaptation to the fresh stream of data so that they fail to retain their performances on the past data which is no longer accessible. This problem is known as *(catastrophic) forgetting*, which is also a term borrowed from modern neuroscience (French, 1999; McClelland et al., 1995; McCloskey and Cohen, 1989; Scoville and Milner, 1957). Balancing between plasticity and memory stability (opposite of forgetting) is a longstanding dilemma in the research of CL (De Lange et al., 2021; Wang et al., 2024a).

To mitigate the loss of plasticity and catastrophic forgetting, it is of great importance but demanding to understand the underlying dynamics behind them. Plenty of possible reasons for losing plasticity have been proposed: for an extensive survey, refer to Klein et al. (2024). Also, there have recently been a few advances towards a mathematically rigorous understanding of it (Gallici et al., 2024). However, our understanding is at the very initial phases, still not clear nor thorough.

3.2 Intermediate Results

Both empirically and theoretically, we made some progress towards understanding for learning dynamics of continually or incrementally evolving agents.

In Lee et al. (2023a), we study the loss of plasticity phenomenon in sample-efficient deep RL. We first argue that there are two key aspects of DNN's plasticity: input plasticity (i.e., adaptability to input distribution shifts) and label plasticity (i.e., adaptability to changing conditional distribution of label for a given input). By a set of careful ablation studies with synthetic experiments, we reveal that these two factors can be well-separated because several existing methods for maintaining plasticity and improving generalization can be categorized into two. The methods for making the loss landscape smoother and more benign, such as sharpness-aware minimization (SAM) optimizer (Foret et al., 2021)[1] and layer normalization (LayerNorm) (Ba et al., 2016), help DNN maintain input plasticity but not label plasticity. On the other hand, the methods that facilitate the neuron activations, such as occasional and partial re-initialization of neural networks (D'Oro et al., 2023; Nikishin et al., 2022; Zhou et al., 2022) and concatenated ReLU activation (CReLU) (Abbas et al., 2023), help DNN maintain label plasticity rather than label plasticity. Based on these findings, we introduce a training recipe "PLASTIC" for sample-efficient RL, which harmoniously combines all these techniques to address both types of plasticity. As main empirical results, we showcase that PLASTIC and its computation-efficient variant (PLASTIC†, combining LayerNorm and last-layer re-initialization) achieves competitive performance on benchmarks including Atari-100K (Bellemare et al., 2013) and Deepmind Control Suite (Tassa et al., 2018).

Now, let us move our attention to a mathematical analysis of CL with a simple linear model $f(\boldsymbol{x}; \boldsymbol{w}) = \boldsymbol{x}^\top \boldsymbol{w}$ (Jung et al., 2025). In particular, we focus on the learning dynamics of the gradient descent (GD) algorithm sequentially run on a stream of binary classification tasks ($y \in \{\pm 1\}$). As observed in many real-world problems (Gultekin and Gultekin, 1983; Verwimp et al., 2023; Yang et al., 2022b), we assume that every task is chosen from a finite collection of tasks, either in a cyclic or random order. This is an interesting and novel problem setting because of the following two reasons: one reason is that most theoretical works largely focus on regression problems based on quadratic loss functions (Asanuma et al., 2021; Bennani et al., 2020; Doan et al., 2021; Evron et al., 2022; Goldfarb and Hand, 2023; Lee et al., 2021; Li et al., 2023), while we consider learning multiple binary classifications with logistic loss $\ell(z) = \log(1 + e^{-z})$. Another reason is that a notable work on continual linear classification by Evron et al. (2023) assumes a non-realistic projection-based algorithm to obtain convergence guarantees, whereas we analyze a gradient-based optimizer, which makes the exact training dynamics much more difficult to characterize entirely. Our theoretical contributions can be summarized into three parts as below, where we denote by J the number of cycles in cyclic task ordering cases:

1. *Cyclic Ordering & Jointly Separable Tasks*: We first consider the cyclically-revealed tasks that are jointly solvable with a single parameter vector $\boldsymbol{w} \in \mathbb{R}^d$. In this case, we showed the asymptotic convergence of the joint training loss, the parameter's directional convergence towards

1. To the best of our knowledge, our work has empirically verified the efficacy of SAM optimizer for the first time in RL literature.

the joint ℓ_2 max-margin direction (where the parameter norm diverges at a rate of $O(\log(J))$), and a non-asymptotic loss convergence of rate $O(\log^2(J)/J)$. We remark that the asymptotic loss convergence and directional convergence can be proved without convexity. On top of that, using the non-asymptotic loss convergence rate, we derive a $O(\log^4(J)/J^2)$ diminishing rate of catastrophic forgetting which occurs every cycle. With this forgetting-per-cycle analysis, we discovered that the data alignment between different tasks impacts forgetting: in particular, an upper bound of cycle-averaged forgetting decreases as the negative alignment between tasks gets smaller.

2. *Random Ordering & Jointly Separable Tasks*: Even when the tasks are randomly revealed with replacement at every stage, we can still prove similar asymptotic loss convergence and the implicit bias result but in an almost-sure sense.

3. *Cyclic Ordering & Jointly (Strictly) Non-Separable Tasks*: We also considered the case when a joint solution (perfectly classifying every data point) never exists and the joint training loss has a unique non-zero minimum over its unconstrained domain (i.e., where a certain amount of forgetting some data points is inevitable). In this case, we proved the non-asymptotic convergence rate of $O(\log^2(J)/J^2)$, in terms of both the squared parameter distance and the joint training loss.

