# THEORETICAL UNDERSTANDING OF DYNAMIC MIXTURE OF EXPERTS IN CONTINUAL LEARNING

**Anonymous authors**
Paper under double-blind review

## ABSTRACT

The use of Mixture-of-Experts (MoE) architectures to mitigate catastrophic forgetting in continual learning (CL) has attracted increasing research interest. In particular, dynamic MoE frameworks that progressively expand the expert pool achieve strong empirical performance by allocating dedicated capacity to incoming tasks. Despite this success, a rigorous theoretical understanding of their underlying mechanisms remains largely underexplored. This work provides a formal analysis of dynamic MoE in CL, introducing a framework that integrates expert expansion, freezing, and inheritance. Our analysis reveals that as task diversity grows, expert specialization proceeds in distinct phases, motivating a phased expansion mechanism. We prove that only a finite set of experts is required to cover an infinite task stream within any prescribed error tolerance, justifying an efficient expansion strategy. Furthermore, we derive explicit bounds on forgetting and generalization error, showing that dynamic expansion with expert inheritance ensures stable learning and fast adaptation. Experiments on both synthetic and real-world benchmarks validate our theoretical insights and offer principled guidelines for designing scalable MoE-based CL systems.

## 1 INTRODUCTION

Continual learning (CL), also known as lifelong learning, is a paradigm that enables intelligent systems to operate in dynamic real-world environments by learning from a sequence of tasks over time (Golab & Özsu, 2003; Wang et al., 2021; Geng et al., 2020). A central challenge in CL is catastrophic forgetting (Wang et al., 2024; French, 1999; McCloskey & Cohen, 1989), where adapting to new tasks causes a severe drop in performance on previously learned ones. To mitigate forgetting, many methods have been proposed, such as selectively stabilizing important parameters (Kumar et al., 2025; Auddy et al., 2026) or replaying stored old data (Shen et al., 2024; Ding et al., 2026). While empirically effective, these approaches typically rely on a single, fixed-capacity set of shared parameters. This structural constraint introduces persistent representational interference when tasks are learned sequentially, fundamentally limiting the model's ability to accumulate and retain knowledge across diverse data distributions (Abraham & Robins, 2005; Jung et al., 2023; Cai et al., 2025).

Mixture-of-Experts (MoE) models, with their inherent sparse routing mechanism, have recently emerged as a promising architectural solution to this limitation (Zhou et al., 2022; Le et al., 2024; Lu et al., 2025; Cheng et al., 2026b). By distributing distinct tasks to specialized experts, MoE architectures naturally mitigate catastrophic forgetting, which means that previously acquired knowledge is shielded from being overwritten by new task data. Capitalizing on this advantage, a growing body of MoE-based methods has been developed specifically for the CL setting (Yu et al., 2024a; Lu et al., 2025; Liu et al., 2026b). In particular, dynamic expansion strategies are widely adopted, where the pool of experts progressively grows to accommodate novel tasks, allowing the model to scale its capacity as new information arrives (Ye & Bors, 2023; Wang et al., 2025; Yu et al., 2026).

Despite the practical success of dynamic expansion strategies in MoE-based CL, the theoretical mechanisms driving their effectiveness remain underexplored. Existing theoretical analyses, though insightful, primarily focus on static scenarios with a fixed number of tasks and a fixed model architecture (Chen et al., 2022; Li et al., 2025; 2026). This limits their applicability to the realistic and dynamic CL paradigm, where tasks arrive continually and unpredictably, necessitating adaptive model growth. To bridge this gap, we provide a theoretical formulation for dynamic MoE in CL.



Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

Concretely, we consider a CL scenario with a sequentially arriving stream of tasks and propose a dynamic expansion framework built on three key mechanisms: expert expansion, freezing, and inheritance (i.e., initializing new experts from relevant frozen ones). This formulation enables a rigorous characterization of the interaction between task diversity, expert specialization, and dynamic growth of the model. The core of our analysis is to establish formal conditions under which a dynamically expanding MoE can stably learn and retain knowledge from a growing task pool.

Building on this theoretical foundation, our analysis yields principled design guidelines by elucidating the inner workings of dynamic MoE in CL. First, we demonstrate that under substantial task diversity, experts do not specialize all at once; rather, specialization proceeds in distinct phases as tasks accumulate. This finding motivates a phased expansion mechanism that aligns critical operations such as expert expansion and freezing with identifiable developmental stages of the expert pool. Second, we establish an upper bound on the number of experts required to cover a task space of bounded intrinsic complexity. Although the task stream may be infinite, we prove that only a finite number of experts is sufficient for any prescribed error tolerance. This result justifies an efficient expansion strategy in which new experts are added only when necessary and the total number of experts never exceeds the number of successfully frozen ones. Finally, we derive explicit bounds for both the expected forgetting rate and the generalization error throughout the dynamic expansion process. These bounds demonstrate that our framework effectively allocates capacity to emergent task clusters, and that expert inheritance is essential for stable learning and fast convergence on new tasks. Experiments on both synthetic and real-world datasets validate these theoretical insights.

In summary, our contributions are as follows:

- We propose a dynamic MoE expansion framework integrating expert expansion, freezing, and inheritance. Based on this, we reveal that task diversity drives phased expert specialization, motivating a phased expansion mechanism.
- We prove that a bounded number of experts suffices to cover an infinite stream of tasks under any prescribed error tolerance, providing explicit bounds on forgetting and generalization error. These results highlight the necessity of expert freezing and inheritance for stable and efficient learning.
- We validate our theoretical findings on both synthetic and real-world datasets, demonstrate controlled expansion in practice, and translate the analytical insights into actionable guidelines for designing dynamic MoE systems in CL.

## 2 Related Work

**Continual Learning.** Existing approaches to CL can be broadly classified into three categories: regularization-based, replay-based, and architecture-based methods. Regularization-based methods (Kirkpatrick et al., 2017; Kumar et al., 2025; Zheng et al., 2026b) selectively impose constraints on parameter updates to protect knowledge from previous tasks. Replay-based methods (Rebuffi et al., 2017; Zhao et al., 2020; Wang et al., 2022; Smith et al., 2024; Zhang et al., 2024a; Ding et al., 2026) mitigate catastrophic forgetting by storing and rehearsing a subset of past samples. Architecture-based methods (Qin et al., 2021; Lu et al., 2024; Veniat et al., 2020; Zhang et al., 2024b) reduce interference by allocating dedicated or sparsely shared model capacity to different tasks. Although existing CL methods mainly rely on conventional dense architectures, this work instead focuses on the increasingly popular MoE architecture.

**Mixture of Experts.** The MoE model is a scalable architecture for increasing the representational capacity of deep networks (Yuksel et al., 2012; Masoudnia & Ebrahimpour, 2014). To maintain training efficiency and stability, modern implementations employ sparse gating mechanisms that activate only a subset of experts per input (Wang et al., 2020; Dai et al., 2022; Liu et al., 2026a). Parallel to these engineering efforts, theoretical analyses have emerged to elucidate the statistical behavior and estimation properties of MoE systems (Chen et al., 2022; Li et al., 2025; Liu et al., 2026b). These studies clarify how the gating network governs expert specialization and training dynamics (He et al., 2022; Cheng et al., 2026a), as well as how functional partitioning among experts emerges (Nguyen et al., 2024; Guo et al., 2026). Differing from these works, our theoretical analysis focuses specifically on the role of MoE within CL settings.

**CL based on MoE.** Recently, MoE has attracted increasing attention in CL for mitigating interference across tasks. Empirically, early work introduced MoE into task-free CL, allowing experts to specialize

in different data regions (Aljundi et al., 2019), while subsequent studies explored dynamically expanding expert architectures as new tasks arrive (Lee et al., 2020; Chen et al., 2023; Jung & Kim, 2024). On the theoretical side, general CL studies have analyzed forgetting and generalization through optimization dynamics (Zheng et al., 2026a), feature rank and regularization in linearized networks (Prakash et al., 2025), and feature learning for plasticity (Wang et al., 2026). However, MoE-specific analyses typically consider a fixed task set rather than sequentially arriving and potentially new tasks (Chen et al., 2022; Li et al., 2025). To address this gap, we study a dynamically expanding MoE structure under sequential new tasks and provide corresponding theoretical guarantees.

# 3 Problem Setting and Framework Design

Unless otherwise specified, the $L_2$ norm is used throughout. For simplicity, we omit the subscript $t$ when not focusing on the current round. For any variable $x$, we write $x = 1, 2, \ldots, N$ as $x \in [N]$. We define $x = \Omega(y)$ if $\exists\ c_2 > 0$ such that $x > c_2|y|$; $x = \Theta(y)$ if $\exists\ c_1, c_2 > 0$ such that $c_1|y| < x < c_2|y|$; $x = \mathcal{O}(y)$ if $\exists\ c_1 > 0$ such that $x < c_1|y|$; and $x = o(y)$ if $x/y \to 0$.

## 3.1 CL with Linear Regression Tasks

**Task Definition.** We consider a CL process indexed by discrete training rounds $t \in [T]$. At each round $t$, a linear regression task arrives, characterized by a ground truth vector $w_t \in \mathbb{R}^d$. The corresponding dataset $\mathcal{D}_t = (X_t, y_t)$ is generated according to $y_t = X_t^\top w_t$, where $X_t \in \mathbb{R}^{d\times s}$ is the feature matrix of $s$ samples and $y_t \in \mathbb{R}^s$ is the target vector. The over-parameterization ratio is denoted by $r = s/d < 1$. Details of the data generation process are provided below.

**Ground Truth Pool.** To model a stream of continuously arriving tasks, we assume each ground truth $w_t$ is randomly drawn from a growing pool $W_t = \{w_1, w_2, \ldots, w_{N_t}\} \subset \mathbb{R}^d$, where $\|w_n\| = \mathcal{O}(1), \forall n \in [N_t]$. The pool expands periodically: at intervals of $\Delta t$ rounds, a new ground truth $w_{N_t}$ is added, so that $W_t = W_{t-\Delta t} \cup \{w_{N_t}\}$. Each task $n \in [N_t]$ is associated with a feature signal $v_n \in \mathbb{R}^d$ satisfying $\|v_n\| = \mathcal{O}(1)$, which is obtained by normalizing $w_n$. This signal informs the gating mechanism for task identification (see Lemma 2 in Appendix D).

**Task Diversity.** We set the task diversity to approximate a more realistic CL scenario. Since the ground truths give the tasks, the diversity is set by imposing some restrictions on ground truths. Specifically, we assume these ground truths can be grouped into $K_t \ll N_t$ clusters $\{C_i\}_{i=1}^{K_t}$. Within the same cluster, the distance satisfies $\forall n \neq n', \|w_n - w_{n'}\| = \mathcal{O}(\sigma_0^\alpha)$ with $\alpha > 2$. For ground truths from different clusters, we set $\|w_n - w_{n'}\| = \Theta(\sigma_0)$, where $\sigma_0 \in (0, 1)$ is a fixed constant.

**Dataset Generation.** At round $t$, a ground truth $w_n$ is sampled uniformly from $W_t$ and assigned as $w_t$. The dataset $\mathcal{D}_t = (X_t, y_t)$ is then generated as follows, following existing setups (Chen et al., 2022; Li et al., 2025; Kanemitsu et al., 2016). A random feature vector (column) is set to $cv_n$, where $c = \mathcal{O}(1)$ is a bounded constant and $v_n$ is the signal vector. The remaining $s - 1$ noise vectors are drawn independently from $\mathcal{N}(0, \sigma_t^2 I_{d\times d})$, where $\sigma_t$ denotes the noise level. The targets are given by $y_t = X_t^\top w_t$, completing the dataset. We can address this binary classification sub-problem over $X_t$ using a gating network in MoE (Shazeer et al., 2017; Fedus et al., 2022; Chen et al., 2022), which is the context of our framework.

## 3.2 Structure of the MoE Model

**Experts and Gating Network.** As illustrated in Figure 1, we employ a MoE model to address the continual sequence of linear regression tasks described in Section 3.1. At training round $t$, the model consists of $M_t$ experts, denoted by the set $E_t = \{m_1, m_2, \ldots, m_{M_t}\}$. The parameters of these experts are collected into a matrix $[w_t^{(1)}, w_t^{(2)}, \ldots, w_t^{(M_t)}]$, where each $w_t^{(m)} \in \mathbb{R}^d$ corresponds to expert $m \in [M_t]$. The routing of inputs to experts is governed by a gating network parameterized by $\Theta_t = [\theta_t^{(1)}, \theta_t^{(2)}, \ldots, \theta_t^{(M_t)}]$, with each $\theta_t^{(m)} \in \mathbb{R}^d$ associated with expert $m$.

