On the Convergence of Continual Learning with Adaptive Methods

Seungyub Han[1] **Yeongmo Kim**[1] **Taehyun Cho**[1] **Jungwoo Lee**[1]

[1]Seoul National University, Seoul, Republic of Korea

Abstract

One of the objectives of continual learning is to prevent catastrophic forgetting in learning multiple tasks sequentially, and the existing solutions have been driven by the conceptualization of the plasticity-stability dilemma. However, the convergence of continual learning for each sequential task is less studied so far. In this paper, we provide a convergence analysis of memory-based continual learning with stochastic gradient descent and empirical evidence that training current tasks causes the cumulative degradation of previous tasks. We propose an adaptive method for nonconvex continual learning (NCCL), which adjusts step sizes of both previous and current tasks with the gradients. The proposed method can achieve the same convergence rate as the SGD method when the catastrophic forgetting term which we define in the paper is suppressed at each iteration. Further, we demonstrate that the proposed algorithm improves the performance of continual learning over existing methods for several image classification tasks.

1 INTRODUCTION

Learning new tasks without forgetting previously learned tasks is a key aspect of artificial intelligence to be as versatile as humans. Unlike the conventional deep learning that observes tasks from an i.i.d. distribution, continual learning train sequentially a model on a non-stationary stream of data [Ring, 1995, Thrun, 1994]. The continual learning AI systems struggle with catastrophic forgetting when the data access of previously learned tasks is restricted [French and Chater, 2002]. Although novel continual learning methods successfully learn the non-stationary stream sequentially, studies on the theoretical convergence analysis of both previous tasks and a current task have not yet been addressed. In this line of research, nonconvex stochastic optimization problems have been well studied on a single task to train deep neural networks and prove theoretical guarantees of good convergence.

Previous continual learning algorithms have introduced novel methods such as a replay memory to store and replay the previously learned examples [Lopez-Paz and Ranzato, 2017, Aljundi et al., 2019b, Chaudhry et al., 2019a], regularization methods that penalize neural networks [Kirkpatrick et al., 2017, Zenke et al., 2017], Bayesian methods that utilize the uncertainty of parameters or data points [Nguyen et al., 2018, Ebrahimi et al., 2020], and other recent approaches [Yoon et al., 2018, Lee et al., 2019]. The study of continual learning in Bayesian frameworks formulate a trained model for previous tasks parameter into an approximate posterior to learn a probabilistic model which have empirically good performance on entire tasks. However, Bayesian approaches can fail in practice and it can be hard to analyze the rigorous convergence due to the approximation. The memory-based methods are more straightforward approaches, where the learner stores a small subset of the data for previous tasks into a memory and utilizes the memory by replaying samples to keep a model staying in a feasible region without losing the performance on the previous tasks. Gradient episodic memory (GEM) [Lopez-Paz and Ranzato, 2017] first formulated the replay based continual learning as a constrained optimization problem. This formulation allows us to rephrase the constraints on objectives for previous tasks as inequalities based on the inner product of loss gradient vectors for previous tasks and a current task. However, the gradient update by GEM variants cannot guarantee both theoretical and empirical convergence of its constrained optimization problem. The modified gradient updates do not always satisfy the loss constraint theoretically, and we can also observe the forgetting phenomenon occurs empirically. It also implies that this intuitive reformulation violates the constrained optimization problem and cannot provide theoretical guarantee to prevent catastrophic forgetting without a rigorous convergence analysis.

Accepted for the 39th Conference on Uncertainty in Artificial Intelligence (UAI 2023).

In this work, we explain the cause of catastrophic forgetting by describing continual learning with a smooth nonconvex finite-sum optimization problem. In the standard single task case, SGD [Ghadimi and Lan, 2013], ADAM [Reddi et al., 2018], YOGI [Zaheer et al., 2018], SVRG [Reddi et al., 2016], and SCSG [Lei et al., 2017] are the algorithms for solving nonconvex problems that arise in deep learning. To analyze the convergence of those algorithms, previous works study the following nonconvex finite-sum problem

$$\min_{x \in \mathbb{R}^d} f(x) = \frac{1}{n} \sum_{i=1}^{n} f_i(x), \tag{1}$$

where we assume that **each objective** $f_i(x)$ **with a model** x **and a data point index** $i \in [n]$ for a dataset with size n (by the convention for notations in nonconvex optimization literature [Reddi et al., 2016]) is nonconvex with L-smoothness assumption. In general, we denote $f_i(x)$ as $f(x; d_i)$ where d_i is a datapoint tuple (INPUT, OUTPUT) with index i. We expect that a stochastic gradient descent based algorithm reaches a stationary point instead of the global minimum in nonconvex optimization. Unlike the convex case, the convergence is generally measured by the expectation of the squared norm of a gradient $\mathbb{E}\|\nabla f(x)\|^2$. The theoretical computational complexity is derived from the ϵ-accurate solution, which is also known as a stationary point with $\mathbb{E}\|\nabla f(x)\|^2 \leq \epsilon$. The general nonconvex finite-sum problems assume that all data points can be sampled during training iterations. This fact is an obstacle to directly apply (1) for continual learning problem.

We provide a solution of the above issue by leveraging memory-based methods, which allow models to access a partial access to the dataset of previous tasks. In this setting, we can analyze nonconvex stochastic optimization problems on the convergence of previous tasks with limited access. Similar with adaptive methods for noncovex optimization, we apply adaptive step sizes during optimization to minimize forgetting with theoretical guarantee. Specifically, we make the following contributions:

- We decompose the finite-sum problem of entire tasks into two summation terms for previous tasks and a current task, respectively. We theoretically show that small random subsets of previous tasks lead to analyzing the expected convergence rate of both tasks while learning a current task.
- We study the convergence of gradient methods under a small memory where the backward transfer performance degrades, and propose a new formulation of continual learning problem with the forgetting term. We then show why catastrophic forgetting occurs theoretically and empirically.
- Though memory-based methods mitigate forgetting, previous works does not fully exploit the gradient information of memory. We introduce a novel adaptive method and its extension which adjust step sizes between tasks at each step with theoretical ground, and demonstrate that both methods show remarkable performance on image classification tasks.

2 RELATED WORK

Memory-based methods. Early memory-based methods utilize memory by the distillation [Rebuffi et al., 2017, Li and Hoiem, 2017] or the optimization constraint [Lopez-Paz and Ranzato, 2017, Chaudhry et al., 2019a]. Especially, A-GEM [Chaudhry et al., 2019a] simplifies the approach for constraint violated update steps as the projected gradient on a reference gradient which ensures that the average memory loss over previous tasks does not increase. Recent works [Chaudhry et al., 2019b, 2020a, Riemer et al., 2018] have shown that updating the gradients on memory directly, which is called experience replay, is a light and prominent approach. We focus on convergence of continual learning, but the above methods focus on increasing the empirical performance without theoretical guarantee. Our analysis provides a legitimate theoretical convergence analysis under the standard smooth nonconvex finite-sum optimization problem setting. Further, [Knoblauch et al., 2020] shows the perfect memory for optimal continual learning is NP-hard by using set-theory, but the quantitative analysis of performance degradation is less studied.