3.3 Future Directions

Rigorous Understanding of Re-Learning. It is a common observation that *re-learning* (i.e., re-initializing and then resuming the training) is strikingly effective for enhancing the adaptability of a learner, especially when the model is a DNN. This has been examined in various learning setups including CL and RL and facilitated a lot of learning methods leveraging this idea (Ash and Adams, 2020; Dohare et al., 2024; D'Oro et al., 2023; Frati et al., 2024; Mhammedi et al., 2024; Nikishin et al., 2022; Shin et al., 2024; Sokar et al., 2023; Zhou et al., 2022). Not only that, the re-learning technique is shown to be effective for simple generalization of vision models because (arguably) it helps mitigate the problem of spurious correlation between the foreground and the background of the image (Alabdulmohsin et al., 2021; Kirichenko et al., 2023; Le et al., 2023; Taha et al., 2021; Zhao et al., 2018). Then, it is natural to ask: why is re-learning so powerful in various domains and problem settings? When is it beneficial? Several works aim to uncover the reason for the effectiveness of re-learning, but some of them still rely on empirical proxies rather than rigorous math or a careful causal analysis (Zaidi et al., 2023). Even though a work by Mhammedi et al. (2024) rigorously proves some learnability guarantees in an online Markov decision process (MDP) setting, the re-learning algorithm proposed in it seems a bit different from the practical resetting methods. Thus, there is still a huge gap in our theoretical understanding of the effectiveness of practical re-learning methods. It is worth mentioning that it might be interesting to consider not only the re-initialization of weight entries but also the resetting of the optimizer states (e.g., momentum, second moments in adaptive optimizers).

4. Multifaceted Learning: Learning with Multiple Conflicting Goals

This section contains two largely different sub-topics of multifaceted learning problems that are not directly relevant to each other but might become fortuitously connected to any other topics mentioned in this article. In Section 4.1, we study the convergence analyses of minimax optimization algorithms. Next, in Section 4.2, we study an algorithm for fair streaming principal component analysis (PCA) as an instance of learning problems with multiple constraints.

We would like to pinpoint these topics to be particularly intriguing because such a problem indeed appears in the real world, concerning the trade-off, tension, and/or balance among multiple goals.

4.1 Minimax Optimization: Learning Problems Beyond Minimization

4.1.1 Backgrounds

Minimax optimization is a problem setting with an objective function having variables for both minimization and maximization of it, described as $\min_{\boldsymbol{x}\in\mathcal{X}} \max_{\boldsymbol{y}\in\mathcal{Y}} f(\boldsymbol{x};\boldsymbol{y})$ (von Neumann, 1928). It can be used to formulate several ML/DL problems including but not limited to generative adversarial networks (GANs) (Goodfellow et al., 2020), adversarial training (Madry et al., 2018; Sinha et al., 2018), and area-under-the-ROC[2]-curve (AUROC) maximization (Ying et al., 2016; Yuan et al., 2021).

One of the main challenges in minimax optimization is *trainability*, or *convergence* itself. In usual minimization problems, optimization of the training loss is regarded as arguably an easy problem; the choice of optimizer determines the convergence *speed* towards at least a local minimum under mild assumptions. In stark contrast, there are two main issues in terms of optimization of minimax problems. One issue is the conceptually non-intuitive local optimality criteria. Although the minimax problems are a strict generalization of minimization problems,[3] the notion of minima does not trivially generalize to minimax problems, especially when the problem is nonconvex-nonconcave. Instead, several non-trivial notions of equilibria (and their tractability) have been proposed, such as (local) Nash equilibrium, (local) minimax point, correlated equilibrium, and Φ-equilibrium; refer to a recent paper by Cai et al. (2024) for a broad survey. Another issue is the difficulty in convergence of minimax algorithms towards a (local) equilibrium (Hsieh et al., 2021). Even for a convex-concave problem with deterministic (i.e., full-batch) gradient oracles, naive algorithms like gradient descent-ascent (GDA) often fail to converge to a Nash equilibrium (Bailey et al., 2020; Gidel et al., 2019; Zhang et al., 2022).

4.1.2 Intermediate Results & Possible Future Works

Below, we study convergent algorithms for minimax optimization assuming benign structures of the problem enabling the convergence to a (local) optimum.

In Cho and Yun (2023), we study the convergence acceleration of stochastic gradient descent-ascent (SGDA) thanks to without-replacement sampling (i.e., shuffling). We consider finite-sum minimax optimization, where the total objective function $f(\boldsymbol{x};\boldsymbol{y})$ is an average of L-Lipschitz-gradient[4] component functions $f_i(\boldsymbol{x};\boldsymbol{y})$'s, i.e., $f(\boldsymbol{x};\boldsymbol{y}) = \frac{1}{n}\sum_{i=1}^{n} f_i(\boldsymbol{x};\boldsymbol{y})$. We compare two SGDA algorithms whose only difference is the sampling method of component indices at every gradient step: `SGDA` (with-replacement sampling of components, previously studied by Lin et al. 2020; Yang et al. 2020, 2022a) and `SGDA-RR` (*random reshuffling*[5] of components). In terms of the sufficient number of gradient steps until ϵ-convergence to equilibrium, we show that the convergence speed of `SGDA-RR` is faster than that of `SGDA`. Here, our analysis is conducted under the benign assumption that the total objective function $f(\boldsymbol{x};\boldsymbol{y})$ is either *nonconvex-PŁ* (i.e., $-f(\boldsymbol{x};\cdot)$ is PŁ[6] for any choice of $\boldsymbol{x}$) or *primal-PŁ-PŁ* (i.e., additionally satisfies that $\max_{\boldsymbol{y}} f(\cdot;\boldsymbol{y})$ is PŁ).