**Routing Mechanism.** Given an input feature matrix $X_t \in \mathbb{R}^{d\times s}$, the gating network computes a score for each expert as $h_t^{(m)} = \sum_{i\in[s]} (\theta_t^{(m)})^\top X_{t,i}$, where $X_{t,i}$ denotes the $i$-th column of $X_t$. In practice, we add a perturbation term to the scores to smooth routing; details are provided in

Appendix A. The scores are normalized via a softmax function to produce routing probabilities:

$$\pi_t^{(m)} = \frac{\exp(h_t^{(m)})}{\sum_{j=1}^{M_t} \exp(h_t^{(j)})},\ m \in [M_t]. \tag{1}$$

For analytical tractability, we adopt a top-1 routing scheme.

### 3.3 Dynamic Expansion Framework for MoE

**Expert and Gating Network Updates.** Once the gating network selects expert $m_t$ according to the top-1 routing scheme, that expert receives the dataset $\mathcal{D}_t$ and updates its parameter $w_t^{(m_t)}$ on the mean squared error loss:

$$\mathcal{L}_t^{\text{MSE}}(w_t^{(m_t)}, \mathcal{D}_t) = \frac{1}{s}\|X_t^\top w_t^{(m_t)} - y_t\|^2. \tag{2}$$

After updating the expert, the gating network parameters $\Theta_t$ are updated, using gradient descent with learning rate $\eta$. We adopt the following composite loss, which comprises a local term and an auxiliary term:

$$\begin{aligned} \mathcal{L}_t^{\text{task}}(\Theta_t, \mathcal{D}_t) &= \mathcal{L}_t^{\text{loc}}(\Theta_t, \mathcal{D}_t) + \mathcal{L}_t^{\text{aux}}(\Theta_t, \mathcal{D}_t), \\ \mathcal{L}_t^{\text{loc}}(\Theta_t, \mathcal{D}_t) &= \pi_{m_t}(X_t, \Theta_t)\|w_t^{(m_t)} - w_{t-1}^{(m_t)}\|, \\ \mathcal{L}_t^{\text{aux}}(\Theta_t, \mathcal{D}_t) &= \mu M_t \sum_{m \in [M_t]} f_t^{(m)} \cdot P_t^{(m)}. \end{aligned} \tag{3}$$

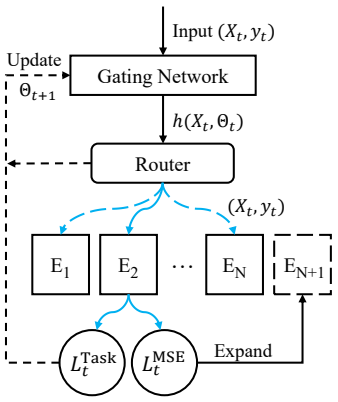



Figure 1: Illustration of the MoE architecture with dynamic expert expansion for CL.

Here $\mu = \mathcal{O}(\sigma_0^{0.5})$ is a constant coefficient; $f_t^{(m)} = \frac{1}{t}\sum_{\tau=1}^{t} \mathbf{1}\{m_\tau = m\}$ denotes the fraction of tasks assigned to expert $m$ up to round $t$; and $P_t^{(m)} = \frac{1}{t}\sum_{\tau=1}^{t} \pi_m(X_\tau, \Theta_\tau) \cdot \mathbf{1}\{m_\tau = m\}$ is the average routing probability assigned to expert $m$ over those rounds. The local loss $\mathcal{L}_t^{\text{loc}}$ accelerates convergence by constraining the magnitude of parameter updates (Li et al., 2025; Sener & Koltun, 2018), while the auxiliary loss $\mathcal{L}_t^{\text{aux}}$ promotes load balancing across experts (Fedus et al., 2022; Shazeer et al., 2017). For each expert $m$, we define $L_t^{(m)} = t f_t^{(m)}$, the cumulative count of tasks processed by expert $m$ by round $t$. Further implementation details are provided in Appendix A.

**Dynamic Expert Expansion.** To adapt to the growing task pool, we employ a dynamic expansion mechanism. When a new task $w_t$ is introduced, the following steps are executed:

(1) Freezing experts that have completed specialization and their associated gating parameters. Following (Li et al., 2025), we define the specialization completion round $T_m$ for expert $m$ as:

$$T_m = \underset{t}{\text{argmin}}\{|h_t^{(m)} - h_t^{(m_t)}| = \mathcal{O}(\sigma_0^{1.25})\}, m \neq m_t; \tag{4}$$

(2) Presenting the new task $w_t$ to all frozen experts, and the mean squared error $\mathcal{L}_t^{\text{MSE}}(w_t^{(m)}, \mathcal{D}_t)$ is computed for each, mimicking evaluation protocols used in empirical CL studies (Chen et al., 2023; Yu et al., 2024b).

(3) If for each frozen expert $m \in [M_t]$, $\mathcal{L}_t^{\text{MSE}}(w_t^{(m)}, \mathcal{D}_t) \geq \epsilon$, a new expert is instantiated. It inherits the parameters of the frozen expert with the highest gating score for the current input.

**Expansion Timeline and Expert Specialization.** We denote the actual expansion rounds as $t_i$ for $i \geq 1$, with $t_0 = 0$. Let $p_i$ be the fraction of experts that are frozen during the interval $[t_i, t_{i+1})$.

Following existing studies (Chen et al., 2022; Li et al., 2025), we introduce a *mark* to describe the association between experts and task clusters. For a given task cluster $n$:

$$\mathcal{M}_n = \{m \in [M_t] | n = \underset{j \in [K_t]}{\text{argmax}}\ (\theta_t^{(m)})^\top v_j\}. \tag{5}$$

Experts in $\mathcal{M}_n$ are said to be specialized for cluster $n$; when needed, we denote them as $m_{n_1}, \ldots, m_{n_k}$. The validity of this definition is supported by Lemma 2 in Appendix D.

# 4 Analysis of Expert Specialization

This section provides a theoretical analysis of the expert specialization mechanism in the presence of diverse tasks. Specifically, we establish that the alignment between task parameters and expert parameters governs the intensity of expert updates (Section 4.1). Building on this, we demonstrate how task diversity, combined with the routing mechanism of the gating network, drives the phased convergence of experts toward specialized configurations (Section 4.2).

## 4.1 Task-Expert Alignment Governs Update Intensity

Consider a task $w_t$ with data $(X_t, y_t)$ and a MoE model with experts $\{m_i\}_{i=1}^M$. As described in Section 3.2, the top-1 expert $(m_t)$ and the gating network are updated in a given round. We begin by stating the following lemma, which characterizes the dominant component of the gradient under different alignment conditions:

**Lemma 4.1.** *(Gradient Domination) Denote $\|w_t^{(m)} - w_{t-1}^{(m)}\| = \mathcal{O}(\sigma_0^{\alpha_m})$. For any training round $t = 1, 2, \cdots$, the gradient norm of expert $m$ satisfies:*

$$\|\nabla_{\theta_t^{(m)}} \mathcal{L}_t^{\text{task}}\| = \begin{cases} \mathcal{O}(\nabla_{\theta_t^{(m)}} \mathcal{L}_t^{\text{loc}}), \ \alpha_m = \mathcal{O}(\ln(\mu M)) \\ \mathcal{O}(\nabla_{\theta_t^{(m)}} \mathcal{L}_t^{\text{aux}}), \alpha_m = \Omega(\ln(\mu M)) \end{cases}, \tag{6}$$

The proof is provided in Appendix C. Lemma 4.1 clarifies the distinct roles of the local and auxiliary loss terms in guiding expert updates. When task diversity is small relative to an expert's current parameters, the gradient is dominated by $\mathcal{L}_t^{\text{aux}}$. In this case, the gating network primarily encourages balanced expert selection to prevent over-specialization. Conversely, when an expert encounters a task with high diversity, the gradient is dominated by $\mathcal{L}_t^{\text{loc}}$. The model then avoids reassigning such tasks in subsequent rounds, allowing the expert to focus on adapting to the current task.

This analysis implies that both the probability of an expert $m_t$ being selected and the magnitude of its update are governed by the degree of its alignment with the current task $w_t$, quantified as $\|w_t - w_t^{(m_t)}\|$. We formalize this intuition in the following proposition, which shows that better-aligned experts receive stronger and more consistent updates, leading to faster convergence toward optimal task parameters. For convenience, we denote the sets of rounds in which the update for expert $m$ is dominated by $\mathcal{L}_t^{\text{loc}}$ and $\mathcal{L}_t^{\text{aux}}$ as $\mathcal{T}_m^{\text{loc}}$ and $\mathcal{T}_m^{\text{aux}}$, respectively.

**Proposition 4.2.** *If $m, m' \in E_t$, $\exists\, t \in \mathbb{N}, \forall\, t' < t, P(m_{t'} = m) = \Omega(P(m_{t'} = m'))$, then*

$$\|h_t^{(m)} - h_t^{(m')}\| = \mathcal{O}((|\mathcal{T}_m^{\text{aux}}| - |\mathcal{T}_{m'}^{\text{aux}}|)\sigma_0) + \mathcal{O}(|\mathcal{T}_{m'}^{\text{loc}}|\sigma_0^{1.5}), \tag{7}$$

The proof is detailed in Appendix D. Proposition 4.2 elucidates the emergence of expert specialization when tasks exhibit significant diversity in early training rounds. According to the definition of expert groups in Eq. equation 5, experts within the same group should produce similar gating outputs. However, the proposition reveals that a discrepancy in task alignment, which is reflected by differing $\alpha_m$ and $\alpha_{m'}$, induces a measurable gap in their gating outputs.

Moreover, Eq. (7) shows that this gap cannot be closed by random future updates. Instead, the model actively reduces the discrepancy by increasing $|\mathcal{T}_{m'}^{\text{aux}}|$, which means that expert $m'$ is assigned more tasks with lower diversity to stabilize its alignment. Meanwhile, the expert with higher alignment, i.e., $m$, maintains more consistent routing behavior. This targeted assignment mechanism leads to divergent convergence rates among experts, which we analyze in the following subsection.

## 4.2 Phased Specialization Driven by Task Diversity

As established in Section 4.1, task diversity induces varying update intensities across experts. Combined with the dynamic routing mechanism of the gating network, this leads to a structured and phased specialization process. To illustrate this, we first characterize the gating network's adjustment mechanism, which directly controls how tasks are distributed among experts.

**Lemma 4.3.** *(Routing Regularization) If $\mathcal{L}_t(\Theta_t, \mathcal{D}_t) = \pi_{m_t}(X_t, \Theta_t) \cdot f$ is the loss for updating the gating network, where $f$ is a positive scalar function independent of $\theta_t^{(m)}$. Then*

$$\langle \theta_{t+1}^{(m)} - \theta_t^{(m)}, v_n \rangle = \begin{cases} -\mathcal{O}(f), m = m_t \\ \mathcal{O}(M_t^{-1} f), m \neq m_t \end{cases} \tag{8}$$

The proof is given in Appendix E. Lemma 4.3 shows how routing probabilities are dynamically regularized. When expert $m_t$ is selected for task $n$, its gating parameters receive a negative update along $v_n$, reducing its future correlation with similar tasks. Conversely, all other experts receive a small positive update, slightly increasing their probability of being selected for analogous inputs. This mechanism prevents over-concentration of similar tasks on a single expert, promoting balanced exposure across the expert pool. Building on this regulated assignment process, the following proposition quantifies the diversity accumulation in task assignment opportunities over time, which ultimately leads to phased specialization of experts.

**Proposition 4.4.** *Let $m$ and $m'$ be experts belonging to the same group $\mathcal{M}_k$, and denote $T_{\min} = \min\{T_m, T_{m'}\}$. For $t < T_{\min}$, let $P(\alpha_m = \Omega(\ln(\mu M))) = p, P(\alpha_{m'} = \Omega(\ln(\mu M))) = 1 - p$. Suppose that the two experts require comparable cumulative effective updates to complete specialization, and that, before their respective specialization completion, their realized effective update frequencies are uniformly of order $p$ and $1 - p$, respectively. Then $\mathbb{E}\left[\frac{T_{m'}}{T_m}\right] = \Theta\left(\frac{p}{1-p}\right)$.*

The proof is detailed in Appendix F. Proposition 4.4 reveals that under significant task diversity, experts within a group converge in a distinct sequential order. Experts with higher initial alignment $m$ receive more frequent relevant task assignments, allowing them to strengthen their specialization rapidly and reach convergence earlier. In contrast, experts with lower initial alignment $m'$ require more rounds to accumulate sufficient targeted updates, resulting in a longer convergence time. This phased convergence has direct implications for model stability and plasticity. Freezing a non-converged expert prematurely disrupts the alignment, as it still requires exposure to similar tasks to stabilize its parameters, whereas freezing a converged expert preserves a stable, specialized functionality. This distinction motivates our segment-based analysis of forgetting and approximation error presented in the following section.