Adaptive step sizes in nonconvex setting. Adaptive step sizes under smooth nonconvex finite-sum optimization problem have been studied on general single task cases [Reddi et al., 2018, Zhang et al., 2020, Zaheer et al., 2018] recently. [Simsekli et al., 2019, Zhang et al., 2020, Simsekli et al., 2020] have revealed that there exists a heavy-tailed noise in some optimization problems for neural networks, such as attention models, and [Zhang et al., 2020] shows that adaptive methods are helpful to achieve the faster convergence under the heavy-tailed distribution where stochastic gradients are poorly concentrated around the mean. In this work, we treat the continual learning problem where stochastic gradients of previous tasks are considered as the out-of-distribution samples in regard to a current task, and develop adaptive methods which are well-performed in continual learning.

3 PRELIMINARIES

Suppose that we observe the learning procedure on a data stream of continual learning at some arbitrary observation point. Let us consider time step $t = 0$ as given observation point. We define the previous task $\mathcal{P}$ for $t < 0$ as all visited data points and the current task $\mathcal{C}$ for $t \geq 0$ as all data points which will face in the future. Then, P and C can be defined as the sets of data points in $\mathcal{P}$ and $\mathcal{C}$ at time step $t = 0$, respectively. Note that the above task description is based on not a sequence of multiple tasks, but two separate sets to

analyze the convergence of each of P and C when starting to update the given batch at the current task $\mathcal{C}$ at some arbitrary observation point. We consider a continual learning problem as a smooth nonconvex finite-sum optimization problem with two decomposed objectives

$$\min_{x\in\mathbb{R}^d} h(x) = \frac{1}{n_f+n_g}\sum_{i\in P\cup C} h_i(x), \tag{2}$$

where n_f and n_g are the numbers of elements for P and C, and $h(x)$ can be decomposed into as follows:

$$h(x) = \frac{n_f}{n_f+n_g}f(x) + \frac{n_g}{n_f+n_g}g(x).$$

For clarity, we use $f(x) = h(x)|_P$ and $g(x) = h(x)|_C$ for the restriction of h to each dataset P and C, respectively. $f_i(x)$ and $g_j(x)$ also denotes the objective terms induced from data where each index is $i \in P$ and $j \in C$, respectively.

Suppose that the replay memories M_t for time step $\in [0, T]$ are random variables which are the subsets of $P \cup C$ to cover prior memory-based approaches [Chaudhry et al., 2019b,a]. To formulate an algorithm for memory-based approaches, we define mini-batches I_t which are sampled from a memory M_t at step t. We now define the stochastic update of memory-based method

$$x^{t+1} = x^t - \alpha_{H_t}\nabla f_{I_t}(x^t) - \beta_{H_t}\nabla g_{J_t}(x^t), \tag{3}$$

where $I_t \subset M_t$ and $J_t \subset C$ denote the mini-batches from the replay memory and the current data stream, respectively. Here, H_t is the union of I_t and J_t. In addition, for a given set S, $\nabla f_S(x^t), \nabla g_S(x^t)$ denote the loss gradient of a model x^t with the mini-batch S at time step t. The adaptive step sizes (learning rates) of $\nabla f_{I_t}(x^t)$ and $\nabla g_{J_t}(x^t)$ are denoted by α_{H_t} and β_{H_t} which are the functions of H_t.

It should be noted the mini-batch I_t from M_t might contain a datapoint $j \in C$ for some cases, such as ER-Reservoir.

Throughout the paper, we assume L-smoothness and the following statements.

Assumption 3.1. f_i is L-smooth that there exists a constant $L > 0$ such that for any $x, y \in \mathbb{R}^d$,

$$\|\nabla f_i(x) - \nabla f_i(y)\| \le L\|x-y\| \tag{4}$$

where $\|\cdot\|$ denotes the Euclidean norm. Then the following inequality directly holds that

$$\begin{aligned} -\frac{L}{2}\|x-y\|^2 &\le f_i(x) - f_i(y) - \langle \nabla f_i(y), x-y\rangle \\ &\le \frac{L}{2}\|x-y\|^2. \end{aligned} \tag{5}$$

We derive Equation 5 in Appendix C. Assumption 3.1 is a well-known and useful statement in nonconvex finite-sum optimization problem [Reddi et al., 2016, 2018, Zhang et al., 2020, Zaheer et al., 2018], and also helps us to describe the convergence of continual learning. We also assume the supremum of loss gap between an initial point x^0 and a global optimum x^* as Δ_f, and the upper bound on the variance of the stochastic gradients as σ_f in the following.

$$\begin{aligned} \Delta_f &= \sup_{x^0} f(x^0) - f(x^*), \\ \sigma_f^2 &= \sup_x \frac{1}{n_f}\sum_{i=1}^{n_f}\|\nabla f_i(x) - \nabla f(x)\|^2. \end{aligned}$$

It should be noted that $g_j(x), \nabla g_j(x)$, which denote the loss and the gradient for a current task, also satisfy all three above assumptions and the following statement.

To measure the efficiency of a stochastic gradient algorithm, we define the Incremental First-order Oracle (IFO) framework [Ghadimi and Lan, 2013]. IFO call is defined as a unit of computational cost by taking an index i which gets the pair $(\nabla f_i(x), f_i(x))$, and IFO complexity of an algorithm is defined as the summation of IFO calls during optimization. For example, a vanilla stochastic gradient descent (SGD) algorithm requires computational cost as much as the batch size b_t at each step, and the IFO complexity is the sum of batch sizes $\sum_{t=1}^{T} b_t$. Let $T(\epsilon)$ be the minimum number of iterations to guarantee ϵ-accurate solutions. The average bound of IFO complexity is less than or equal to $\sum_{t=1}^{T(\epsilon)} b_t = O(1/\epsilon^2)$ [Reddi et al., 2016].

4 CONTINUAL LEARNING AS NONCONVEX OPTIMIZATION

We first present a theoretical convergence analysis of memory-based continual learning in nonconvex setting. We aim to understand why catastrophic forgetting occurs in terms of the convergence rate, and reformulate the optimization problem of continual learning into a nonconvex setting with theoretical guarantee. For completeness we present all proofs in Appendix C.