Let us move our attention to deterministic (i.e., non-stochastic) minimax algorithms. In Lee et al. (2024a), we elucidate the strict superiority of alternating GDA (`Alt-GDA`) over its simultaneous-update counterpart (`Sim-GDA`) by rigorously characterizing the global convergence rates (toward the Nash equilibrium) of both algorithms for strongly-convex-strongly-concave (SCSC) and Lipschitz-gradient objective functions. On top of that, exceedingly leveraging the acceleration due to alternating updates between minimization and maximization variables, we propose a better algorithm called

2. ROC is an abbreviation of 'receiver-operating characteristic'.
3. Consider the case where the domain of maximization variables is a singleton set.
4. A differentiable function is said to be Lipschitz-gradient if its gradient is Lipschitz continuous.
5. We refer to a without-replacement sampling method of indices that uniformly randomly shuffles the order of the indices at the beginning of every epoch as *random reshuffling* or *random-shuffling*.
6. A differentiable function $f(\cdot)$ is said to be a Polyak-Łojasiewicz (PŁ) function with a constant $\mu > 0$ when it satisfies $\|\nabla f(\boldsymbol{z})\|^2 \geq 2\mu(f(\boldsymbol{z}) - \inf_{\tilde{\boldsymbol{z}}} f(\tilde{\boldsymbol{z}}))$. A PŁ function is not necessarily convex.

alternating-extrapolation GDA (`Alex-GDA`). Although it is a general framework that can subsume `Sim-GDA` and `Alt-GDA` as special cases, our theory proves that (1) certain configurations of `Alex-GDA` result in a faster convergence rate than `Alt-GDA` for SCSC and Lipschitz-gradient objectives, and (2) some other configurations of `Alex-GDA` result in a successful convergence for bilinear minimax problems where both `Sim-GDA` and `Alt-GDA` fails to converge.

Future Directions. Some questions raised from the results above remain open.

- A recent work by Cha et al. (2023) proposed an advanced convergence rate lower bound for general permutation-based stochastic gradient descent (SGD) for finite-sum minimization problems. Their bound implies that a permutation sampling method called `GraB` (Lu et al., 2022) has a near-optimal convergence upper bound that matches the lower bound. Would this bound-matching result successfully extend to finite-sum minimax problems? Even when it is provably impossible, it can explain a strict separation between minimization and minimax optimization.
- Although we proved that a collection of configurations of `Alex-GDA` exhibits a faster convergence rate than `Alt-GDA`, it is still unclear exactly which configuration is optimal among them. Can we extend our previous analysis to exactly characterize the optimal instance of `Alex-GDA`?

4.2 Principal Component Analysis with Fairness and Memory Constraints

Principal component analysis (PCA) is a popular dimensionality reduction technique using projection onto a low-dimensional linear subspace. An exact PCA of a pre-defined d-dimensional dataset needs a computation of sample covariance matrix, thereby resulting in a $\mathcal{O}(d^2)$ memory requirement. However, if the target dimension k is much smaller than the full dimension d, we have some other approximate alternatives based on iterative algorithms (e.g., noisy power method (NPM), Hardt and Price, 2014) that only require $\mathcal{O}(dk)$ memory consumption, which can handle relatively constrained memory budget.

In Lee et al. (2023b), we additionally consider (representational) fairness as a constraint: we aim to find an orthogonal projection maximizing the variance while making the projected data points indistinguishable in terms of their sensitive attributes. To tackle this problem, we propose an algorithm called the fair noisy power method (FNPM), a two-phase modification of NPM. To provide a statistical guarantee of the algorithm in both fairness and optimality, we rigorously characterize a sample complexity upper bound that is sufficient to achieve both near-perfect fairness and nearly maximized projection variance. Moreover, we numerically verified that FNPM is the most memory-efficient one among all existing fair PCA algorithms by running them to process full-resolution full-colored CelebA dataset (Liu et al., 2015) on a computer with a moderate-sized memory.

Future Direction. It is still unclear whether our proposed FNPM is optimal in sample complexity. To investigate the optimality, we should prove a sample complexity lower bound of fair streaming PCA, under appropriate assumptions on the true data distribution and the sampling methods.

References

Zaheer Abbas, Rosie Zhao, Joseph Modayil, Adam White, and Marlos C Machado. Loss of plasticity in continual deep reinforcement learning. In *Conference on Lifelong Learning Agents*, pages 620–636. PMLR, 2023. 4, 5

Emmanuel Abbe, Samy Bengio, Aryo Lotfi, and Kevin Rizk. Generalization on the unseen, logic reasoning and degree curriculum. In *Proceedings of the International Conference on Machine Learning (ICML)*, pages 31–60. PMLR, 2023. 2

Kartik Ahuja and Amin Mansouri. On provable length and compositional generalization. *arXiv preprint arXiv:2402.04875*, 2024. 3

Ibrahim Alabdulmohsin, Hartmut Maennel, and Daniel Keysers. The impact of reinitialization on generalization in convolutional neural networks. In *Proceedings of the Conference on Artificial Intelligence (AAAI)*, 2021. 6