## 5 Analysis of Forgetting and Generalization

Building on the routing mechanism and expert convergence properties established in Section 4, we further explore the role of dynamic expert expansion for mitigating forgetting and enhancing generalization. In particular, we derive explicit expressions for the forgetting rate and generalization error within this framework, highlighting advantages over fixed-expert models. Note that the following analysis relies fundamentally on the mechanisms introduced in Section 4. Here, we first introduce necessary notation and metrics.

**Metrics.** For each training round $t$, the *parameter accuracy* of the selected expert $m_t$ is defined as $\mathcal{E}_t$, which measures its squared distance to the current ground truth $w_t$; the *generalization error* in round $t$ is defined as $\mathcal{G}(w_t, m_t)$, which reflects the prediction gap relative to the true target:

$$\mathcal{E}_t(w_t^{(m_t)}) = \|w_t^{(m_t)} - w_t\|^2,\ \mathcal{G}(w_t, m_t) = \frac{\|X_t^\top w_t^{(m_t)} - y_t\|}{\|X_t\|}, \tag{9}$$

where $X_t^\top w_t = y_t$ by construction.

To evaluate long-term CL performance, we adopt two standard metrics (Lin et al., 2023; Chaudhry et al., 2019): the *forgetting rate* $F_t$ and the *average generalization error* $G_t$, defined respectively as:

$$F_t = \frac{1}{t-1}\sum_{\tau=1}^{t-1}\left(\mathcal{E}_\tau(w_t^{(m_\tau)}) - \mathcal{E}_\tau(w_\tau^{(m_\tau)})\right), G_t = \frac{1}{t}\sum_{\tau=1}^{t}\mathcal{G}(w_{n_\tau}, m_\tau). \tag{10}$$

**Quotient Space.** For a given task $n$ with data $(X_n, y_n)$, different parameter vectors may yield identical predictions. To capture this equivalence, we introduce the quotient space $Q(X_n^\top) = \mathbb{R}^d / \ker(X_n^\top)$, where $\ker(X_n^\top)$ denotes the null space of $X_n^\top$. Let $\phi : \mathbb{R}^d \to Q(X_n)$ be the natural projection and $\varphi : Q(X_n) \to \mathbb{R}^s$ an isomorphism. For any vector $x$ or set $\mathcal{A}$, we denote $\bar{x} = \phi(x)$ and $\bar{\mathcal{A}} = \{\bar{x} \mid x \in \mathcal{A}\}$.

Since the expert marks partition the set of experts, i.e., $\bigcup_{n=1}^{N_t} \mathcal{M}_n = E_t$ with $\mathcal{M}_n \cap \mathcal{M}_{n'} = \varnothing$ for $n \neq n'$, each expert $m_{n_i} \in \mathcal{M}_n$ is uniquely associated with a task cluster. We define a mapping $g(m_{n_i}) = \mathcal{S}_i(w_n)$, where $\mathcal{S}_i(w_n) = \left\{w_{n'} \in \mathbb{R}^d \;\middle|\; \mathcal{G}(w_{n'}, m_{n_i}) < \epsilon_0\right\}$ is the set of ground truth parameter vectors whose generalization error on expert $m_{n_i}$ is below a threshold $\epsilon_0 > 0$. By definition, $w_n \in \mathcal{S}_i(w_n)$.

### 5.1 Benefits and Limitations of Expanding Experts

**Proposition 5.1.** $\forall\, \epsilon_0 > 0$ *and* $t \in \mathbb{N}^+, \exists\, N = \mathcal{O}(1)$, $\{m_n\}_{n=1}^{N},\ s.t.\ W_t \subset \bigcup_{n=1}^{N} g(m_n)$.

The proof is given in Appendix G. By the definition of the mapping $g$, the proposition depends on the gating mechanism's ability to route tasks accurately. Proposition 5.1 implies that although the task space may contain infinitely many tasks, a finite expert set suffices to cover all tasks effectively: for any task, at least one expert in the pool achieves a generalization error below a given threshold $\epsilon_0$. A larger $\epsilon_0$ corresponds to greater error tolerance and thus requires fewer experts, indicating an upper bound on the number of experts during dynamic expansion.

The next proposition quantifies how increasing the number of experts reduces generalization error.

**Proposition 5.2.** *Let* $\bar{S}_m = \bigcup_{i=1}^{m} \bar{S}_i(w_n), n \geqslant 1, 1 \leqslant m \leqslant k_n$. *Let* $\Delta_m = \varphi(\bar{S}_m) - \varphi(\bar{S}_{m-1})$ *and* $\varphi(\bar{S}_0) = 0$. *Then* $\mathbb{E}[\Delta_m]$ *is non-increasing.*

The proof is provided in Appendix H. Proposition 5.2 shows that adding more experts for the same task reduces its generalization error, but the marginal benefit diminishes as $m$ grows. Moreover, by Proposition 5.1, a larger expert pool can blur task boundaries. Together, these observations imply that indiscriminate expansion does not necessarily improve overall CL performance. This justifies the efficient expansion strategy described in Section 3.3, which enforces a strict bound on the total number of experts via an $\epsilon$-threshold and a maximum number of new experts per round.

### 5.2 Forgetting and Generalization under Dynamic Expansion Framework

We derive explicit bounds for the forgetting rate and generalization error under the dynamic expansion framework and compare them with the setting where the total number of experts is directly preset.

**Proposition 5.3.** *Let* $M_t = \Omega(K_t \ln K_t)$, *then for all* $t \in [t_i, t_{i+1})$:

*(1) The expected forgetting rate satisfies*

$$\mathbb{E}[F_t] < \frac{1}{t-1}\sum_{j=0}^{i-1}\Big\{p_j \sum_{\tau=t_{(j)}}^{t_{(j+1)}} \frac{r^{L_t^{(m_\tau)}} - r^{L_\tau^{(m_\tau)}}}{N_t} \sum_{n=1}^{N_t} \|w_n\|^2 + (1-p_j) \sum_{\tau=t_{(j)}}^{t_{(j+1)}} \frac{r^{L_\tau^{(m_\tau)}} - r^{L_t^{(m_\tau)}}}{N_t^2 / \sum_{n=1}^{N_t} \|w_{n'} - w_n\|^2}$$
$$+ \psi p_j \sum_{\tau=t_{(j)}}^{t} (1 - r^{L_t^{(m_\tau)} - L_\tau^{(m_\tau)}}) r^{L_t^{(m_\tau)} - L_{t_{(j-1)}}^{(m_\tau)} - 1}\Big\}, \tag{11}$$

*where* $\psi = \frac{1}{N_t}\sum_{n=1}^{N_t}\sum_{n,n' \in W_k} \frac{\|w_{n'} - w_n\|^2}{|C_k|}$ *represents the average difference between two tasks in a certain cluster. The symbols* $r, L_t^{(m_t)}$ *and* $p_j$ *are defined in Section 3.*

*(2) The expected generalization error satisfies*

$$\mathbb{E}[G_t] < \frac{1}{t}\sum_{j=1}^{i}\Big\{ \sum_{\tau=t_{(j)}}^{t_{(j+1)}} \frac{r^{L_t^{(m_\tau)}}}{N_t} \sum_{n=1}^{N_t} \|w_n\|^2 + \Psi\Big((1-p_j) \sum_{\tau=t_{(j-1)}}^{t_{(j)}} (1 - r^{L_t^{(m_\tau)} - L_{t_{(j)}}^{(m_\tau)}})$$
$$+ p_j \sum_{\tau=t_{(j)}}^{t_{(j+1)}} r^{L_t^{(m_\tau)} - L_{t_{(j)}}^{(m_\tau)}} (1 - r^{L_{t_{(j)}}^{(m_\tau)}})\Big) + \psi p_j \sum_{\tau=t_{(j)}}^{t} (1 - r^{L_t^{(m_\tau)} - L_{t_{(j)}}^{(m_\tau)}})$$
$$+ (1-p_j) \sum_{\tau=t_{(j)}}^{t-1} \frac{r^{L_t^{(m_\tau)} - L_{t_{(j)}}^{(m_\tau)}} (1 - r^{L_{t_{(j)}}^{(m_\tau)}})}{N_t^2 / \sum_{n=1}^{N_t} \|w_{n'} - w_n\|}\Big\}, \tag{12}$$

*where* $\Psi = \frac{1}{N_t K_t}\sum_{n=1}^{N_t}\sum_{k=1}^{K_t} \frac{1}{|C_k|} \sum_{n' \in W_k} \|w_{n'} - w_n\|^2$ *represents the average difference between a randomly selected task in a cluster and a task in another specific cluster.*

The proof is given in Appendix I. Proposition 5.3 and its proof process yield several insights. **Forgetting.** As more tasks arrive, $p_j$ increases and the contribution of the $(1 - p_j)$ terms diminishes, while forgetting becomes dominated by the $p_j$ terms, which capture errors caused by routing tasks to

frozen experts. Under the dynamic expansion and freezing strategy, these contributions gradually decay due to the factor $r < 1$, leading to a reduction in overall forgetting. The remaining forgetting mainly arises from transient inconsistencies in task allocation and the adaptation of newly introduced experts, rather than persistent interference in previously learned knowledge. The parameter inheritance strategy for expanded experts is crucial because $\psi$ would be replaced by a larger quantity $\psi' = \frac{1}{N_t}\sum_{n=1}^{N_t}\sum_{n\in C_k}\frac{\|w_n\|^2}{|C_k|}$, significantly increasing the upper bound on forgetting.

**Generalization error.** The generalization bound additionally depends on the differences between clusters, which is manifested as $\Psi$. Both expert expansion and parameter inheritance help allocate experts appropriately across task clusters, improving generalization. Unlike forgetting, which arises from new-task interference on old experts, generalization error captures fitting accuracy on new task clusters. As noted in Proposition 5.2, increasing the number of tasks strengthens the knowledge coverage of the MoE system, weakening the impact of task diversity and yielding a more robust generalization error under dynamic expansion.

# 6 Empirical Validation

## 6.1 Synthetic-data Experiments

**Settings.** We fix the noise level to $\sigma_0 = 0.4$, and set the total number of training rounds to $T = 2{,}000$. New tasks are introduced every $\Delta t = 100$ rounds. The feature dimension is $d = 10$, and each task contains $s = 6$ samples. The learning rate is set to $\eta = 0.5$. We consider a model with a maximum of 20 experts. The task stream consists of 3 clusters, each containing 10 tasks. Task diversity is controlled by the parameter $\alpha$, where smaller values of $\alpha$ correspond to higher diversity. To evaluate the balance of gating decisions, we compute the average output entropy $H = -\frac{1}{t}\sum_{\tau=1}^{t}\sum_{i=1}^{M}\pi_\tau^{(i)}\log\pi_\tau^{(i)}$. Finally, we vary the error threshold $\epsilon$ that triggers expert expansion, which in turn indirectly determines the final number of experts in the model.

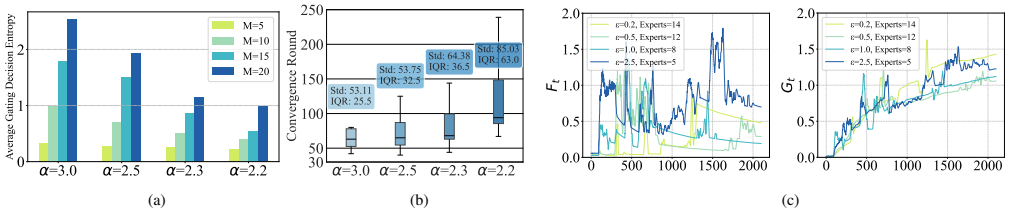
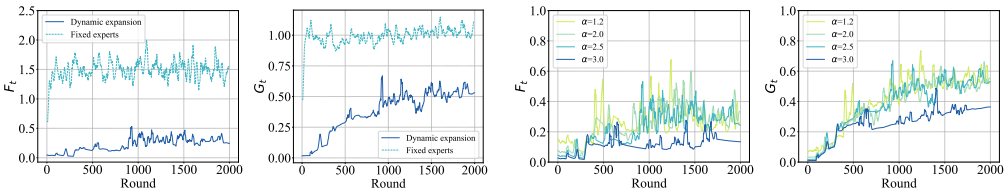



Figure 2: (a) Impact of task diversity on gating decision entropy; (b) Convergence rounds of experts, where smaller $\alpha$ corresponds to greater diversity; (c) Effect of the number of expanded experts on forgetting rate and generalization error, where $\epsilon$ is the error threshold for expansion.