4.1 MEMORY-BASED NONCONVEX CONTINUAL LEARNING

Unlike conventional smooth nonconvex finite-sum optimization problems where each mini-batch is i.i.d-sampled from the whole dataset $P \cup C$, the replay memory based continual learning encounters a non-i.i.d stream of data C with access to a small sized memory M_t. Algorithm 1 provides the pseudocode for memory-based approach with the iterative update rule 3. Now, we can analyze the convergence on P and C during a learning procedure on an arbitrary data

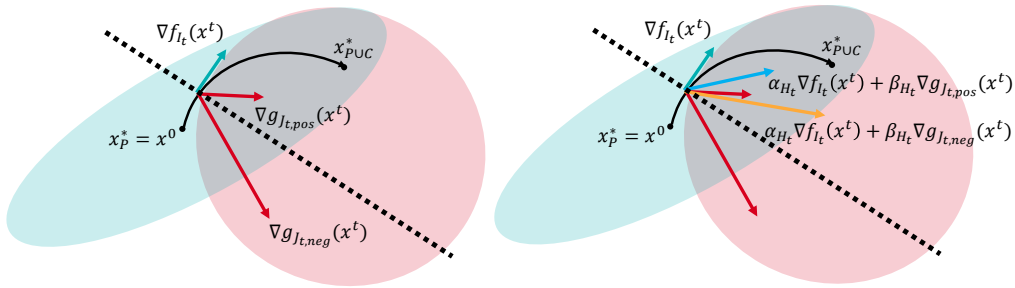


Figure 1: Geometric illustration of Non-Convex Continual Learning (NCCL). In continual learning, a model parameter x^t starts from a local optimal point for the previously learned tasks x^*_P. Over T iterations, we expect to reach a new optimal point $x^*_{P\cup C}$ which has a good performance on both P and C. In the t-th iteration, x^t encounters either $\nabla g_{J_t,pos}(x^t)$ or $\nabla g_{J_t,neg}(x^t)$. These two cases indicate whether $\langle f_{I_t}(x^t), \nabla g_{J_t}(x^t)\rangle$ is positive or not. To prevent x^t from escaping the feasible region, i.e., catastrophic forgetting, we impose a theoretical condition on learning rates for f and g.

Algorithm 1 Nonconvex Continual Learning (NCCL)

Require: Previous task set P, current task set C, initial model x^0.
Sample a initial memory $M_0 \subset P$ ▷ By replay schemes, the selection dist. of M_0 are different.
for $t = 0$ **to** $T-1$ **do**
  Sample a mini-batch $I_t \subset M_t$
  Sample a mini-batch $J_t \subset C$
  Compute step sizes $\alpha_{H_t}, \beta_{H_t}$ by $\nabla f_{I_t}(x^t), \nabla g_{J_t}(x^t)$
  $x^{t+1} \leftarrow x^t - \alpha_{H_t}\nabla f_{I_t}(x^t) - \beta_{H_t}\nabla g_{J_t}(x^t)$
  Update M_{t+1} by the rule of replay scheme with J_t.
end for

stream from two consecutive sets P and C for continual learning [Chaudhry et al., 2019a,b, 2020b].

By limited access to P, the expectation of gradient update $\mathbb{E}_{I_t\subset M_t}[\nabla f_{I_t}(x^t)]$ in Equation 3 for $f(x)$ is a biased estimate of the gradient $\nabla f(x^t)$. At the timestep t, we have

$$\nabla f_{M_t}(x^t) = \mathbb{E}_{I_t}\left[\nabla f_{I_t}(x^t)|M_t\right] = \mathbb{E}_{I_t}\left[\nabla f(x^t) + e_t|M_t\right] = \nabla f(x^t) + e_{M_t},$$

where e_t and e_{M_t} denote the error terms, $\nabla f_{I_t}(x^t) - \nabla f(x^t)$ and the expectation over I_t given M_t, respectively. It should be noted that a given replay memory M_t with small size at timestep t introduces an inevitable overfitting bias.

For example, there exist two popular memory schemes, episodic memory and ER-reservoir. The episodic memory $M_t = M_0$ for all t is uniformly sampled once from a random sequence of P, and ER-reservoir iteratively samples the replay memory M_t by the selection rule $M_t \subset M_{t-1} \cup J_t$. Here, we denote the history of M_t as $M_{[0:t]} = (M_0, \cdots, M_t)$. To compute the expectation over all stochasticities of NCCL, we need to derive the expectation of $\nabla f_{M_t}(x^t)$ over the randomness of M_t. We formalize the expectation over all learning trials with the selection randomness as follows.

Lemma 4.1. *If M_0 is uniformly sampled from P, then both episodic memory and ER-reservoir satisfies*

$$\mathbb{E}_{M_{[0:t]}}\left[\nabla f_{M_t}(x^t)\right] = \nabla f(x^t) \quad \textit{and} \quad \mathbb{E}_{M_{[0:t]}}\left[e_{M_t}\right] = 0.$$

Note that taking expectation iteratively with respect to the history $M_{[0:t]}$ is needed to compute the expected value of gradients for M_t. Surprisingly, taking the expectation of overfitting error over memory selection gets zero. However, it does not imply $e_t = 0$ for each learning trial with some $M_{[0:t]}$.

4.2 THEORETICAL CONVERGENCE ANALYSIS

We now propose two terms of interest in a gradient update of nonconvex continual learning (NCCL). We define the overfitting term B_t and the catastrophic forgetting term Γ_t as follows:

$$B_t = (L\alpha^2_{H_t} - \alpha_{H_t})\langle \nabla f(x^t), e_t\rangle + \beta_{H_t}\langle \nabla g_{J_t}(x^t), e_t\rangle,$$
$$\Gamma_t = \frac{\beta^2_{H_t}L}{2}\|\nabla g_{J_t}(x^t)\|^2 - \beta_{H_t}(1-\alpha_{H_t}L)\langle \nabla f_{I_t}(x^t), \nabla g_{J_t}(x^t)\rangle.$$

The amount of effect on convergence by a single update can be measured by using Equation 5 as follows:

$$f(x^{t+1}) \le f(x^t) - \langle \nabla f(x^t), \alpha_{H_t}\nabla f_{I_t}(x^t) + \beta_{H_t}\nabla g_{J_t}(x^t)\rangle + \frac{L}{2}\|\alpha_{H_t}\nabla f_{I_t}(x^t) + \beta_{H_t}\nabla g_{J_t}(x^t)\|^2 \quad (6)$$

by letting $x \leftarrow x^{t+1}$ and $y \leftarrow x^t$. Note that the above inequality can be rewritten as

$$f(x^{t+1}) \le f(x^t) - \left(\alpha_{H_t} - \frac{L}{2}\alpha^2_{H_t}\right)\|\nabla f(x^t)\|^2 + \Gamma_t + B_t + \frac{L}{2}\alpha^2_{H_t}\|e_t\|^2.$$

A NCCL algorithm update its model with two additional terms B_t, Γ_t compared to conventional SGD. An overfitting

term B_t and a catastrophic forgetting term Γ_t are obtained by grouping terms that contain e_t and $\nabla g_{J_t}(x^t)$, respectively. These two terms inevitably degrade the performance of NCCL with respect to time. It should be noted that Γ_t has $\langle \nabla f_{I_t}(x^t), \nabla g_{J_t}(x^t) \rangle$, which is a key factor to determine interference and transfer [Riemer et al., 2018]. On the other hand, B_t includes e_t, which is an error gradient between the batch from M_t and the entire dataset P.