Cem Anil, Yuhuai Wu, Anders Andreassen, Aitor Lewkowycz, Vedant Misra, Vinay Ramasesh, Ambrose Slone, Guy Gur-Ari, Ethan Dyer, and Behnam Neyshabur. Exploring length generalization in large language models. *Advances in Neural Information Processing Systems (NeurIPS)*, 35: 38546–38556, 2022. 2, 3

Haruka Asanuma, Shiro Takagi, Yoshihiro Nagano, Yuki Yoshida, Yasuhiko Igarashi, and Masato Okada. Statistical mechanical analysis of catastrophic forgetting in continual learning with teacher and student networks. *Journal of the Physical Society of Japan*, 90(10):104001, 2021. 5

Jordan Ash and Ryan P Adams. On warm-starting neural network training. *Advances in neural information processing systems*, 33:3884–3894, 2020. 6

Jimmy Lei Ba, Jamie Ryan Kiros, and Geoffrey E Hinton. Layer normalization. *arXiv preprint arXiv:1607.06450*, 2016. 5

James P Bailey, Gauthier Gidel, and Georgios Piliouras. Finite regret and cycles with fixed step-size via alternating gradient descent-ascent. In *Proceedings of Conference on Learning Theory (COLT)*, pages 391–407. PMLR, 2020. 7

Marc G Bellemare, Yavar Naddaf, Joel Veness, and Michael Bowling. The arcade learning environment: An evaluation platform for general agents. *Journal of Artificial Intelligence Research*, 47: 253–279, 2013. 5

Mehdi Abbana Bennani, Thang Doan, and Masashi Sugiyama. Generalisation guarantees for continual learning with orthogonal gradient descent. *arXiv preprint arXiv:2006.11942*, 2020. 5

Yang Cai, Constantinos Costis Daskalakis, Haipeng Luo, Chen-Yu Wei, and Weiqiang Zheng. On tractable ϕ-equilibria in non-concave games. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 37, 2024. URL https://openreview.net/forum?id=3CtTMF5zzM. 7

Jaeyoung Cha, Jaewook Lee, and Chulhee Yun. Tighter lower bounds for shuffling SGD: Random permutations and beyond. In Andreas Krause, Emma Brunskill, Kyunghyun Cho, Barbara Engelhardt, Sivan Sabato, and Jonathan Scarlett, editors, *Proceedings of the International Conference on Machine Learning (ICML)*, volume 202 of *Proceedings of Machine Learning Research*, pages 3855–3912. PMLR, 23–29 Jul 2023. URL https://proceedings.mlr.press/v202/cha23a.html. 8

Yang Chen, Yitao Liang, and Zhouchen Lin. Low-dimension-to-high-dimension generalization and its implications for length generalization. *arXiv preprint arXiv:2410.08898*, 2024. 3

Hanseul Cho and Chulhee Yun. SGDA with shuffling: faster convergence for nonconvex-PŁ minimax optimization. In *Proceedings of the International Conference on Learning Representations (ICLR)*, volume 11. OpenReview.net, 2023. URL https://openreview.net/forum?id=6xXtM8bFFJ. 7

Hanseul Cho, Jaeyoung Cha, Pranjal Awasthi, Srinadh Bhojanapalli, Anupam Gupta, and Chulhee Yun. Position coupling: Improving length generalization of arithmetic transformers using task structure. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 37, 2024. 3

Hanseul Cho, Jaeyoung Cha, Srinadh Bhojanapalli, and Chulhee Yun. Arithmetic transformers can length-generalize in both operand length and count. In *Proceedings of the International Conference on Learning Representations (ICLR)*, volume 13. OpenReview.net, 2025. URL https://openreview.net/forum?id=eIgGesYKLG. 3

Charles Darwin. *On the Origin of Species by Means of Natural Selection, or the Preservation of Favoured Races in the Struggle for Life*. John Murray, London, 1 edition, 1859. 4

Matthias De Lange, Rahaf Aljundi, Marc Masana, Sarah Parisot, Xu Jia, Aleš Leonardis, Gregory Slabaugh, and Tinne Tuytelaars. A continual learning survey: Defying forgetting in classification tasks. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPMAI)*, 44(7):3366–3385, 2021. 4, 5

Gregoire Deletang, Anian Ruoss, Jordi Grau-Moya, Tim Genewein, Li Kevin Wenliang, Elliot Catt, Chris Cundy, Marcus Hutter, Shane Legg, Joel Veness, and Pedro A Ortega. Neural networks and the chomsky hierarchy. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2023. URL https://openreview.net/forum?id=WbxHAzkeQcn. 2

Thang Doan, Mehdi Abbana Bennani, Bogdan Mazoure, Guillaume Rabusseau, and Pierre Alquier. A theoretical analysis of catastrophic forgetting through the ntk overlap matrix. In *International Conference on Artificial Intelligence and Statistics*, pages 1072–1080. PMLR, 2021. 5

Shibhansh Dohare, J. Fernando Hernandez-Garcia, Qingfeng Lan, Parash Rahman, A Rupam Mahmood, and Richard S. Sutton. Loss of plasticity in deep continual learning. *Nature*, 632(8026): 768–774, 2024. 4, 6

Pierluca D'Oro, Max Schwarzer, Evgenii Nikishin, Pierre-Luc Bacon, Marc G Bellemare, and Aaron Courville. Sample-efficient reinforcement learning by breaking the replay ratio barrier. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2023. URL https://openreview.net/forum?id=OpC-9aBBVJe. 5, 6