**Results.** Higher task diversity leads to more imbalanced expert routing (lower entropy), while lower diversity results in more synchronized expert convergence. Increasing the number of experts initially reduces forgetting and generalization error, but excessive expansion degrades performance, indicating that the required number of experts is bounded. Moreover, dynamic expert expansion consistently outperforms fixed architectures by better balancing plasticity and stability across different task diversity settings. Additional experimental results are provided in Appendix B.

## 6.2 Experiments on Complex Datasets

**Settings.** We extend our experiments to complex datasets, including MNIST (LeCun et al., 1989), CIFAR-10 (Krizhevsky et al., 2009), CIFAR-100 and Tiny-ImageNet (Le & Yang, 2015). Every 10 rounds, a new binary classification task is introduced, each randomly assigned 500 training and 500 test samples. For MNIST and CIFAR-10, we adopt neural networks with two convolutional layers and three fully connected layers as experts and the learning rate $\eta$ is 0.05. While for CIFAR-100 and Tiny-ImageNet, ResNet-18 is used as experts and $\eta$ is 0.005. The error threshold that triggers the expansion of experts varies from 0.1 to 0.45. Upon arrival of each new task, we apply the dynamic expansion strategy outlined in Section 3.3. To evaluate performance on these sequential classification tasks, we adapt the definitions of forgetting rate and generalization error.




(a) (b)

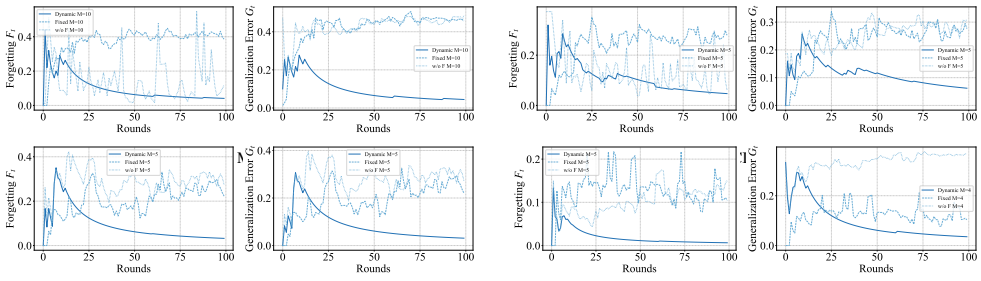
Figure 3: (a) Comparison of forgetting rate and generalization error between the dynamic expansion framework and a fixed-expert model; (b) Performance of the dynamic expansion model under settings of different task diversities.

Let $\mathcal{E}_t(y_t^{(m_t)}) = \frac{1}{N}\sum_{i=1}^{N} \|y_i^{(m_t)} - y_i\|$, where $N$ is the number of samples, $y_i^{(m_t)}$ is the prediction from the expert assigned at round $t$, and $y_i$ is the corresponding one-hot label. The cumulative generalization error up to round $t$ is then defined as $G_t = \frac{1}{t}\sum_{\tau=1}^{t} \mathcal{E}_\tau(y_\tau^{(m_\tau)})$.

**Results.** The results in Table 1 show that as the number of expanded experts increases, the performance shows a trend of increasing first and then declining, indicating that the optimal number of experts required is limited. In Figure 4, the dynamic expansion framework demonstrates superior stability compared to the fixed approach. Specifically, despite different neural networks of experts, the expansion framework substantially mitigates catastrophic forgetting, reducing the rate to below 0.1 in the final stages. Furthermore, it maintains a consistently low generalization error throughout the training process, confirming the effectiveness of the framework on real-world sequential tasks.

Table 1: Test accuracy under different $\epsilon$ settings

| $\epsilon$ | 0.10 | 0.15 | 0.20 | 0.25 | 0.30 | 0.35 | 0.40 | 0.45 |
|---|---|---|---|---|---|---|---|---|
| CIFAR100 | 0.7415 | 0.7464 | 0.7512 | 0.8120 | 0.8153 | 0.7324 | 0.7262 | 0.6464 |
| Tiny-ImageNet | 0.6202 | 0.6675 | 0.6702 | 0.7107 | 0.6821 | 0.6206 | 0.6267 | 0.5561 |




(c) CIFAR-100 (d) Tiny-ImageNet

Figure 4: Comparison of the forgetting rate and the generalization error between dynamic expansion framework and fixed expert framework on MNIST, CIFAR-10, CIFAR-100 and Tiny-ImageNet.

## 7 Conclusion

This work presents a theoretical analysis of dynamic MoE in a CL scenario. By introducing a dynamic expansion framework that integrates expert expansion, freezing, and inheritance, we bridge a critical gap between the empirical success of these methods and their underlying theoretical foundations. Our analysis reveals that task diversity drives phased expert specialization, and that a bounded number of experts suffices to cover an infinite task stream under any prescribed error tolerance, with explicit guarantees on forgetting and generalization. Experiments on both synthetic and real-world datasets validate these theoretical insights.

**Limitations.** The current analysis focuses on the task-incremental setting; extending it to class-incremental scenarios is an important direction for future work.

## AI USE STATEMENT

In this work, generative AI tools were utilized exclusively for assisting with initial language proofreading and LaTeX template formatting. All core conceptual ideas, theoretical derivations, algorithm designs, code implementations, and experimental formulations were independently conceived, verified, and authored by the researchers. The authors take full responsibility for the entirety of the content, claims, and formulations presented in this paper.

## ETHICS STATEMENT

This research focuses on foundational machine learning algorithms. All datasets used are standard, publicly available academic benchmarks. The proposed method does not generate harmful content, involve sensitive personal data, or introduce societal bias.

## REFERENCES


Wickliffe C Abraham and Anthony Robins. Memory retention–the synaptic stability versus plasticity dilemma. *Trends in neurosciences*, 28(2):73–78, 2005.

Rahaf Aljundi, Klaas Kelchtermans, and Tinne Tuytelaars. Task-free continual learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 11254–11263, 2019.

Sayantan Auddy, Jakob Hollenstein, Matteo Saveriano, Antonio Rodríguez-Sánchez, and Justus Piater. Scalable and efficient continual learning from demonstration via a hypernetwork-generated stable dynamics model. *IEEE Transactions on Cognitive and Developmental Systems*, 2026.

Weilin Cai, Juyong Jiang, Fan Wang, Jing Tang, Sunghun Kim, and Jiayi Huang. A survey on mixture of experts in large language models. *IEEE Transactions on Knowledge and Data Engineering*, 2025.

Arslan Chaudhry, Ranzato Marc'Aurelio, Marcus Rohrbach, and Mohamed Elhoseiny. Efficient lifelong learning with a-gem. In *7th International Conference on Learning Representations, ICLR 2019*. International Conference on Learning Representations, ICLR, 2019.

Wuyang Chen, Yanqi Zhou, Nan Du, Yanping Huang, James Laudon, Zhifeng Chen, and Claire Cui. Lifelong language pretraining with distribution-specialized experts. In *International Conference on Machine Learning*, pp. 5383–5395. PMLR, 2023.

Zixiang Chen, Yihe Deng, Yue Wu, Quanquan Gu, and Yuanzhi Li. Towards understanding the mixture-of-experts layer in deep learning. *Advances in neural information processing systems*, 35: 23049–23062, 2022.

Anzhe Cheng, Shukai Duan, Shixuan Li, Chenzhong Yin, Mingxi Cheng, Heng Ping, Tamoghna Chattopadhyay, Sophia I Thomopoulos, Shahin Nazarian, Paul Thompson, et al. Ermoe: Eigen-reparameterized mixture-of-experts for stable routing and interpretable specialization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 12997–13006, 2026a.

Ximeng Cheng, Qiaonan Zhu, Shukun Jia, Yichao Cao, and Xiaobo Lu. Incremental mixture of experts: Continual learning for object detection in forestry scenarios. *Pattern Recognition*, pp. 113001, 2026b.

Damai Dai, Li Dong, Shuming Ma, Bo Zheng, Zhifang Sui, Baobao Chang, and Furu Wei. Stablemoe: Stable routing strategy for mixture of experts. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 7085–7095, 2022.

Meng Ding, Jinhui Xu, and Kaiyi Ji. Provable effects of data replay in continual learning: A feature learning perspective. *arXiv preprint arXiv:2602.02767*, 2026.

William Fedus, Barret Zoph, and Noam Shazeer. Switch transformers: Scaling to trillion parameter models with simple and efficient sparsity. *Journal of Machine Learning Research*, 23(120):1–39, 2022.

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

Chuanxing Geng, Sheng-jun Huang, and Songcan Chen. Recent advances in open set recognition: A survey. *IEEE transactions on pattern analysis and machine intelligence*, 43(10):3614–3631, 2020.

Lukasz Golab and M Tamer Özsu. Issues in data stream management. *ACM Sigmod Record*, 32(2): 5–14, 2003.

Hongcan Guo, Haolang Lu, Guoshun Nan, Bolun Chu, Jialin Zhuang, Yuan Yang, Wenhao Che, Xinye Cao, Sicong Leng, Qimei Cui, et al. Advancing expert specialization for better moe. *Advances in Neural Information Processing Systems*, 38:48767–48809, 2026.

Jiaao He, Jidong Zhai, Tiago Antunes, Haojie Wang, Fuwen Luo, Shangfeng Shi, and Qin Li. Fastermoe: modeling and optimizing training of large-scale dynamic pre-trained models. In *Proceedings of the 27th ACM SIGPLAN Symposium on Principles and Practice of Parallel Programming*, pp. 120–134, 2022.

Dahuin Jung, Dongjin Lee, Sunwon Hong, Hyemi Jang, Ho Bae, and Sungroh Yoon. New insights for the stability-plasticity dilemma in online continual learning. In *11th International Conference on Learning Representations, ICLR 2023*, 2023.

Min Jae Jung and JooHee Kim. Pmoe: Progressive mixture of experts with asymmetric transformer for continual learning. *arXiv preprint arXiv:2407.21571*, 2024.

Hidehiro Kanemitsu, Masaki Hanada, and Hidenori Nakazato. Clustering-based task scheduling in a large number of heterogeneous processors. *IEEE Transactions on Parallel and Distributed Systems*, 27(11):3144–3157, 2016.

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.

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

Saurabh Kumar, Henrik Marklund, and Benjamin Van Roy. Maintaining plasticity in continual learning via regenerative regularization. In *Conference on Lifelong Learning Agents*, pp. 410–430. PMLR, 2025.

Minh Le, An Nguyen, Huy Nguyen, Trang Nguyen, Trang Pham, Linh Van Ngo, and Nhat Ho. Mixture of experts meets prompt-based continual learning. *Advances in Neural Information Processing Systems*, 37:119025–119062, 2024.

Ya Le and Xuan S. Yang. Tiny imagenet visual recognition challenge. In *Stanford CS231n Project Report*, 2015.

Yann LeCun, Bernhard Boser, John Denker, Donnie Henderson, Richard Howard, Wayne Hubbard, and Lawrence Jackel. Handwritten digit recognition with a back-propagation network. *Advances in neural information processing systems*, 2, 1989.

Soochan Lee, Junsoo Ha, Dongsu Zhang, and Gunhee Kim. A neural dirichlet process mixture model for task-free continual learning. In *8th International Conference on Learning Representations, ICLR 2020*, 2020.

Haoran Li, Jingfeng Wu, and Vladimir Braverman. Memory-statistics tradeoff in continual learning with structural regularization. In *International Conference on Learning Representations*, volume 2026, pp. 63493–63534, 2026.

Hongbo Li, Sen Lin, Lingjie Duan, Yingbin Liang, and Ness B Shroff. Theory on mixture-of-experts in continual learning. International Conference on Learning Representations (ICLR), 2025.

Sen Lin, Peizhong Ju, Yingbin Liang, and Ness Shroff. Theory on forgetting and generalization of continual learning. In *International Conference on Machine Learning*, pp. 21078–21100. PMLR, 2023.