Since taking the expectation over all stochasticities of NCCL implies the total expectation, we define the operator of total expectation with respect to $0 \leq t < T$ for ease of exposition as follows:

$$\mathbb{E}_t = \mathbb{E}_{M_{[0:t]}} \left[\mathbb{E}_{I_t} \left[\mathbb{E}_{J_t} \left[\cdot \,|I_t \right] \right] |M_{[0:t]} \right].$$

In addition, we denote $\mathbb{E}_{T-1} = \mathbb{E}$. We first state the stepwise change of upper bound.

Lemma 4.2. *Suppose that Assumption 3.1 holds and $0 < \alpha_{H_t} \leq \frac{2}{L}$. For x^t updated by Algorithm 1, we have*

$$\mathbb{E}_t \|\nabla f(x^t)\|^2 \leq \mathbb{E}_t \left[\frac{f(x^t) - f(x^{t+1}) + B_t + \Gamma_t}{\alpha_{H_t}(1 - \frac{L}{2}\alpha_{H_t})} \right] + \mathbb{E}_t \left[\frac{\alpha_{H_t} L}{2(1 - \frac{L}{2}\alpha_{H_t})} \sigma_f^2 \right]. \quad (7)$$

Surprisingly, we observe $\mathbb{E}_t[B_t] = 0$ by Lemma 4.1. It should be also noted that the individual trial with a randomly given M_0 cannot cancel the effect of B_t. We discuss more details of overfitting to memory in Appendix E.

We now describe a convergence analysis of Algorithm 1. We telescope over training iterations for the current task, which leads to obtain the following theorem.

Theorem 4.3. *Let $\alpha_{H_t} = \alpha = \frac{c}{\sqrt{T}}$ for some $0 < c \leq \frac{2\sqrt{T}}{L}$ and $t \in \{0, \cdots, T-1\}$. By Lemma 4.2, the iterates of NCCL satisfy*

$$\min_t \mathbb{E}\|\nabla f(x^t)\|^2 \leq \frac{A}{\sqrt{T}} \left(\frac{1}{c} \left(\Delta_f + \sum_{t=0}^{T-1} \mathbb{E}\left[\Gamma_t\right] \right) + \frac{Lc}{2}\sigma_f^2 \right)$$

where $A = 1/(1 - L\alpha/2)$.

We also prove the convergence rate of a current task C with the gradient udpates from the replay-memory M in continual learining.

Lemma 4.4. *Suppose that $I_t \cap J_t = \emptyset$, Taking expectation over $I_t \subset M_t$ and $J_t \subset C$, we have*

$$\min_t \mathbb{E}\|\nabla h|_{M\cup C}(x^t)\|^2 \leq \sqrt{\frac{2\Delta_{h|_{M\cup C}} L}{T}} \sigma_{h|_{M\cup C}}, \quad (8)$$

where $\Delta_{h|_{M\cup C}}$ and $\sigma_{h|_{M\cup C}}$ is the version of loss gap and the variance for h on $M \cup C$, respectively.

Thus, the convergence of a current task $\mathcal{C}$ is guaranteed, since its superset $M \cup C$ is converged. Otherwise, the convergence rate might differ from the conventional SGD for $\mathcal{C}$ by the given $\Delta_{h|_{M\cup C}}, \sigma_{h|_{M\cup C}}$ at time 0, but the asymptotic convergence rate is still identical.

One key observation is that $\mathbb{E}[\Gamma_t]$ are cumulatively added on the upper bound of $\mathbb{E}\|\nabla f(x)\|^2$, which is a constant in conventional SGD. The loss gap Δ_f and the variance of gradients σ_f are fixed values. In practice, tightening $\sum_t \mathbb{E}[\Gamma_t]$ appears to be critical for the performance of NCCL. However, $\sum_{t=0}^{T-1} \mathbb{E}[\Gamma_t]/\sqrt{T}$ is not guaranteed to converge to 0. This fact gives rise to catastrophic forgetting in terms of a nondecreasing upper bound. We now show the key condition of the convergence of $\sum_{t=0}^{T-1} \mathbb{E}[\Gamma_t]/\sqrt{T}$.

Lemma 4.5. *Let an upper bound $\beta > \beta_{H_t} > 0$. Consider two cases, $\beta < \alpha$ and $\beta \geq \alpha$ for α in Theorem 4.3. We have the following bound*

$$\sum_{t=0}^{T-1} \frac{\mathbb{E}[\Gamma_t]}{\sqrt{T}} < O\left(1/T^{3/2} + 1/T\right) \quad \textit{when } \beta < \alpha,$$
$$\sum_{t=0}^{T-1} \frac{\mathbb{E}[\Gamma_t]}{\sqrt{T}} < O\left(\sqrt{T} + 1/\sqrt{T}\right), \quad \textit{when } \beta \geq \alpha.$$

With the following theorem, we show that $f(x)$ can converge even if we have limited access to P.

Theorem 4.6. *Let $\beta_{H_t} < \alpha = \frac{c}{\sqrt{T}}$ for all t. Then we have the convergence rate*

$$\min_t \mathbb{E}\|\nabla f(x^t)\|^2 \leq O\left(\frac{1}{\sqrt{T}}\right). \quad (9)$$

Otherwise, $f(x)$ is not guaranteed to converge when $\beta \geq \alpha$ and might diverge at the rate $O(\sqrt{T})$.

Corollary 4.7. *For $\beta_{H_t} < \alpha = \frac{c}{\sqrt{T}}$ for all t, the IFO complexity of Algorithm 1 to obtain an ϵ-accurate solution is:*

$$\textit{IFO calls} = O(1/\epsilon^2). \quad (10)$$

We build intuituions about the convergence condition of the previous tasks $\mathcal{P}$ in Theorem 4.6. As empirically shown in stable A-GEM and stable ER-Reservoir [Mirzadeh et al., 2020], the condition of $\beta_{H_t} < \alpha$ theoretically implies that decaying step size is a key solution to continual learning considering when we pick any arbitrary observation points.