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, et al. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024. 2

Itay Evron, Edward Moroshko, Rachel Ward, Nathan Srebro, and Daniel Soudry. How catastrophic can catastrophic forgetting be in linear regression? In *Conference on Learning Theory*, pages 4028–4079. PMLR, 2022. 5

Itay Evron, Edward Moroshko, Gon Buzaglo, Maroun Khriesh, Badea Marjieh, Nathan Srebro, and Daniel Soudry. Continual learning in linear classification on separable data. In *International Conference on Machine Learning*, pages 9440–9484. PMLR, 2023. 5

Pierre Foret, Ariel Kleiner, Hossein Mobahi, and Behnam Neyshabur. Sharpness-aware minimization for efficiently improving generalization. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2021. URL https://openreview.net/forum?id=6Tm1mposlrM. 5

Lapo Frati, Neil Traft, Jeff Clune, and Nick Cheney. Reset it and forget it: Relearning last-layer weights improves continual and transfer learning. In *ECAI 2024*, pages 2998–3005. IOS Press, 2024. 6

Robert M. French. Catastrophic forgetting in connectionist networks. *Trends in Cognitive Sciences*, 3(4):128–135, 1999. 5

Eberhard Fuchs and Gabriele Flügge. Adult neuroplasticity: More than 40 years of research. *Neural Plasticity*, 2014(1):541870, 2014. doi: https://doi.org/10.1155/2014/541870. URL https://onlinelibrary.wiley.com/doi/abs/10.1155/2014/541870. 4

Matteo Gallici, Mattie Fellows, Benjamin Ellis, Bartomeu Pou, Ivan Masmitja, Jakob Nicolaus Foerster, and Mario Martin. Simplifying deep temporal difference learning. *arXiv preprint arXiv:2407.04811*, 2024. 5

Gemini, Rohan Anil, Sebastian Borgeaud, Yonghui Wu, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M Dai, Anja Hauth, et al. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*, 2023. 2

Gauthier Gidel, Hugo Berard, Gaëtan Vignoud, Pascal Vincent, and Simon Lacoste-Julien. A variational inequality perspective on generative adversarial networks. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2019. URL https://openreview.net/forum?id=r1laEnA5Ym. 7

Daniel Goldfarb and Paul Hand. Analysis of catastrophic forgetting for random orthogonal transformation tasks in the overparameterized regime. In *International Conference on Artificial Intelligence and Statistics*, pages 2975–2993. PMLR, 2023. 5

Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial networks. *Communications of the ACM*, 63(11):139–144, 2020. 7

Mustafa N Gultekin and N Bulent Gultekin. Stock market seasonality: International evidence. *Journal of financial economics*, 12(4):469–481, 1983. 5

Raia Hadsell, Dushyant Rao, Andrei A. Rusu, and Razvan Pascanu. Embracing change: Continual learning in deep neural networks. *Trends in Cognitive Sciences*, 24(12):1028–1040, 2020. doi: 10.1016/j.tics.2020.09.004. 4

Moritz Hardt and Eric Price. The Noisy Power Method: A Meta Algorithm with Applications. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 27, pages 2861–2869. Curran Associates, Inc., 2014. 8

Zhenyu He, Guhao Feng, Shengjie Luo, Kai Yang, Liwei Wang, Jingjing Xu, Zhi Zhang, Hongxia Yang, and Di He. Two stones hit one bird: Bilevel positional encoding for better length extrapolation. In *Proceedings of the International Conference on Machine Learning (ICML)*, 2024. URL https://openreview.net/forum?id=luqH1eL4PN. 4

Byeongho Heo, Song Park, Dongyoon Han, and Sangdoo Yun. Rotary position embedding for vision transformer. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 289–305. Springer, 2025. 4

Ya-Ping Hsieh, Panayotis Mertikopoulos, and Volkan Cevher. The limits of min-max optimization algorithms: Convergence to spurious non-critical sets. In *Proceedings of the International Conference on Machine Learning (ICML)*, pages 4337–4348. PMLR, 2021. 7

Xinting Huang, Andy Yang, Satwik Bhattamishra, Yash Sarrof, Andreas Krebs, Hattie Zhou, Preetum Nakkiran, and Michael Hahn. A formal framework for understanding length generalization in transformers. *arXiv preprint arXiv:2410.02140*, 2024. 3, 4

Samy Jelassi, Stéphane d'Ascoli, Carles Domingo-Enrich, Yuhuai Wu, Yuanzhi Li, and François Charton. Length generalization in arithmetic transformers. *arXiv preprint arXiv:2306.15400*, 2023. 2

Hyunji Jung, Hanseul Cho, and Chulhee Yun. Convergence and implicit bias of gradient descent on continual linear classification. In *Proceedings of the International Conference on Learning Representations (ICLR)*, volume 13. OpenReview.net, 2025. URL https://openreview.net/forum?id=DTqx3iqjkz. 5

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. *arXiv preprint arXiv:2402.01817*, 2024. 2

Amirhossein Kazemnejad, Inkit Padhi, Karthikeyan Natesan Ramamurthy, Payel Das, and Siva Reddy. The impact of positional encoding on length generalization in transformers. *Advances in Neural Information Processing Systems (NeurIPS)*, 36, 2023. 2

Jeonghwan Kim, Giwon Hong, Kyung-min Kim, Junmo Kang, and Sung-Hyon Myaeng. Have you seen that number? investigating extrapolation in question answering models. In *Proceedings of the conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7031–7037, 2021. 2