Jiacheng Liu, Peng Tang, Wenfeng Wang, Yuhang Ren, Xiaofeng Hou, Pheng Ann Heng, Minyi Guo, and Chao Li. A survey on inference optimization techniques for mixture of experts models. *ACM Computing Surveys*, 58(10):1–37, 2026a.

Yang Liu, Toan Nguyen, and Flora D Salim. Cp-moe: Consistency-preserving mixture-of-experts for continual learning. *arXiv preprint arXiv:2605.20247*, 2026b.

Aojun Lu, Tao Feng, Hangjie Yuan, Xiaotian Song, and Yanan Sun. Revisiting neural networks for continual learning: An architectural perspective. *arXiv preprint arXiv:2404.14829*, 2024.

Yue Lu, Shizhou Zhang, De Cheng, Guoqiang Liang, Yinghui Xing, Nannan Wang, and Yanning Zhang. Training consistent mixture-of-experts-based prompt generator for continual learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 19152–19160, 2025.

Saeed Masoudnia and Reza Ebrahimpour. Mixture of experts: a literature survey. *Artificial Intelligence Review*, 42(2):275–293, 2014.

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

Hien Duy Nguyen, TrungTin Nguyen, and Florence Forbes. Bayesian likelihood free inference using mixtures of experts. In *2024 International Joint Conference on Neural Networks (IJCNN)*, pp. 1–8. IEEE, 2024.

Arjun Prakash, Naicheng He, Kaicheng Guo, Saket Tiwari, Ruo Yu Tao, Tyrone Serapio, Amy Greenwald, and George Konidaris. Spectral collapse drives loss of plasticity in deep continual learning. *arXiv preprint arXiv:2509.22335*, 2025.

Qi Qin, Wenpeng Hu, Han Peng, Dongyan Zhao, and Bing Liu. Bns: Building network structures dynamically for continual learning. *Advances in Neural Information Processing Systems*, 34: 20608–20620, 2021.

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*, pp. 2001–2010, 2017.

Ozan Sener and Vladlen Koltun. Multi-task learning as multi-objective optimization. *Advances in neural information processing systems*, 31, 2018.

Noam Shazeer, Azalia Mirhoseini, Krzysztof Maziarz, Andy Davis, Quoc Le, Geoffrey Hinton, and Jeff Dean. Outrageously large neural networks: The sparsely-gated mixture-of-experts layer. In *International Conference on Learning Representations*, 2017.

Jiangrong Shen, Wenyao Ni, Qi Xu, and Huajin Tang. Efficient spiking neural networks with sparse selective activation for continual learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pp. 611–619, 2024.

James Seale Smith, Lazar Valkov, Shaunak Halbe, Vyshnavi Gutta, Rogerio Feris, Zsolt Kira, and Leonid Karlinsky. Adaptive memory replay for continual learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 3605–3615, 2024.

Tom Veniat, Ludovic Denoyer, and Marc'Aurelio Ranzato. Efficient continual learning with modular networks and task-driven priors. *arXiv preprint arXiv:2012.12631*, 2020.

FY Wang, DW Zhou, HJ Ye, and DC Zhan Foster. Feature boosting and compression for class-incremental learning. *ECCV FOSTER*, 2022.

Huiyi Wang, Haodong Lu, Lina Yao, and Dong Gong. Self-expansion of pre-trained models with mixture of adapters for continual learning. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 10087–10098, 2025.

Jiuqi Wang, Jayanth Srinivasa, Claire Chen, Shuze Daniel Liu, Ali Payani, and Shangtong Zhang. Predicting plasticity in deep continual learning: A theoretical perspective. *arXiv preprint arXiv:2605.09044*, 2026.

Liyuan Wang, Mingtian Zhang, Zhongfan Jia, Qian Li, Chenglong Bao, Kaisheng Ma, Jun Zhu, and Yi Zhong. Afec: Active forgetting of negative transfer in continual learning. *Advances in Neural Information Processing Systems*, 34:22379–22391, 2021.

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*, 46(8):5362–5383, 2024.

Xin Wang, Fisher Yu, Lisa Dunlap, Yi-An Ma, Ruth Wang, Azalia Mirhoseini, Trevor Darrell, and Joseph E Gonzalez. Deep mixture of experts via shallow embedding. In *Uncertainty in artificial intelligence*, pp. 552–562. PMLR, 2020.

Fei Ye and Adrian G Bors. Self-evolved dynamic expansion model for task-free continual learning. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 22102–22112, 2023.

Jiazuo Yu, Yunzhi Zhuge, Lu Zhang, Ping Hu, Dong Wang, Huchuan Lu, and You He. Boosting continual learning of vision-language models via mixture-of-experts adapters. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 23219–23230, 2024a.

Ruilong Yu, Mingyan Liu, Fei Ye, Adrian G Bors, Rongyao Hu, et al. Learning expandable and adaptable representations for continual learning. *Advances in Neural Information Processing Systems*, 38:52737–52757, 2026.

Xiaofan Yu, Tajana Rosing, and Yunhui Guo. Evolve: Enhancing unsupervised continual learning with multiple experts. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, pp. 2366–2377, 2024b.

Seniha Esen Yuksel, Joseph N Wilson, and Paul D Gader. Twenty years of mixture of experts. *IEEE transactions on neural networks and learning systems*, 23(8):1177–1193, 2012.

Xiaohui Zhang, Jiangyan Yi, Chenglong Wang, Chu Yuan Zhang, Siding Zeng, and Jianhua Tao. What to remember: Self-adaptive continual learning for audio deepfake detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pp. 19569–19577, 2024a.

Yanyi Zhang, Binglin Qiu, Qi Jia, Yu Liu, and Ran He. Not just object, but state: compositional incremental learning without forgetting. *Advances in Neural Information Processing Systems*, 37: 123182–123206, 2024b.

Bowen Zhao, Xi Xiao, Guojun Gan, Bin Zhang, and Shu-Tao Xia. Maintaining discrimination and fairness in class incremental learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 13208–13217, 2020.

Guodong Zheng, Peng Wang, Shengchao Hu, Quan Zheng, and Li Shen. Understanding the dynamics of forgetting and generalization in continual learning via the neural tangent kernel. In *International Conference on Learning Representations*, volume 2026, pp. 18488–18525, 2026a.

Yaoyue Zheng, Yin Zhang, Joost van de Weijer, Gido van de Ven, Shaoyi Du, Xuetao Zhang, and Zhiqiang Tian. Revisiting weight regularization for low-rank continual learning. In *International Conference on Learning Representations*, volume 2026, pp. 42895–42919, 2026b.

Yanqi Zhou, Tao Lei, Hanxiao Liu, Nan Du, Yanping Huang, Vincent Zhao, Andrew M Dai, Quoc V Le, James Laudon, et al. Mixture-of-experts with expert choice routing. *Advances in Neural Information Processing Systems*, 35:7103–7114, 2022.

## A Details of the Dynamic Expansion Framework

We provide some details of the MoE-based CL framework which is described in Section 3.

**Update of Expert Parameters.** Since $r = s/d < 1$, for each $y_t = X_t^\top w_t$, there are infinitely many solutions for $w_t$. To ensure the uniqueness of the solution in each round $t$, the following update method for expert parameters is adopted:

$$\min_{w_t'} \|w_t' - w_{t-1}^{(m_t)}\|, s.t.\ X_t^\top w_t' = y_t. \tag{13}$$

This method requires expert $m_t$ to select the one with the smallest parameter change (in the sense of the 2-norm) among all possible solutions.

**Perturbations to the Router.** In practice, adding a noise term to the router is a common strategy for MoE design (Fedus et al., 2022; Chen et al., 2022). This design plays a crucial role in the smoothness of routing, which is detailed in Lemma 1. Specifically, for the input feature matrix $X_t$, the gating output for expert $m$ is calculated as follows:

$$h_t^{(m)}(\Theta_t, X_t) = \sum_{i\in[s]} (\theta_t^{(m)})^\top X_{t,i} + r_t^{(m)}, \tag{14}$$

where $r_t^{(m)}$ is a random noise. In our experiments, we adopt this setting.

**Lemma 1**(Smooth Router) Define $h, \hat{h} \in \mathbb{R}^{[M_t]}$ are two gating outputs, $p_m = \mathbb{P}(\underset{m'\in[M_t]}{\text{argmax}}\{h_{m'} + r_{m'}\} = m)$, $\hat{p} = \mathbb{P}(\underset{m'\in[M_t]}{\text{argmax}}\{\hat{h}_{m'} + r_{m'}\} = m)$. The density function of $\{r^{(m)}\}_{m=1}^{M_t}$ has an upper bound $\gamma$. Then $\|p - \hat{p}\|_\infty \leqslant (\gamma M_t^2)\|h - \hat{h}\|_\infty$.

*Proof.* Let $m_1 = \underset{m}{\text{argmax}}\{h_m + r_m\}$, $m_2 = \underset{m}{\text{argmax}}\{\hat{h}_m + r_m\}$. Without loss of generality, we suppose $m_1 \neq m_2$. Then $h_{m_1} - h_{m_2} \geqslant r_{m_2} - r_{m_1} \geqslant \hat{h}_{m_1} - \hat{h}_{m_2}$. Since the distance of $r_{m_2} - r_{m_1}$ from the center does not exceed half the length of the interval, we obtain $|r_{m_1} - r_{m_2} - C(m_1, m_2)| \leqslant \|\hat{h} - h\|_\infty$, where $C(m_1, m_2) = 0.5(h_{m_1} - h_{m_2} + \hat{h}_{m_1} - \hat{h}_{m_2})$. Therefore,

$$\begin{aligned}
\mathbb{P}(m_1 \neq m_2) &= \sum_{a<b} \mathbb{P}(m_1 = a, m_2 = b) && (15)\\
&\leqslant \sum_{m_1<m_2} \mathbb{P}(|r_{m_1} - r_{m_2} - C(m_1, m_2)|) \leqslant \|\hat{h} - h\|_\infty && (16)\\
&= \sum_{m_1<m_2} \mathbb{E}(r_{m_1} \in [r_{m_2} + C(m_1, m_2) \pm \|\hat{h} - h\|_\infty] | r_{m_2}) && (17)\\
&\leqslant \frac{M_t(M_t - 1)}{2} \cdot 2 \cdot \gamma \leqslant \gamma M_t^2 \|\hat{h} - h\|_\infty && (18)
\end{aligned}$$

Finally, we obtain

$$\begin{aligned}
|p_i - \hat{p}_i| &= |\mathbb{P}(\underset{m}{\text{argmax}}\{h_m + r_m\} = i) - \mathbb{P}(\underset{m}{\text{argmax}}\{\hat{h}_m + r_m\} = i)| && (19)\\
&\leqslant \mathbb{E}|\mathbf{1}(\underset{m}{\text{argmax}}\{h_m + r_m\} = i) - \mathbf{1}(\underset{m}{\text{argmax}}\{\hat{h}_m + r_m\} = i)| \leqslant \gamma M^2 \|\hat{h} - h\|_\infty && (20)
\end{aligned}$$

□

**Expand Experts.** Section 3.3 describes a simple case when expanding experts, *i.e.*, if the threshold condition is satisfied, one expert is added to $E_t$. In practice, the number of expanded experts can be further relaxed, but we always require that this number does not exceed the number of experts that have been frozen in the current round. Such a setting ensures that the total number of all unfrozen experts is bounded and will not grow exponentially. In addition, the interval between arriving tasks should not be too short to ensure that there is at least one frozen expert available when each new task arrives.

## B Additional Experimental Results

All experiments are conducted on 4 NVIDIA 3090 GPUs with 24GB memory each.

### B.1 Effect of Parameter Inheritance in Dynamic Expansion

For the dynamic expansion framework, we adopt two initialization methods: 0-initialization and parameter inheritance initialization. We compare the changes in forgetting rate and generalization error under these two scenarios. It can be seen from Figure 5 that parameter inheritance is significantly better in terms of forgetting rate and generalization error.

### B.2 Ablation on the Number of Fixed Experts

In the experiment in Section 6.1, 8 experts were finally obtained using the dynamic expansion framework. Considering that dynamically expanding experts may learn redundant features, the fixed expert model may achieve better results when there are fewer than 8 experts. Therefore, we conduct an experiment with the number of fixed experts set to 1, 4, 8, and 12. Figure 6 shows that the impact caused by a smaller number of fixed experts is not obvious, which further illustrates that the experimental results in Section 6.1 are reliable.