Remark 4.8. To prevent catastrophic forgetting, the step size of $g(x)$, β_{H_t} should be lower than the step size of $f(x)$, α_{H_t}. It should also be noted that $\mathbb{E}_{M_{[1:t]}}[B_t|M_0]$ is not always 0 for any M_0. This implies that, from time step 0, each trial with different given M_0 also has the non-zero cumulative sum $\sum \mathbb{E}_{M_{[1:T]}}[B_t|M_0]$, which occurs overestimating bias theoretically.

The convergence rate with respect to the marginalization on M_0 in Theorem 4.6 exactly match the usual nonconvex SGD rates. The selection rules for M_0 with various memory schemes are important to reduce the variance of convergence rate with having the mean convergence rate as Equation 9 among trials. This is why memory schemes matters in continual learning in terms of variance. Please see more details in Appendix E.

4.3 REFORMULATED PROBLEM OF CONTINUAL LEARNING

The previous section showed the essential factors in continual learning to observe the theoretical convergence rate. The overfitting bias term B_t has a strong dependence on the memory selection rule and can be computed exactly only if we can access the entire dataset P during learning on $\mathcal{C}$. In terms of expectation, we have shown that the effect of B_t is negligible. We also show that its empirical effect is less important than Γ_t in Figure 2. Then we focus on the performance degradation by the catastrophic forgetting term Γ_t. For every trial, the worst-case convergence is dependent on $\Delta_f + \sum_{t=0}^{T-1} \mathbb{E}\left[\Gamma_t\right]$ by Theorem 4.3. To tighten the upper bound and keep the model to be converged, we should minimize the cumulative sum of Γ_t. We now reformulate the continual learning problem 2 as follows.

$$\begin{aligned} \underset{\alpha_{H_t},\, \beta_{H_t}}{\text{minimize}} \quad & \sum_{t=0}^{T-1} \mathbb{E}[\Gamma_t] \\ \text{subject to} \quad & 0 < \beta_{H_t} < \alpha_{H_t} \leq 2/L \text{ for all } t < T \end{aligned} \tag{11}$$

It is noted that the above reformulation presents a theoretically guaranteed continual learning framework for memory-based approaches in nonconvex setting and the constraint is to guarantee the convergence of both $f(x)$ and $g(x)$.

5 ADAPTIVE METHODS FOR CONTINUAL LEARNING

As discussed in the above Section, we can solve a memory-based continual learning by minimizing $\sum_{t=0}^{T-1} \mathbb{E}[\Gamma_t]$. Adaptive methods are variants of SGD, which automatically adjust the step size (learning rate) on a per-feature basis. In this section, we review A-GEM in terms of adaptive methods, and also propose a new algorithm (NCCL) for achieving adaptivity in continual learning. For brevity, we denote the inner product $\langle \nabla f_{I_t}(x^t), \nabla g_{J_t}(x^t) \rangle$ as Λ_{H_t}.

5.1 A-GEM

A-GEM [Chaudhry et al., 2019a] propose a surrogate of $\nabla g_{J_t}(x^t)$ as the following equation to avoid violating the constraint when the case of interference, $\Lambda_{H_t} \leq 0$:

$$\nabla g_{J_t}(x^t) - \left\langle \frac{\nabla f_{I_t}(x^t)}{\|\nabla f_{I_t}(x^t)\|}, \nabla g_{J_t}(x^t) \right\rangle \frac{\nabla f_{I_t}(x^t)}{\|\nabla f_{I_t}(x^t)\|}.$$

Let β be the step size for $g(x)$ when the constraint is not violated. Then we can interpret the surrogate as an adaptive learning rate α_{H_t}, which is $\alpha(1 - \frac{\langle \nabla f_{I_t}(x^t), \nabla g_{J_t}(x^t) \rangle}{\|\nabla f_{I_t}(x^t)\|^2})$ to cancel out the negative component of $\nabla f_{I_t}(x^t)$ on $\nabla g_{J_t}(x^t)$.

For the transfer case $\Lambda_{H_t} > 0$, A-GEM use $\alpha_{H_t} = 0$. After applying the surrogate, $\mathbb{E}[\Gamma_t]$ is reduced as shown in Appendix D. It is noted that A-GEM theoretically violates the constraints of (11) to prevent catastrophic forgetting by letting $\alpha_{H_t} = 0$ and does not utilize the better transfer effect. Then, A-GEM is an adaptive method without theoretical guarantee.

5.2 NCCL

As discussed above, we note that $\mathbb{E}[\Gamma_t]$ is a quadratic polynomial of β_{H_t}. For the interference case $\Lambda_{H_t} \leq 0$, the minimum point of polynomial, $\beta^*_{H_t}$ has a negative value which violates the constraint $\beta_{H_t} > 0$, and $\mathbb{E}[\Gamma_t]$ is monotonically increasing on $\beta_{H_t} > 0$. Then, we instead adapt α_{H_t} to reduce the value of $\mathbb{E}[\Gamma_t]$ at time t by adopting the scheme of A-GEM. The minimum of the polynomial on $\mathbb{E}[\Gamma_t]$ can be obtained when the case of transfer, $\Lambda_{H_t} > 0$ by differentiating on β_{H_t}. Then the minimum $\mathbb{E}[\Gamma^*_t]$ and the optimal step size $\beta^*_{H_t}$ can be obtained as

$$\beta^*_{H_t} = \frac{(1 - \alpha_{H_t} L)\Lambda_{H_t}}{L\|\nabla g_{J_t}(x^t)\|^2}, \quad \mathbb{E}[\Gamma^*_t] = -\frac{(1 - \alpha_{H_t} L)\Lambda_{H_t}}{2L\|\nabla g_{J_t}(x^t)\|^2}.$$

To satisfy the constraints of (11), we should update $\nabla f_{I_t}(x^t)$ with non-zero step size and $\beta_{H_t} < \alpha_{H_t}$ for all t. Then the proposed adaptive method for memory-based approaches is given by