Polina Kirichenko, Pavel Izmailov, and Andrew Gordon Wilson. Last layer re-training is sufficient for robustness to spurious correlations. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2023. URL https://openreview.net/forum?id=Zb6c8A-Fghk. 6

Timo Klein, Lukas Miklautz, Kevin Sidak, Claudia Plant, and Sebastian Tschiatschek. Plasticity loss in deep reinforcement learning: A survey. *arXiv preprint arXiv:2411.04832*, 2024. 4, 5

Phuong Quynh Le, Jörg Schlötterer, and Christin Seifert. Is last layer re-training truly sufficient for robustness to spurious correlations? *IJCAI Workshop on XAI*, 2023. 6

Hojoon Lee, Hanseul Cho, Hyunseung Kim, Daehoon Gwak, Joonkee Kim, Jaegul Choo, Se-Young Yun, and Chulhee Yun. PLASTIC: Improving input and label plasticity for sample efficient reinforcement learning. In A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine, editors, *Advances in Neural Information Processing Systems (NeurIPS)*, volume 36, pages 62270–62295. Curran Associates, Inc., 2023a. URL https://proceedings.neurips.cc/paper_files/paper/2023/file/c464fc4516aca4e68f2a14e67c6f0402-Paper-Conference.pdf. 5

Jaewook Lee, Hanseul Cho, and Chulhee Yun. Fundamental benefit of alternating uptades in minimax optimization. In Ruslan Salakhutdinov, Zico Kolter, Katherine Heller, Adrian Weller, Nuria Oliver, Jonathan Scarlett, and Felix Berkenkamp, editors, *Proceedings of the International Conference on Machine Learning (ICML)*, volume 235 of *Proceedings of Machine Learning Research*, pages 26439–26514. PMLR, 21–27 Jul 2024a. URL https://proceedings.mlr.press/v235/lee24e.html. 7

Junghyun Lee, Hanseul Cho, Se-Young Yun, and Chulhee Yun. Fair streaming principal component analysis: Statistical and algorithmic viewpoint. In A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine, editors, *Advances in Neural Information Processing Systems (NeurIPS)*, volume 36, pages 5126–5167. Curran Associates, Inc., 2023b. URL https://proceedings.neurips.cc/paper_files/paper/2023/file/1074541383db5ef12d6ac66d2f8e8d34-Paper-Conference.pdf. 8

Nayoung Lee, Kartik Sreenivasan, Jason D. Lee, Kangwook Lee, and Dimitris Papailiopoulos. Teaching arithmetic to small transformers. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2024b. URL https://openreview.net/forum?id=dsUB4bst9S. 2

Sebastian Lee, Sebastian Goldt, and Andrew Saxe. Continual learning in the teacher-student setup: Impact of task similarity. In *International Conference on Machine Learning*, pages 6109–6119. PMLR, 2021. 5

Haoran Li, Jingfeng Wu, and Vladimir Braverman. Fixed design analysis of regularization-based continual learning. In *Conference on Lifelong Learning Agents*, pages 513–533. PMLR, 2023. 5

Jia-Nan Li, Jian Guan, Wei Wu, Zhengtao Yu, and Rui Yan. 2d-tpe: Two-dimensional positional encoding enhances table understanding for large language models. *arXiv preprint arXiv:2409.19700*, 2024. 4

Tianyi Lin, Chi Jin, and Michael Jordan. On gradient descent ascent for nonconvex-concave minimax problems. In Hal Daumé III and Aarti Singh, editors, *Proceedings of the International Conference on Machine Learning (ICML)*, volume 119 of *Proceedings of Machine Learning Research*, pages 6083–6093. PMLR, 13–18 Jul 2020. URL https://proceedings.mlr.press/v119/lin20a.html. 7

Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. Deep Learning Face Attributes in the Wild. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pages 3730–3738, 2015. 8

Yucheng Lu, Wentao Guo, and Christopher M De Sa. Grab: Finding provably better data permutations than random reshuffling. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh, editors, *Advances in Neural Information Processing Systems (NeurIPS)*, volume 35, pages 8969–8981. Curran Associates, Inc., 2022. URL https://proceedings.neurips.cc/paper_files/paper/2022/file/3acb49252187efa352a1ae0e4b066ced-Paper-Conference.pdf. 8

Clare Lyle, Zeyu Zheng, Evgenii Nikishin, Bernardo Avila Pires, Razvan Pascanu, and Will Dabney. Understanding plasticity in neural networks. In *Proceedings of the International Conference on Machine Learning (ICML)*, pages 23190–23211. PMLR, 2023. 4

Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. Towards deep learning models resistant to adversarial attacks. In *Proceedings of the International Conference on Learning Representations (ICLR)*. OpenReview.net, 2018. 7

James L McClelland, Bruce L McNaughton, and Randall C O'Reilly. Why there are complementary learning systems in the hippocampus and neocortex: insights from the successes and failures of connectionist models of learning and memory. *Psychological review*, 102(3):419, 1995. 5

Michael McCloskey and Neal J Cohen. Catastrophic interference in connectionist networks: The sequential learning problem. In *Psychology of learning and motivation*, volume 24, pages 109–165. Elsevier, 1989. 5

Sean McLeish, Arpit Bansal, Alex Stein, Neel Jain, John Kirchenbauer, Brian R. Bartoldson, Bhavya Kailkhura, Abhinav Bhatele, Jonas Geiping, Avi Schwarzschild, and Tom Goldstein. Transformers can do arithmetic with the right embeddings. In *Advances in Neural Information Processing Systems*, volume 37, 2024. 2