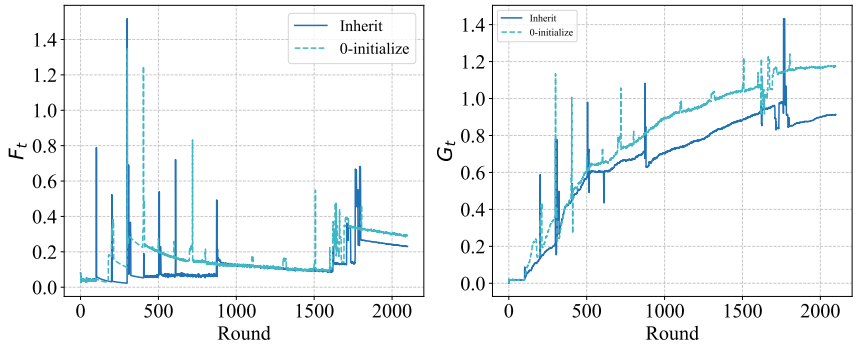



Figure 5: The impact of parameter inheritance mechanism on dynamically expanding experts.

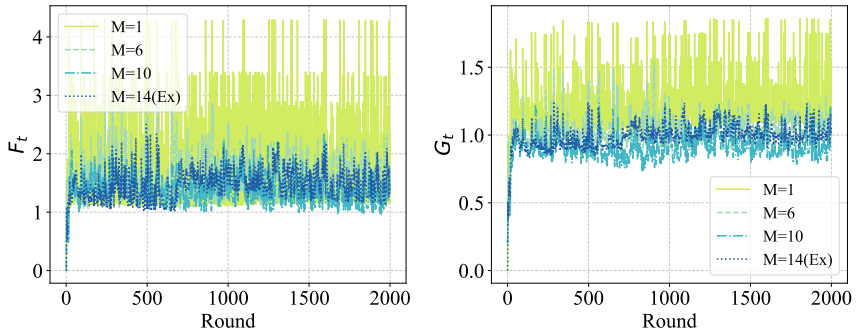



Figure 6: Ablation on the number of fixed experts. A smaller number of fixed experts cannot significantly improve performance.

### B.3 Extended Experiments in Class-Incremental Learning

Although this work focuses mainly on the theoretical analysis of task incremental learning, we also extend the insights in the theory to class incremental learning (CIL) scenarios. Specifically, we conduct standard CIL experiments in fixed expert and dynamic expert settings using CIFAR-100 and Tiny-ImageNet datasets. ResNet-18 is adopted as the expert network, with the learning rate set to 0.02, the weight decay set to 0.9, and the expansion threshold set to 0.85. The results in Figure 7

on the test accuracy and forgetting rate demonstrate that the dynamic expansion framework exhibits great potential in the later stage. Therefore, we believe that migrating the theory of this work to class-incremental scenarios will be a promising research direction.

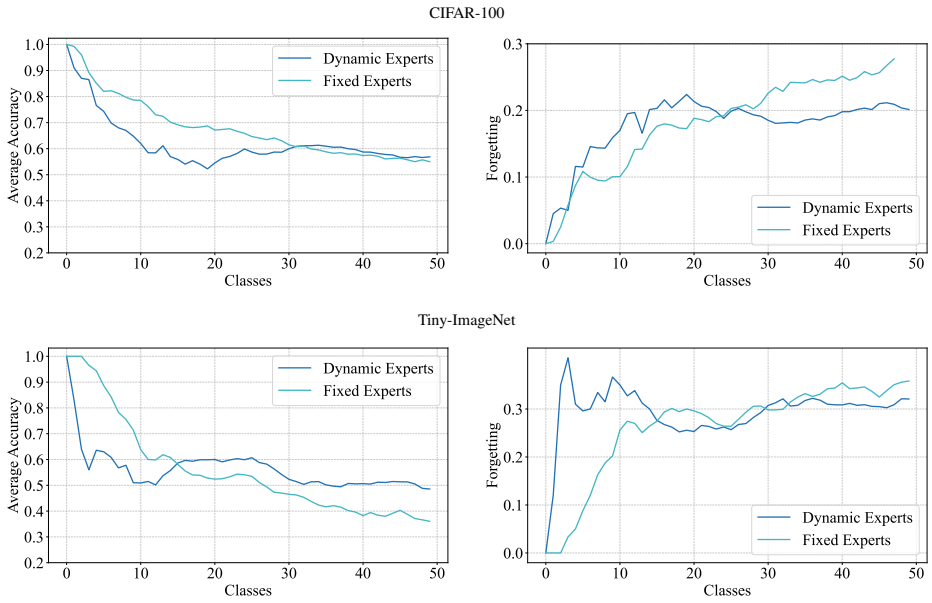



Figure 7: Experiments on the CIL scenarios.

## C Proof of Lemma 4.1

*Proof.* It can be known from the method of Lagrange multipliers that the solution to Eq. (13) is

$$w_t^{(m_t)} = w_{t-1}^{(m_t)} + X_t(X_t^\top X_t)^{-1}(y_t - X_t^\top w_{t-1}^{(m_t)}) \tag{21}$$

Given that $\|X_t\| = \mathcal{O}(1)$, therefore

$$\|w_t^{(m_t)} - w_{t-1}^{(m_t)}\| = \|X_t(X_t^\top X_t)^{-1}X_t^\top(w_t - w_{t-1}^{(m_t)})\| = \mathcal{O}(\sigma_0^{\alpha_m}). \tag{22}$$

Considering that

$$\nabla_{\theta_t^{(m)}}\mathcal{L}_t^{\text{loc}} = \begin{cases} \pi_{m_t}\sum\limits_{m'\neq m_t}\pi_{m'}\sum\limits_{i\in[s]}X_{t,i}\cdot\|w_t^{(m_t)} - w_{t-1}^{(m_t)}\|, & m = m_t \\ -\pi_{m_t}\cdot\pi_m\cdot\sum\limits_{i\in[s]}X_{t,i}\cdot\|w_t^{(m_t)} - w_{t-1}^{(m_t)}\|, & m \neq m_t \end{cases} \tag{23}$$

and

$$\|\nabla_{\theta_t^{(m)}}\mathcal{L}_t^{\text{aux}}\| \leqslant \|\frac{\mu M}{T}f_t^{(m_t)}\frac{\partial\pi_{m_t}(X_t,\Theta_t)}{\partial\theta_t^{(m)}}\| = \mathcal{O}(\mu M), \tag{24}$$

since $\|X_t\| = \mathcal{O}(1)$ and $\pi_m \leqslant 1$, by comparing Eq. (23) and Eq. (24), it can be seen that when $\alpha_m = \mathcal{O}(\ln(\mu M))$, $\mathcal{L}_t^{\text{task}}$ is dominated by $\mathcal{L}_t^{\text{loc}}$; otherwise, it degenerates into being always dominated by $\mathcal{L}_t^{\text{aux}}$. □

## D Proof of Proposition 4.2

We first provide the following Lemma 2.

**Lemma 2.** Let $X$ and $\tilde{X}$ be feature matrices with the feature signals $v_n$ and $v_{n'}$. If $\alpha = \mathcal{O}(\ln(\mu M))$, then for the same expert $m$, the following equation holds with a probability of at least $1 - o(1)$:

$$|h_m(X, \theta_t^{(m)}) - h_m(\tilde{X}, \theta_t^{(m)})| = \mathcal{O}(\sigma_0^2). \tag{25}$$

*Proof.* According to the parameter update formula, it can be obtained that

$$\|\theta_t^{(m)}\| = \|\theta_t^{(m)} - \theta_{t-1}^{(m)} + \cdots + \theta_1^{(m)} - \theta_0^{(m)}\| \tag{26}$$

$$\leqslant \|\theta_t^{(m)} - \theta_{t-1}^{(m)}\| + \cdots + \|\theta_1^{(m)} - \theta_0^{(m)}\| \tag{27}$$

$$\leqslant \eta \sum_t \|\nabla_{\theta_{t-1}^{(m)}} \mathcal{L}_t^{\text{task}}\| = \mathcal{O}(\sigma_0^\alpha) \tag{28}$$

Since

$$P(|\sum_{i=2}^{s}\sum_{j=1}^{d} X_{t,(i,j)}| < s \cdot d \cdot \sigma_0) \geqslant 1 - 2\exp(-\frac{\sigma_0^2 s^2 d^2}{\|X_{t,i}\|}), \tag{29}$$

and $X_{t,(i,j)} \sim N(0, \sigma_t^2)$, so $\|X_{t,i}\| = \mathcal{O}(\sigma_0)$. Thus, we obtain that it holds with a probability exceeding $1 - o(1)$ that

$$|\sum_{i=2}^{s}\sum_{j=1}^{d} X_{t,(i,j)}| = \mathcal{O}(\sigma_0). \tag{30}$$

Based on these results above, we get

$$|h_m(X_t, \theta_t^{(m)}) - h_m(\tilde{X}_t, \theta_t^{(m)})| = |(\theta_t^{(m)})^\top \sum_{i=1}^{s} X_{t,i}| \tag{31}$$

$$\leqslant |(\theta_t^{(m)})^\top \sum_{i=1}^{s} X_{t,i}| + |(\theta_t^{(m)})^\top (v_n - v_{n'})| \tag{32}$$

$$\leqslant |\max_{t,j}\{\theta_{t,j}^{(m)}\}| \cdot |\sum_{i=2}^{s}\sum_{j=1}^{d} X_{t,(i,j)}| + \|v_n - v_{n'}\|| \tag{33}$$

$$\leqslant \mathcal{O}(\sigma_0^2) + \mathcal{O}(\sigma_0^{\alpha+1}) = \mathcal{O}(\sigma_0^2), \tag{34}$$

which completes the proof. This lemma shows that the gating output of each expert is determined by the input feature signal $v_n$ of $X_t$. □

**Proof of Proposition** 4.2.

*Proof.* For $m, m' \in \mathcal{M}_k$, consider

$$\|h_t^{(m)} - h_t^{(m')}\| \sim \|\langle v_t, \theta_t^{(m)} - \theta_t^{(m')}\rangle\| \tag{35}$$

$$\leqslant \|v_t\|\|\theta_t^{(m)} - \theta_t^{(m')}\| \tag{36}$$

$$= \|v_t\|\|\sum_{i=1}^{t}[(\theta_i^{(m)} - \theta_{i-1}^{(m)}) - (\theta_i^{(m')} - \theta_{i-1}^{(m')})]\| \tag{37}$$

$$\leqslant \|v_t\|\sum_{i=1}^{t}\|(\theta_i^{(m)} - \theta_{i-1}^{(m)}) - (\theta_i^{(m')} - \theta_{i-1}^{(m')})\|, \tag{38}$$

where Eq. (35) is due to Lemma 1. Since we have the parameter update formula

$$\theta_t^{(m)} - \theta_{t-1}^{(m)} = \eta \nabla_{\theta_t^{(m)}} \mathcal{L}_t^{\text{task}}, \tag{39}$$

Eq. (38) can be written as

$$\eta\|v_t\|\sum_{i=1}^{t}\|\nabla_{\theta_{i-1}^{(m)}} \mathcal{L}_{i-1}^{\text{task}} - \nabla_{\theta_{i-1}^{(m')}} \mathcal{L}_{i-1}^{\text{task}}\| \tag{40}$$

$$\sim \eta\|v_t\|\|\sum_{i\in t_m^{\text{loc}}} \nabla_{\theta_{i-1}^{(m)}} \mathcal{L}_{i-1}^{\text{loc}} - \sum_{i\in t_{m'}^{\text{loc}}} \nabla_{\theta_{i-1}^{(m')}} \mathcal{L}_{i-1}^{\text{loc}}$$

$$+ \sum_{j\in t_m^{\text{aux}}} \nabla_{\theta_{j-1}^{(m)}} \mathcal{L}_{j-1}^{\text{aux}} - \sum_{j\in t_{m'}^{\text{aux}}} \nabla_{\theta_{j-1}^{(m')}} \mathcal{L}_{j-1}^{\text{aux}}\|. \tag{41}$$

According to the assumption in Section 3.1 and the definition in Lemma 4.1, we obtain

$$\|h_t^{(m)} - h_t^{(m')}\| \tag{42}$$
$$= \mathcal{O}(\sigma_0^{0.5}) \left[\mathcal{O}(|\mathcal{T}_{m'}^{\text{loc}}|\sigma_0^{\alpha_{m'}}) + \mathcal{O}((|\mathcal{T}_m^{\text{aux}}| - |\mathcal{T}_{m'}^{\text{aux}}|)\mu M)\right] \tag{43}$$
$$= \mathcal{O}(|\mathcal{T}_{m'}^{\text{loc}}|\sigma_0^{1.5}) + \mathcal{O}((|\mathcal{T}_m^{\text{aux}}| - |\mathcal{T}_{m'}^{\text{aux}}|)\sigma_0), \tag{44}$$

where the last equation is because the update of expert parameters is expected not to exceed the gap between any two ground truths. □

## E Proof of Lemma 4.3

*Proof.* Consider the n-dimensional softmax function $\pi(x_i) = \frac{e^{x_i}}{\sum_{j=1}^n e^{x_j}}$. It has the property that

(1) $\frac{\partial \pi(x_i)}{\partial x_j} = \pi(x_i) \cdot (\delta_{ij} - \pi(x_j))$, where $\delta_{ij}$ is the Kronecker symbol. (2)$\sum_{j=1}^n \frac{\partial \pi(x_i)}{\partial x_j} = 0$.