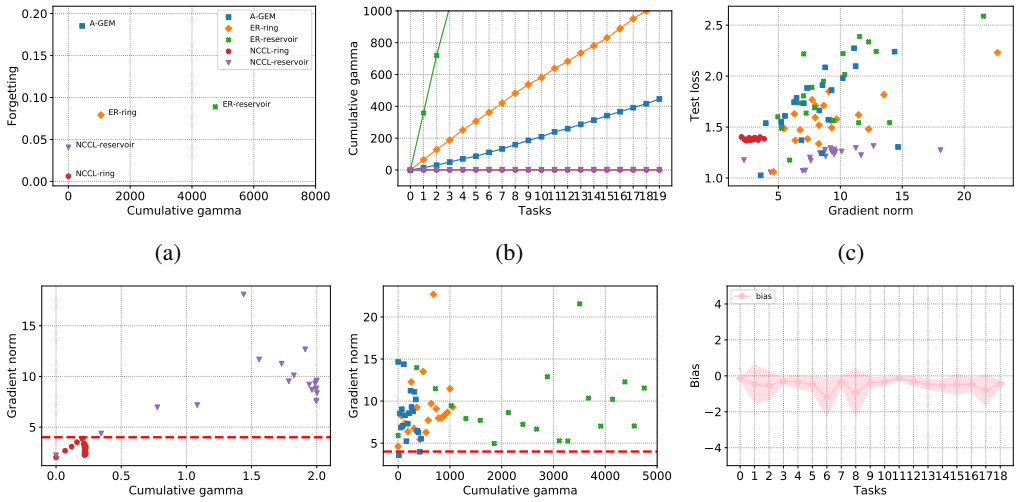
$$\alpha_{H_t} = \begin{cases} \alpha(1 - \frac{\Lambda_{H_t}}{\|\nabla f_{I_t}(x^t)\|^2}), & \Lambda_{H_t} \leq 0 \\ \alpha, & \Lambda_{H_t} > 0, \end{cases}$$

$$\beta_{H_t} = \begin{cases} \alpha, & \Lambda_{H_t} \leq 0 \\ \min\left(\alpha(1-\delta), \frac{(1-\alpha L)\Lambda_{H_t}}{L\|\nabla g_{J_t}(x^t)\|^2}\right), & \Lambda_{H_t} > 0 \end{cases}$$

where $\alpha = c/\sqrt{T}$ and δ is some constant $0 < \delta \ll 1$. Note that our two adaptive learning rates are a stepwise greedy perspective choice of memory-based continual learning.

6 EXPERIMENTS

We use two following metrics to evaluate algorithms. **(1) Average accuracy** is defined as $\frac{1}{T}\sum_{j=1}^{T} a_{T,j}$, where $a_{i,j}$ denotes the test accuracy on task j after training on task i. **(2) Forgetting** is the average maximum forgetting is defined



(a) (b) (c)

(d) (e) (f)

Figure 2: Metrics for continual learning (CL) algorithms trained on split-CIFAR100 with different 5 seeds . (a) Forgetting versus $\sum \mathbb{E}[\Gamma_t]$ at the end of training. (b) Evolution of $\sum \mathbb{E}[\Gamma_t]$ during continual learning. (c) Empirical verification of the relation between $\|\nabla f(x)\|$ for the first task and test loss of the first task in split CIFAR-100. (d)-(e) are the empirical verification of $\sum \mathbb{E}[\Gamma_t]$ versus $\|\nabla f(x)\|$ for the first task in CL algorithms. The red horizontal line indicates the empirical $\|\nabla f(x)\|$ right after training the first task. (f) Illustration of empirical B_t at the end of each task.

as $\frac{1}{T-1}\sum_{j=1}^{T-1} \max_{l\in[T-1]} (a_{l,j} - a_{T,j})$. Due to limited space, we report the details of architecture and learning procedure and missing results with additional datasets in Appendix B.

6.1 EXPERIMENTAL SETUP

Datasets. We demonstrate the experimental results on standard continual learning benckmarks: **Permuted-MNIST** [Kirkpatrick et al., 2017] is a MNIST [LeCun et al., 1998] based dataset, where each task has a fixed permutation of pixels and transform data points by the permutation to make each task distribution unrelated. **Split-MNIST** [Zenke et al., 2017] splits MNIST dataset into five tasks. Each task consists of two classes, for example (1, 7), (3, 4), and has approximately 12K images. **Split-CIFAR10, 100, and Mini-Imagenet** also split versions of CIFAR-10, 100 [Krizhevsky et al., 2009], and MiniImagenet [Vinyals et al., 2016] into five tasks and 20 tasks.

Baselines. We report the experimental evaluation on the online continual setting which implies a model is trained with a single epoch. We compare with the following continual learning baselines. **Fine-tune** is a simple method that a model trains observed data naively without any support, such as replay memory. **Elastic weight consolidation (EWC)** is a regularization based method by Fisher Information [Kirkpatrick et al., 2017]. **ER-Reservoir** chooses samples to store from a data stream with a probability proportional to the number of observed data points. The replay memory returns a random subset of samples at each iteration for experience replay. ER-Reservoir [Chaudhry et al., 2019b] shows a powerful performance in continual learning scenario. **GEM and A-GEM** [Lopez-Paz and Ranzato, 2017, Chaudhry et al., 2019a] use gradient episodic memory to overcome forgetting. The key idea of GEM is gradient projection with quadratic programming and A-GEM simplifies this procedure. We also compare with iCarl, MER, ORTHOG-SUBSPACE [Chaudhry et al., 2020b], stable SGD [Mirzadeh et al., 2020], and MC-SGD [Mirzadeh et al., 2021].

6.2 EXPERIMENT RESULTS

The following tables show our main experimental results, which is averaged over 5 runs. We denote the number of examples per class per task at the top of each column. Overall, NCCL + memory schemes outperform baseline methods especially in the forgetting metric. Our goal is to demonstrate the usefulness of the adaptive methods to reduce the catastrophic forgetting, and to show empirical evidence for our convergence analysis. We remark that NCCL successfully suppress **forgetting by a large margin** compared to baselines. It is noted that NCCL also outperforms A-GEM, which does not maximize transfer when $\Lambda_{H_t} > 0$ and violates the proposed constraints in (11).

We now investigate the proposed terms with regard to memory-based continual learning, $\sum \mathbb{E}[\Gamma_t]$ and B_t. To verify our theoretical analysis, in Figure 2 we show the cumulative catastrophic forgetting term $\sum_t \mathbb{E}[\Gamma_t]$ is the key factor of the convergence of the first task in split-CIFAR100. Dur-

Table 1: Accuary and Forgetitng results between the proposed methods (NCCL+Ring, NCCL+Reservoir) and other baselines in task-incremental learning. When the replay-memory is used, we denote the memory size as the number of examples per class per task. The additional results and the detailed setting with different memory size is in Appendix B.