Zakaria Mhammedi, Dylan J Foster, and Alexander Rakhlin. The power of resets in online reinforcement learning. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 37, 2024. URL https://openreview.net/forum?id=7sACcaOmGi. 6

Iman Mirzadeh, Keivan Alizadeh, Hooman Shahrokhi, Oncel Tuzel, Samy Bengio, and Mehrdad Farajtabar. GSM-symbolic: Understanding the limitations of mathematical reasoning in large language models. *arXiv preprint arXiv:2410.05229*, 2024. 2

Evgenii Nikishin, Max Schwarzer, Pierluca D'Oro, Pierre-Luc Bacon, and Aaron Courville. The primacy bias in deep reinforcement learning. In *Proceedings of the International Conference on Machine Learning (ICML)*, pages 16828–16847. PMLR, 2022. 5, 6

Rodrigo Nogueira, Zhiying Jiang, and Jimmy Lin. Investigating the limitations of transformers with simple arithmetic tasks. *arXiv preprint arXiv:2102.13019*, 2021. 2

OpenAI. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023. 2

Santiago Ramón y Cajal. Regeneración de los nervios.[translation and edited by j. bresler (1908) studien über nervenregeneration, johann ambrosius barth]. 1907. 4

Santiago Ramón y Cajal. *Estudios sobre la degeneración y regeneración del sistema nerviosa*, volume 1. Imprenta de Hijos de Nicolás Moya, 1913. 4

Nikhila Ravi, Valentin Gabeur, Yuan-Ting Hu, Ronghang Hu, Chaitanya Ryali, Tengyu Ma, Haitham Khedr, Roman Rädle, Chloe Rolland, Laura Gustafson, et al. Sam 2: Segment anything in images and videos. *arXiv preprint arXiv:2408.00714*, 2024. 4

William Beecher Scoville and Brenda Milner. Loss of recent memory after bilateral hippocampal lesions. *Journal of neurology, neurosurgery, and psychiatry*, 20(1):11, 1957. 5

Ruoqi Shen, Sébastien Bubeck, Ronen Eldan, Yin Tat Lee, Yuanzhi Li, and Yi Zhang. Positional description matters for transformers arithmetic. *arXiv preprint arXiv:2311.14737*, 2023. 2

Haizhou Shi, Zihao Xu, Hengyi Wang, Weiyi Qin, Wenyuan Wang, Yibin Wang, Zifeng Wang, Sayna Ebrahimi, and Hao Wang. Continual learning of large language models: A comprehensive survey. *arXiv preprint arXiv:2404.16789*, 2024. 4

Baekrok Shin, Junsoo Oh, Hanseul Cho, and Chulhee Yun. DASH: Warm-starting neural network training in stationary settings without loss of plasticity. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 37, 2024. 6

Aman Sinha, Hongseok Namkoong, and John Duchi. Certifying some distributional robustness with principled adversarial training. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2018. URL https://openreview.net/forum?id=Hk6kPgZA-. 7

Ghada Sokar, Rishabh Agarwal, Pablo Samuel Castro, and Utku Evci. The dormant neuron phenomenon in deep reinforcement learning. In *International Conference on Machine Learning*, pages 32145–32168. PMLR, 2023. 6

Herbert Spencer. *The Principles of Biology*, volume 1. Williams and Norgate, London, 1864. 4

Frank W. Stahnisch and Robert Nitsch. Santiago ramón y cajal's concept of neuronal plasticity: the ambiguity lives on. *Trends in Neurosciences*, 25(11):589–591, 2002. 4

Jianlin Su, Murtadha Ahmed, Yu Lu, Shengfeng Pan, Wen Bo, and Yunfeng Liu. Roformer: Enhanced transformer with rotary position embedding. *Neurocomputing*, 568:127063, 2024. 4

Ahmed Taha, Abhinav Shrivastava, and Larry S Davis. Knowledge evolution in neural networks. In *Proceedings of the IEEE/CVF conference on Computer Vision and Pattern Recognition (CVPR)*, pages 12843–12852, 2021. 6

Yuval Tassa, Yotam Doron, Alistair Muldal, Tom Erez, Yazhe Li, Diego de Las Casas, David Budden, Abbas Abdolmaleki, Josh Merel, Andrew Lefrancq, et al. Deepmind control suite. *arXiv preprint arXiv:1801.00690*, 2018. 5

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in Neural Information Processing Systems (NeurIPS)*, 30, 2017. 2

Eli Verwimp, Kuo Yang, Sarah Parisot, Lanqing Hong, Steven McDonagh, Eduardo Pérez-Pellitero, Matthias De Lange, and Tinne Tuytelaars. Clad: A realistic continual learning benchmark for autonomous driving. *Neural Networks*, 161:659–669, 2023. 5

John von Neumann. Zur theorie der gesellschaftsspiele. *Mathematische Annalen*, 100(1):295–320, 1928. doi: 10.1007/BF01448847. URL https://doi.org/10.1007/BF01448847. 7

Liyuan Wang, Xingxing Zhang, Hang Su, and Jun Zhu. A comprehensive survey of continual learning: theory, method and application. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPMAI)*, 2024a. 4, 5

Zhenyi Wang, Enneng Yang, Li Shen, and Heng Huang. A comprehensive survey of forgetting in deep learning beyond continual learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPMAI)*, 2024b. 4