Then the composite function $h_i(X_t, \theta_t^{(i)})$ is used to replace $x_i$ in the softmax function, and considering the linearity of the gating network, we obtain

$$\nabla_{\theta_t^{(m)}} \mathcal{L}_t(\Theta_t, \mathcal{D}_t) = \begin{cases} \pi_{m_t} \sum\limits_{m \neq m_t} \pi_{m'} \sum\limits_{i \in [S_t]} X_{t,i} \cdot f, \ m = m_t \\ -\pi_{m_t} \cdot \pi_m \cdot \sum\limits_{i \in [S_t]} X_{t,i} \cdot f, \ m \neq m_t \end{cases} \tag{45}$$

and

$$\sum_{m=1}^{M} \nabla_{\theta_t^{(m)}} \mathcal{L}_t(\Theta_t, \mathcal{D}_t) = 0. \tag{46}$$

Since $\pi_m = \mathcal{O}(1)$, according to Eq. (45) and the parameter update formula $\Delta\theta_t = -\eta \nabla_\theta \mathcal{L}_t$, Eq. (8) is obtained. □

## F Proof of Proposition 4.4

*Proof.* Let $\xi$ denote a training trajectory drawn from the underlying probability space, and write $T_m = T_m(\xi)$ and $T_{m'} = T_{m'}(\xi)$. Before their respective specialization completion, the realized effective update frequencies of $m$ and $m'$ are uniformly of order $p$ and $1-p$, respectively.

Along a fixed trajectory $\xi$, the cumulative effective update of expert $m$ up to $T_m(\xi)$ is therefore of order

$$U_m(T_m) = \Theta(pT_m(\xi)\mu M + (1-p)T_m(\xi)\sigma_0^\alpha)\,. \tag{47}$$

For expert $m'$, since the probability that $\alpha_{m'} = \Omega(\ln(\mu M))$ is $1-p$, its cumulative effective update up to $T_{m'}(\xi)$ is of order

$$U_{m'}(T_{m'}) = \Theta((1-p)T_{m'}(\xi)\mu M + pT_{m'}(\xi)\sigma_0^\alpha)\,. \tag{48}$$

Using

$$\sigma_0^\alpha = o(\mu M),$$

we obtain

$$U_m(T_m) = \Theta(pT_m(\xi)\mu M)\,, \qquad U_{m'}(T_{m'}) = \Theta((1-p)T_{m'}(\xi)\mu M)\,. \tag{49}$$

Since the two experts belong to the same group and, by assumption, require comparable cumulative effective updates to complete specialization, these two quantities are of the same order:

$$\Theta(pT_m(\xi)\mu M) = \Theta((1-p)T_{m'}(\xi)\mu M)\,. \tag{50}$$

Canceling the common factor $\mu M > 0$, we get

$$pT_m(\xi) = \Theta((1-p)T_{m'}(\xi))\,. \tag{51}$$

Equivalently,

$$\frac{T_{m'}(\xi)}{T_m(\xi)} = \Theta\left(\frac{p}{1-p}\right), \tag{52}$$

where the constants hidden in $\Theta(\cdot)$ are uniform over the training trajectory $\xi$.

Therefore, taking expectation over $\xi$ preserves the same scaling:

$$\mathbb{E}\left[\frac{T_{m'}}{T_m}\right] = \int \frac{T_{m'}(\xi)}{T_m(\xi)}\,\mathrm{d}\mathbb{P}(\xi) = \Theta\left(\frac{p}{1-p}\right). \tag{53}$$

This completes the proof. □

## G Proof of Proposition 5.1

*Proof.* Since each column of $X_n$ is generated by a multivariate standard Gaussian distribution, the probability that the columns of $X_n$ are linearly independent is 1. Therefore, for all $n \in [N]$, $Q(X_n) \simeq \mathbb{R}^s$, and this isomorphism is denoted as $\varphi$. Since for all $w \in W_t$, $\|w\| = \mathcal{O}(1)$, we can set $L = \max_{w \in W_t} \|w\|_\infty$. Thus, $\bar{W}_t \subset [-L, L]^s$. For simplicity, we write $[-L, L]^s$ as $L^s$. According to the definition of $g(m_{n_i}) = \mathcal{S}_i(w_n)$, $\varphi(\bar{S}_i(w_n))$ is an open set in $L^s$ (Proposition 5.2 will further show that this is an ellipsoid). Note that $w_{n_i} \in \mathcal{S}_i(w_n)$; by the definition of an open set, there exists an open ball $B(\varphi(\bar{w}_{n_i}), r_{n_i})$ centered at $\varphi(\bar{w}_{n_i})$ with radius $r_{n_i} > 0$ such that

$$B(\varphi(\bar{w}_{n_i}), r_{n_i}) \subset \varphi(\bar{S}_i(w_n)). \tag{54}$$

Since the parameter inheritance strategy ensures the lower bound of each expert's ability, we can set $\inf_{n,i}\{r_{n_i}\} = r_0 > 0$. For any $\bar{w} \in L^s$, construct an open ball $B(\bar{w}, r_0)$ with $\bar{w}$ as the center and $r_0$ as the radius. Denote $\mathcal{U} = \bigcup_{\bar{w} \in L^s} B(\bar{w}, r_0)$. It is obvious that $L^s \subset \varphi(\mathcal{U})$. Note that $L^s$ is a bounded closed set in the Euclidean space, so according to the Heine-Borel theorem, it is also a compact set. Therefore, there exists a finite number $N$ of open balls in $\varphi(\mathcal{U})$ that can cover $L^s$, *i.e*,

$$L^s \subset \bigcup_{n=1}^{N} \varphi(B(\bar{w}_n, r_0)). \tag{55}$$

Let all $\bar{w}_n$ be the ground truth and consider the preimage sets $g^{-1}(\phi^{-1}(\bar{w}_n))$, where $n = 1, \cdots, N$. Take one expert $m_{n_i}$ from each preimage set to obtain the expert sequence $\{m_{n_i}\}_{n=1}^N$, where $i$ can be any index not exceeding $k_n$. Thus, we can denote $n_i$ as $n$. Based on the above analysis, we finally obtain

$$\varphi(\bar{W}_t) \subset L^s \subset \bigcup_{n=1}^{N} B(\varphi(\bar{w}_n), r_0) \subset \bigcup_{n=1}^{N} B(\varphi(\bar{w}_n), r_n) \subset \bigcup_{n=1}^{N} \varphi(\bar{S}_i(w_n)). \tag{56}$$

Due to the image-preserving property of the union, $\bigcup_{n=1}^N \varphi(\bar{S}_i(w_n)) = \varphi(\bigcup_{n=1}^N \bar{S}_i(w_n))$. Moreover, since $\varphi$ is a linear isomorphism, we obtain $W_t \subset \bigcup_{n=1}^N g(m_n)$. □

## H Proof of Proposition 5.2

*Proof.* Let $X'$ be any matrix induced by $X_{n'}^\top$ under $\varphi \circ \phi$ (determined by the chosen basis), and let the images of $w_n^{(m_{n_i})}$ and $y_{n'}$ under this mapping be $w'_n$ and $y'$. It is obvious that $X' \in \mathbb{R}^{s \times s}$. Then $\varphi(\bar{S}_i(w_n)) = \{w' \in \mathbb{R}^s \mid \|X'w'_n - y'\|/\|X'\| < \epsilon_0 \|X'\|\}$. Let $h = w' - w'_n$, then it can be rewritten as

$$\varphi(\bar{S}_i(w_n)) = w'_n + \{h | h^\top X'^\top X' h < \epsilon_0^2 \|X'\|^2\}. \tag{57}$$

Since $\mathrm{Rank}(X) = s$, the set $\varphi(\bar{S}_i(w_n))$ determines an ellipsoid containing $w'_n$. Let $\mu_e$ denote the distribution of all ellipsoids in $\mathbb{R}^s$, and $\mu_w$ denote the conditional distribution of all ellipsoids containing $w'_n$. Since experts in the same cluster are obtained through exactly the same training method, we have $\varphi(\bar{S}_m) \sim \mu_w$. For any $w \in \mathbb{R}^s$, let $p(w) = P(w \in \varphi(\bar{S}_1))$. Thus,

$$\Delta_m = \int_{\mathbb{R}^s} \mathbf{1}\{w \in \varphi(\bar{S}_m), w \notin \bigcup_{i=1}^{m-1} \varphi(\bar{S}_i)\} dw. \tag{58}$$

It is known from Fubini's theorem that

$$\mathbb{E}[\Delta_m] = \mathbb{E}[\int_{\mathbb{R}^s} \mathbf{1}\{w \in \varphi(\bar{S}_m), w \notin \bigcup_{i=1}^{m-1} \varphi(\bar{S}_i)\} dw] \tag{59}$$

$$= \int_{\mathbb{R}^s} P(w \in \varphi(\bar{S}_m), w \notin \bigcup_{i=1}^{m-1} \varphi(\bar{S}_i)\}) dw \tag{60}$$

$$= \int_{\mathbb{R}^s} p(w)(1-p(w))^{m-1} dw, \tag{61}$$

where the last equation relies on exchangeability, which can be demonstrated more rigorously through the dependency coefficient. According to the Dominated Convergence Theorem, $\mathbb{E}[\Delta_m]$ is non-increasing, and

$$\lim_{m \to +\infty} \mathbb{E}[\Delta_m] = 0, \tag{62}$$

which completes the proof. □

# I Proof of Proposition 5.3

We first provide Lemma 3 and Lemma 4.

**Lemma 3.** Assume that a fixed number of task pools and experts are set at $t = 0$, and this number is equal to the number of tasks and experts using the dynamic expansion framework in round $t$. If the freezing strategy in 4 is adopted, then there exists $t_* = \Omega(\eta^{-1}M)$, such that when $t > t_*$, the forgetting rate and the generalization error satisfy

$$\begin{aligned}
\mathbb{E}[F_t] < & \frac{1}{t-1} \sum\nolimits_{\tau=1}^{t-1} \frac{r^{L_t^{(m_\tau)}} - r^{L_\tau^{(m_\tau)}}}{N_t} \sum\nolimits_{n=1}^{N_t} \|w_n\|^2 \\
& + \frac{1}{t-1} \sum\nolimits_{\tau=1}^{t_*} \frac{r^{L_i^{(m_\tau)}} - r^{L_{t_*}^{(m_\tau)}}}{N_t^2} \sum\nolimits_{n \neq n'} \|w_{n'} - w_n\|^2 \\
& + \frac{\psi}{t-1} \sum\nolimits_{\tau=t_*+1}^{t} \left(1 - r^{L_t^{(m_\tau)} - L_\tau^{(m_\tau)}}\right) r^{L_t^{(m_\tau)} - L_{t_*}^{(m_\tau)} - 1},
\end{aligned} \tag{63}$$

where $\psi = \frac{1}{N_t} \sum_{n=1}^{N_t} \sum_{n,n' \in C_k} \frac{\|w_{n'} - w_n\|^2}{|w_k|}$ denotes the average gap between two tasks in a certain task cluster.

$$\begin{aligned}
\mathbb{E}[G_t] < & \frac{1}{t} \sum_{\tau=1}^{t} \frac{r^{L_t^{(m_\tau)}}}{N_t} \sum_{n=1}^{N_t} \|w_n\|^2 + \frac{\psi}{T} \sum_{\tau=t_*+1}^{t} \left(1 - r^{L_t^{(m_\tau)} - L_{t_1}^{(m_\tau)}}\right) \\
& + \frac{\Psi}{t} \left\{ \sum_{\tau=1}^{t_1} \left(1 - r^{L_t^{(m_\tau)} - L_{t_*}^{(m_\tau)}}\right) + \sum_{\tau=t_*+1}^{t} r^{L_t^{(m_\tau)} - L_{t_*}^{(m_\tau)}} \left(1 - r^{L_{t_*}^{(m_\tau)}}\right) \right\} \\
& + \frac{1}{t} \sum_{\tau=1}^{t_*} \frac{r^{L_t^{(m_\tau)} - L_{t_*}^{(m_\tau)}} (1 - r^{L_{t_*}^{(m_\tau)}})}{N^2} \sum_{n \neq n'} \|w_{n'} - w_n\|^2,
\end{aligned} \tag{64}$$

where $\Psi = \frac{1}{N_t} \sum_{n=1}^{N_t} \frac{1}{K_t} \sum_{k=1}^{K_t} \frac{1}{|C_k|} \sum_{n' \in C_k} \|w_{n'} - w_n\|^2$ represents the average difference between a randomly selected task in one cluster and a task in another specific cluster.