Method	**dataset**	**Permuted-MNIST**		**split-CIFAR 100**		**split-MiniImagenet**	
	memory size	**5**		**5**		**1**	
	memory	accuracy	forgetting	accuracy	forgetting	accuracy	forgetting
Fine-tune	✗	47.9	0.29(0.01)	40.4(2.83)	0.31(0.02)	36.1(1.31)	0.24(0.03)
EWC	✗	63.1(1.40)	0.18(0.01)	42.7(1.89)	0.28(0.03)	34.8(2.34)	0.24(0.04)
stable SGD	✗	80.1 (0.51)	0.09 (0.01)	59.9(1.81)	0.08(0.01)	-	-
MC-SGD	✗	85.3 (0.61)	0.06 (0.01)	63.3 (2.21)	0.06 (0.03)	-	-
A-GEM	✓	64.1(0.74)	0.19(0.01)	59.9(2.64)	0.10(0.02)	42.3(1.42)	0.17(0.01)
ER-Ring	✓	75.8(0.24)	0.07(0.01)	62.6(1.77)	0.08(0.02)	49.8(2.92)	0.12(0.01)
ER-Reservoir	✓	76.2(0.38)	0.07(0.01)	**65.5(1.99)**	0.09(0.02)	44.4(3.22)	0.17(0.02)
ORHOG-subspace	✓	84.32(1.1)	0.11(0.01)	64.38(0.95)	0.055(0.007)	**51.4(1.44)**	0.10(0.01)
NCCL + Ring	✓	84.41(0.32)	0.053(0.002)	61.09(1.47)	**0.02(0.01)**	45.5(0.245)	**0.041(0.01)**
NCCL+Reservoir	✓	**88.22(0.26)**	**0.028(0.003)**	63.68(0.18)	0.028(0.009)	41.0(1.02)	0.09 (0.01)
Multi-task		91.3	0	71	0	65.1	0

ing continual learning, $\sum_t \mathbb{E}[\Gamma_t]$ increases in all methods of Figure 2b. Figrure 2a, 2d, 2e show that the larger $\sum_t \mathbb{E}[\Gamma_t]$ causes the larger forgetting and $\|\nabla f(x)\|$ for the first task. We can observe that $\|\nabla f(x)\|$ gets larger than 4, which is for the red line, when $\sum_t \mathbb{E}[\Gamma_t]$ becomes larger than 2. We also verify that the theoretical result $\mathbb{E}_t[B_t] = 0$ is valid in Figure 2f. It implies that the empirical results of Lemma 4.1, which show the effect of B_t on Equation 7. Furthermore, the memory bias helps to tighten the convergence rate of P by having negative values in practice. Even with tiny memory, the estimated B_t has much smaller value than $\mathbb{E}[\Gamma_t]$ as we can observe in Figure 2. For experience replay, we need not to worry about the degradation by memory bias and would like to emphasize that tiny memory can slightly help to keep the convergence on P empirically. We conclude that the overfitting bias term might not be a major factor in degrading the performance of continual learning agent when it is compared to the catastrophic forgetting term Γ_t. Next, we modify the clipping bound of β_{H_t} in Section of adaptive methods to resolve the lower performance in terms of average accuracy. In Table 1, NCCL+Ring does not have the best average accuracy score, even though it has the lowest value of $\sum \mathbb{E}[\Gamma_t]$. As we discussed earlier, it is because the convergence rate of C is slower than vanilla ER-Ring with the fixed step sizes. Now, we remove the restriction of β_{H_t}, $\min\left(\alpha(1-\delta), \frac{(1-\alpha L)\Lambda_{H_t}}{L\|\nabla g_{J_t}(x^t)\|^2}\right)$ for $\Lambda_{H_t} > 0$, and instead apply the maximum clipping bound β_{max} to maximize the transfer effect, which occurs if $\Lambda_{H_t} > 0$, by getting $\mathbb{E}[\Gamma_t^*]$. In the original version, we force $\beta_{H_t} < \alpha$ to reduce theoretical catastrophic forgetting term completely. However, replacing with β_{max} is helpful in terms of average accuracy as shown in Appendix B. It means that β_{max} is a hyperparameter to increase the average accuracy by balancing between forgetting on P and learning on C. In Appendix B, we add more results with larger sizes of memory, which shows that NCCL outperforms in terms of average accuracy. It means that estimating transfer and interference in terms of Λ_{H_t} to alleviate forgetting by the small memory for NCCL is less effective.

Table 2: Results of class incremental split-CIFAR100 with Memory size = 10,000.

Methods	**accuracy**
Finetune	3.06(0.2)
A-GEM	2.40(0.2)
GSS-Greedy [Aljundi et al., 2019b]	19.53(1.3)
MIR [Aljundi et al., 2019a]	20.02(1.7)
ER + GMED [Jin et al., 2020]	20.93(1.6)
MIR + GMED [Jin et al., 2020]	21.22(1.0)
NCCL-Reservoir (ours)	**21.95(0.3)**

7 CONCLUSION

We have presented a theoretical convergence analysis of continual learning. Our proof shows that a training model can circumvent catastrophic forgetting by suppressing catastrophic forgetting term in terms of the convergence on previous task. We demonstrate theoretically and empirically that adaptive methods with memory schemes show the better performance in terms of forgetting. It is also noted that there exist two factors on the convergence of previous task: catastrophic forgetting and overfitting to memory. Finally, it is expected the proposed nonconvex framework is helpful to analyze the convergence rate of CL algorithms.

Author Contributions

Seungyub Han initated and led this work, implemented the proposed method, and wrote the paper. Yeongmo Kim set up and ran baseline tests. Taehyun Cho helped to write the paper and verified mathematical proofs. Jungwoo Lee advised on this work. Corresponding author: Jungwoo Lee (e-mail: junglee@snu.ac.kr)

Acknowledgements

This work was supported in part by National Research Foundation of Korea(NRF) grants funded by the Korea Government(MSIT) (No. 2021R1A2C2014504 and No. 2021M3F3A2A02037893), in part by Institute of Information & communications Technology Planning & Evaluation(IITP) grants funded by the Korea Government(MSIT) (No. 2021-0-00106 (20%), No. 2021-0-02068 (20%), and No. 2021-0-01059 (20%)), and in part by INMAC and BK21 FOUR program.

References

Rahaf Aljundi, Lucas Caccia, Eugene Belilovsky, Massimo Caccia, Min Lin, Laurent Charlin, and Tinne Tuytelaars. Online continual learning with maximally interfered retrieval. *CoRR*, abs/1908.04742, 2019a. URL `http://arxiv.org/abs/1908.04742`.

Rahaf Aljundi, Min Lin, Baptiste Goujaud, and Yoshua Bengio. Gradient based sample selection for online continual learning. In *Advances in Neural Information Processing Systems*, pages 11816–11825, 2019b.

Arslan Chaudhry, Marc'Aurelio Ranzato, Marcus Rohrbach, and Mohamed Elhoseiny. Efficient lifelong learning with A-GEM. In *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*. OpenReview.net, 2019a.

Arslan Chaudhry, Marcus Rohrbach, Mohamed Elhoseiny, Thalaiyasingam Ajanthan, Puneet K Dokania, Philip HS Torr, and Marc'Aurelio Ranzato. On tiny episodic memories in continual learning. *arXiv preprint arXiv:1902.10486*, 2019b.