Zhaofeng Wu, Linlu Qiu, Alexis Ross, Ekin Akyürek, Boyuan Chen, Bailin Wang, Najoung Kim, Jacob Andreas, and Yoon Kim. Reasoning or reciting? exploring the capabilities and limitations of language models through counterfactual tasks. *arXiv preprint arXiv:2307.02477*, 2023. 2

Andy Yang and David Chiang. Counting like transformers: Compiling temporal counting logic into softmax transformers. In *Proceedings of Conference on Language Modeling (COLM)*, volume 1, 2024. URL https://openreview.net/forum?id=FmhPg4UJ9K. 3

Junchi Yang, Negar Kiyavash, and Niao He. Global convergence and variance reduction for a class of nonconvex-nonconcave minimax problems. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems (NeurIPS)*, volume 33, pages 1153–1165. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper_files/paper/2020/file/0cc6928e741d75e7a92396317522069e-Paper.pdf. 7

Junchi Yang, Antonio Orvieto, Aurelien Lucchi, and Niao He. Faster single-loop algorithms for minimax optimization without strong concavity. In Gustau Camps-Valls, Francisco J. R. Ruiz, and Isabel Valera, editors, *Proceedings of International Conference on Artificial Intelligence and Statistics (AISTATS)*, volume 151 of *Proceedings of Machine Learning Research*, pages 5485–5517. PMLR, 28–30 Mar 2022a. URL https://proceedings.mlr.press/v151/yang22b.html. 7

Yingxiang Yang, Zhihan Xiong, Tianyi Liu, Taiqing Wang, and Chong Wang. Fourier learning with cyclical data. In *International Conference on Machine Learning*, pages 25280–25301. PMLR, 2022b. 5

Yuan Yang, Siheng Xiong, Ali Payani, Ehsan Shareghi, and Faramarz Fekri. Can LLMs reason in the wild with programs? In Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen, editors, *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 9806–9829, Miami, Florida, USA, 11 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.findings-emnlp.573. URL https://aclanthology.org/2024.findings-emnlp.573/. 2

Nicolas Yax, Hernán Anlló, and Stefano Palminteri. Studying and improving reasoning in humans and machines. *Communications Psychology*, 2(1):51, 2024. doi: 10.1038/s44271-024-00091-8. URL https://doi.org/10.1038/s44271-024-00091-8. 2

Yiming Ying, Longyin Wen, and Siwei Lyu. Stochastic online AUC maximization. *Advances in Neural Information Processing Systems (NeurIPS)*, 29, 2016. 7

Zhuoning Yuan, Yan Yan, Milan Sonka, and Tianbao Yang. Large-scale robust deep AUC maximization: A new surrogate loss and empirical studies on medical image classification. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pages 3040–3049, 2021. 7

Sheheryar Zaidi, Tudor Berariu, Hyunjik Kim, Jorg Bornschein, Claudia Clopath, Yee Whye Teh, and Razvan Pascanu. When does re-initialization work? In Javier Antorán, Arno Blaas, Fan Feng, Sahra Ghalebikesabi, Ian Mason, Melanie F. Pradier, David Rohde, Francisco J. R. Ruiz, and Aaron Schein, editors, *Proceedings on "I Can't Believe It's Not Better! - Understanding Deep Learning Through Empirical Falsification" at NeurIPS 2022 Workshops*, volume 187 of *Proceedings of Machine Learning Research*, pages 12–26. PMLR, 2023. 6

Guodong Zhang, Yuanhao Wang, Laurent Lessard, and Roger B. Grosse. Near-optimal local convergence of alternating gradient descent-ascent for minimax optimization. In *Proceedings of International Conference on Artificial Intelligence and Statistics (AISTATS)*, 2022. 7

Kechi Zhang, Ge Li, Huangzhao Zhang, and Zhi Jin. HiRoPE: Length extrapolation for code models using hierarchical position. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL)*, volume 1, pages 13615–13627, 2024. 4

Yi Zhang, Arturs Backurs, Sebastien Bubeck, Ronen Eldan, Suriya Gunasekar, and Tal Wagner. Unveiling transformers with LEGO: A synthetic reasoning task, 2023. URL https://openreview.net/forum?id=1jDN-RfQfrb. 2

Kaikai Zhao, Tetsu Matsukawa, and Einoshin Suzuki. Retraining: A simple way to improve the ensemble accuracy of deep neural networks for image classification. In *International conference on pattern recognition (ICPR)*, pages 860–867. IEEE, 2018. 6

Hattie Zhou, Ankit Vani, Hugo Larochelle, and Aaron Courville. Fortuitous forgetting in connectionist networks. *Proceedings of the International Conference on Learning Representations (ICLR)*, 2022. URL https://openreview.net/forum?id=ei3SY1_zYsE. 5, 6

Hattie Zhou, Arwen Bradley, Etai Littwin, Noam Razin, Omid Saremi, Joshua M. Susskind, Samy Bengio, and Preetum Nakkiran. What algorithms can transformers learn? a study in length generalization. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2024a. URL https://openreview.net/forum?id=AssIuHnmHX. 2

Yongchao Zhou, Uri Alon, Xinyun Chen, Xuezhi Wang, Rishabh Agarwal, and Denny Zhou. Transformers can achieve length generalization but not robustly. *ICLR Workshop on Understanding of Foundation Models (ME-FoMo)*, 2024b. 2