Arslan Chaudhry, Albert Gordo, Puneet K Dokania, Philip Torr, and David Lopez-Paz. Using hindsight to anchor past knowledge in continual learning. *arXiv preprint arXiv:2002.08165*, 3, 2020a.

Arslan Chaudhry, Naeemullah Khan, Puneet Dokania, and Philip Torr. Continual learning in low-rank orthogonal subspaces. *Advances in Neural Information Processing Systems*, 33, 2020b.

Sayna Ebrahimi, Mohamed Elhoseiny, Trevor Darrell, and Marcus Rohrbach. Uncertainty-guided continual learning with bayesian neural networks. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net, 2020.

Robert M. French and Nick Chater. Using noise to compute error surfaces in connectionist networks: A novel means of reducing catastrophic forgetting. *Neural Computation*, 14(7):1755–1769, 2002.

Saeed Ghadimi and Guanghui Lan. Stochastic first-and zeroth-order methods for nonconvex stochastic programming. *SIAM Journal on Optimization*, 23(4):2341–2368, 2013.

Xisen Jin, Arka Sadhu, Junyi Du, and Xiang Ren. Gradient-based editing of memory examples for online task-free continual learning. *arXiv preprint arXiv:2006.15294*, 2020.

James Kirkpatrick, Razvan Pascanu, Neil Rabinowitz, Joel Veness, Guillaume Desjardins, Andrei A Rusu, Kieran Milan, John Quan, Tiago Ramalho, Agnieszka Grabska-Barwinska, et al. Overcoming catastrophic forgetting in neural networks. *Proceedings of the national academy of sciences*, 114(13):3521–3526, 2017.

Jeremias Knoblauch, Hisham Husain, and Tom Diethe. Optimal continual learning has perfect memory and is np-hard. *arXiv preprint arXiv:2006.05188*, 2020.

Alex Krizhevsky, Geoffrey Hinton, et al. Learning multiple layers of features from tiny images. 2009.

Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998.

Kibok Lee, Kimin Lee, Jinwoo Shin, and Honglak Lee. Overcoming catastrophic forgetting with unlabeled data in the wild. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 312–321, 2019.

Lihua Lei, Cheng Ju, Jianbo Chen, and Michael I Jordan. Non-convex finite-sum optimization via scsg methods. In *Advances in Neural Information Processing Systems*, pages 2348–2358, 2017.

Zhizhong Li and Derek Hoiem. Learning without forgetting. *IEEE transactions on pattern analysis and machine intelligence*, 40(12):2935–2947, 2017.

David Lopez-Paz and Marc'Aurelio Ranzato. Gradient episodic memory for continual learning. In *Advances in Neural Information Processing Systems*, pages 6467–6476, 2017.

Seyed Iman Mirzadeh, Mehrdad Farajtabar, Razvan Pascanu, and Hassan Ghasemzadeh. Understanding the role of training regimes in continual learning. In *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020*, 2020.

Seyed Iman Mirzadeh, Mehrdad Farajtabar, Dilan Gorur, Razvan Pascanu, and Hassan Ghasemzadeh. Linear mode connectivity in multitask and continual learning. In *International Conference on Learning Representations*, 2021.

Cuong V. Nguyen, Yingzhen Li, Thang D. Bui, and Richard E. Turner. Variational continual learning. In *6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Conference Track Proceedings*. OpenReview.net, 2018.

Sylvestre-Alvise Rebuffi, Alexander Kolesnikov, Georg Sperl, and Christoph H Lampert. icarl: Incremental classifier and representation learning. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, pages 2001–2010, 2017.

Sashank J Reddi, Ahmed Hefny, Suvrit Sra, Barnabás Póczos, and Alex Smola. Stochastic variance reduction for nonconvex optimization. In *International conference on machine learning*, pages 314–323, 2016.

Sashank J Reddi, Satyen Kale, and Sanjiv Kumar. On the convergence of adam and beyond. In *International Conference on Learning Representations*, 2018.

Matthew Riemer, Ignacio Cases, Robert Ajemian, Miao Liu, Irina Rish, Yuhai Tu, and Gerald Tesauro. Learning to learn without forgetting by maximizing transfer and minimizing interference. In *International Conference on Learning Representations*, 2018.

Mark B. Ring. *Continual learning in reinforcement environments*. PhD thesis, University of Texas at Austin, TX, USA, 1995. URL `http://d-nb.info/945690320`.

Umut Simsekli, Levent Sagun, and Mert Gurbuzbalaban. A tail-index analysis of stochastic gradient noise in deep neural networks. In *International Conference on Machine Learning*, pages 5827–5837. PMLR, 2019.

Umut Simsekli, Lingjiong Zhu, Yee Whye Teh, and Mert Gurbuzbalaban. Fractional underdamped langevin dynamics: Retargeting sgd with momentum under heavy-tailed gradient noise. In *International Conference on Machine Learning*, pages 8970–8980. PMLR, 2020.

Sebastian Thrun. A lifelong learning perspective for mobile robot control. In *Intelligent Robots and Systems, Selections of the International Conference on Intelligent Robots and Systems 1994, IROS 94, Munich, Germany, 12-16 September 1994*, pages 201–214, 1994. doi: 10.1016/b978-044482250-5/50015-3. URL `https://doi.org/10.1016/b978-044482250-5/50015-3`.

Oriol Vinyals, Charles Blundell, Timothy Lillicrap, koray kavukcuoglu, and Daan Wierstra. Matching networks for one shot learning. In D. Lee, M. Sugiyama, U. Luxburg, I. Guyon, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 29. Curran Associates, Inc., 2016. URL `https://proceedings.neurips.cc/paper/2016/file/90e1357833654983612fb05e3ec9148c-Paper.pdf`.

Jaehong Yoon, Eunho Yang, Jeongtae Lee, and Sung Ju Hwang. Lifelong learning with dynamically expandable networks. In *6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Conference Track Proceedings*. OpenReview.net, 2018.

Manzil Zaheer, Sashank Reddi, Devendra Sachan, Satyen Kale, and Sanjiv Kumar. Adaptive methods for nonconvex optimization. In *Advances in neural information processing systems*, pages 9793–9803, 2018.

Friedemann Zenke, Ben Poole, and Surya Ganguli. Continual learning through synaptic intelligence. *Proceedings of machine learning research*, 70:3987, 2017.

Jingzhao Zhang, Sai Praneeth Karimireddy, Andreas Veit, Seungyeon Kim, Sashank Reddi, Sanjiv Kumar, and Suvrit Sra. Why are adaptive methods good for attention models? *Advances in Neural Information Processing Systems*, 33, 2020.