*Proof.* Let $P_t = X_t(X_t^\top X_t)^{-1} X_t^\top$, then $P_t$ is an orthogonal projection matrix. Based on the rotational symmetry of orthogonal projection matrices, we obtain

$$E\|P_t x\|^2 = \frac{tr(P_t)}{d} = \frac{s}{d}\|x\|^2 = r\|x\|^2. \tag{65}$$

Since

$$\begin{aligned}
& \|w_t^{m_\tau} - w_{n_i}\|^2 = \|w^{(m_\tau)}_{\tau^{(m_\tau)}(L_t^{(m_\tau)})} - w_{n_i}\|^2 \\
& = \|(I - P_t)(w^{m_\tau}_{\tau^{(m_\tau)}(L_t^{(m_\tau)}-1)} - w_{n_\tau}) + P_t(w_{\tau^{(m_\tau)}(L_t^{(m_\tau)})})\|,
\end{aligned} \tag{66}$$

It can be derived from the above properties that

$$\mathbb{E}[\|w_t^{(m_\tau)} - w_{n_\tau}\|^2] = r^{L_t^{(m_\tau)}}\mathbb{E}[\|w_{n_\tau}\|^2] + \sum_{l=1}^{L_t^{(m_\tau)}}(1-r)r^{L_t^{(m_\tau)}-l}\mathbb{E}[\|w_{\tau^{(m_\tau)}(l)} - w_{n_\tau}\|^2]. \quad (67)$$

According to the uniformity of expert routing given in Lemma 4.3, we have

$$\begin{aligned}
\mathbb{E}[F_t] &= \frac{1}{t-1}\sum_{\tau=1}^{t}\mathbb{E}[\|w_t^{(m_i)} - w_{n_\tau}\|^2 - \|w_\tau^{(m_\tau)} - w_{n_i}\|^2] \quad (68)\\
&= \frac{1}{t-1}\sum_{\tau=1}^{t_*}\{(r^{L_t^{(m_\tau)}})\mathbb{E}[\|w_{n_\tau}\|^2] \\
&\quad + \sum_{l=1}^{L_t^{(m_\tau)}}(1-r)r^{L_t^{(m_\tau)}-l}\mathbb{E}[\|w_{\tau^{(m_\tau)}(l)} - w_{n_\tau}\|]^2 \\
&\quad - \sum_{j=1}^{L_\tau^{(m_\tau)}}(1-r)r^{L_\tau^{(m_\tau)}-j}\mathbb{E}[\|w_{\tau^{(m_\tau)}}(j)\|^2]\} \\
&\quad + \frac{1}{t-1}\sum_{\tau=t_*+1}^{t}\{(r^{L_t^{(m_\tau)}} - r^{L_\tau^{(m_\tau)}})\mathbb{E}[\|w_{n_\tau}\|^2] + \Delta z\}, \quad (69)
\end{aligned}$$

where

$$\begin{aligned}
\Delta z &= \sum_{l=L_{t_*}^{(m_\tau)}+1}^{L_t^{(m_\tau)}}(1-r)r^{L_t^{(m_\tau)}-l}\mathbb{E}[\|w_{\tau^{(m_\tau)}(l)} - w_{n_\tau}\|^2] \\
&\quad - \sum_{j=L_{t_*}^{(m_\tau)}+1}^{L_\tau^{(m_i)}}(1-r)r^{L_\tau^{(m_\tau)}-j}\mathbb{E}[\|w_{\tau^{(m_i)}(j)} - w_{n_i}\|^2] \quad (70)\\
&= \frac{r^{L_t^{(m_\tau)}-L_{t_*}^{(m_\tau)}-1}(1-r^{L_t^{(m_\tau)}-L_\tau^{(m_\tau)}})}{N_t}\sum_{n=1}^{N_t}\sum_{n,n'\in C_k}\frac{\|w_{n'}-w_n\|^2}{|C_k|} \quad (71)
\end{aligned}$$

By organizing the above two formulas, we obtain the conclusion equation. The proof for the generalization error is similar. □

**Lemma 4.** Let $t_\tau$ denote the round when expert $m_\tau$ is frozen. Then, when $t < t_\tau$,

$$\mathbb{E}[\|w_{\tau^{(m_\tau)}(l)} - w_{n_i}\|^2] < \frac{1}{N_t^2}\sum_{n\neq n'}\|w_n - w_{n'}\|^2. \quad (72)$$

and when $t > t_\tau$,

$$\mathbb{E}[\|w_{\tau^{(m_\tau)}(l)} - w_{n_i}\|^2] = \frac{1}{N_t}\sum_{n=1}^{N_t}\sum_{n,n'\in C_k}\frac{\|w_{n'}-w_n\|^2}{|C_k|}. \quad (73)$$

*Proof.* Since each expert will converge to an expert mark $\mathcal{M}_n$ defined in Eq. (5) before $t_\tau$, we obtain

$$\begin{aligned}
\mathbb{E}[\|w_{\tau^{(m_i)}(l)} - w_{n_i}\|^2] &= \mathbb{E}[\frac{1}{N_t}\sum_{n=1}^{N_t}\|w_{\tau^{(m_i)}(l)} - w_n\|^2] \quad (74)\\
&< \frac{1}{N_t}\sum_{n=1}^{N_t}\frac{1}{N_t}\sum_{n'=1}^{N_t}\|w_{n'} - w_n\|^2 \quad (75)\\
&= \frac{1}{N_t^2}\sum_{n\neq n'}\|w_{n'} - w_n\|^2. \quad (76)
\end{aligned}$$

The inequality is because the expected error $\mathbb{E}[\|w_{\tau^{(m_i)}(l)} - w_n\|^2]$ per round is smaller than the uniformly random routing strategy with expected error $\mathbb{E}[\|w_{\tau^{(m_i)}(l)} - w_n\|^2] = \frac{1}{N_t}\sum_{n'=1}^{N_t}\|w_n - w_{n'}\|^2$. When $t > t_\tau$, the router always routes task $w_{\tau^{(m_i)}(l)} = w_{n'}$ within the same cluster to expert $m_i$, which causes the form of Eq. (73). □

**Proof of Proposition** 5.3.

*Proof.* Take any $j \leqslant i$ and consider $t \in [t_{(j)}, t_{(j+1)})$. Let $\mathcal{N}_i$ be the number of newly added experts during the period $[t_{(i)}, t_{(i+1)}]$. The experts in round $t \in [t_{(i)}, t_{(i+1)})$ are arranged in the order of freezing as $\mathcal{N}_1, \cdots, \mathcal{N}_i$. According to the uniformity of expert selection given in Lemma 4.3, during this period, $P(A) = \frac{\sum_{k=1}^{i}\mathcal{N}_k}{\sum_{k=0}^{i}\mathcal{N}_k} = p_j$, then $P(\bar{A}) = \frac{\mathcal{N}_0}{\sum_{k=0}^{i}\mathcal{N}_k} = 1 - p_j$. That is to say, in the rounds where $t \in [t_{(j)}, t_{(j+1)})$, the probabilities that the gating network routes to the already frozen experts and the unfrozen experts are $p_j$ and $1 - p_j$ respectively. Let the local forgetting in this segment of rounds be $F_t^j$, and we can calculate

$$\mathbb{E}[F_t^j] = \mathbb{E}[F_t^j|A]P(A) + \mathbb{E}[F_t^j|\bar{A}]P(\bar{A}). \tag{77}$$

Since we randomly select ground truth from the pool each time, we still have the estimation of the ground truth: $\mathbb{E}[\|w_{n_i}\|^2|A] = \frac{1}{N_t}\sum_{n=1}^{N_t}\|w_n\|^2$. Therefore,

$$\mathbb{E}[F_t^j|A]P(A) = \frac{p_j}{t-1}\sum_{\tau=1}^{t}\mathbb{E}[\|w_t^{(m_i)} - w_{n_\tau}\|^2 - \|w_\tau^{(m_\tau)} - w_{n_i}\|^2|A] \tag{78}$$

$$= \frac{p_j}{t-1}\sum_{\tau=t_{(j)}}^{t_{(j+1)}}\{(r^{L_t^{(m_\tau)}} - r^{L_\tau^{(m_\tau)}})\mathbb{E}[\|w_{n_\tau}\|^2] + \Delta z\} \tag{79}$$

Similarly, we obtain

$$\begin{aligned}\mathbb{E}[F_t^j|\bar{A}]P(\bar{A}) = \frac{1-p_j}{t-1}\Big\{&r^{L_t^{(m_\tau)}}\mathbb{E}[\|w_{n_\tau}\|^2] \\ &+ (1-r)r^{L_\tau^{(m_\tau)}}\mathbb{E}[\|w_{\tau^{(m_\tau)}(l)} - w_{n_\tau}\|^2] \\ &-(1-r)r^{L_\tau^{m_\tau}}\mathbb{E}[\|w_{\tau^{(m_\tau)}}(j)\|^2]\Big\}\end{aligned} \tag{80}$$

According to the proof of Lemma 3, we obtain

$$\mathbb{E}[F_t] = \sum_{j=0}^{i-1}\mathbb{E}[F_t^j] = \sum_{j=0}^{i-1}\{\mathbb{E}[F_t^j|A]P(A) + \mathbb{E}[F_t^j|\bar{A}]P(\bar{A})\} \tag{81}$$

$$\begin{aligned}&= \sum_{j=0}^{i-1}\Bigg\{\frac{p_j}{t-1}\sum_{\tau=t_{(j)}}^{t_{(j+1)}}\{(r^{L_t^{(m_\tau)}} - r^{L_\tau^{(m_\tau)}})\mathbb{E}[\|w_{n_\tau}\|^2] + \Delta z\} \\ &+ \frac{1-p_j}{t-1}\Big\{r^{L_t^{(m_\tau)}}\mathbb{E}[\|w_{n_\tau}\|^2] + (1-r)r^{L_\tau^{(m_\tau)}}\mathbb{E}[\|w_{\tau^{(m_\tau)}(l)} - w_{n_\tau}\|^2] \\ &-(1-r)r^{L_\tau^{m_\tau}}\mathbb{E}[\|w_{\tau^{(m_\tau)}}(j)\|^2]\Big\}\Bigg\}\end{aligned} \tag{82}$$

$$\begin{aligned}&< \frac{1}{t-1}\sum_{j=0}^{i-1}\Big\{p_j\sum_{\tau=t_{(j)}}^{t_{(j+1)}}\frac{r^{L_t^{(m_\tau)}} - r^{L_\tau^{(m_\tau)}}}{N_t}\sum_{n=1}^{N_t}\|w_n\|^2 \\ &+ (1-p_j)\sum_{\tau=t_{(j)}}^{t_{(j+1)}}\frac{r^{L_\tau^{(m_\tau)}} - r^{L_t^{(m_\tau)}}}{N_t^2}\sum_{n=1}^{N_t}\|w_{n'} - w_n\|^2 \\ &+ \psi p_j\sum_{\tau=t_{(j)}}^{t}(1 - r^{L_t^{(m_\tau)} - L_\tau^{(m_\tau)}})r^{L_t^{(m_\tau)} - L_{t_{(j-1)}}^{(m_\tau)} - 1}\Big\}\end{aligned} \tag{83}$$

Finally, we calculate the generalization error. Due to the conclusion of Proposition 5.1, when the number of rounds is sufficiently large, newly arriving tasks will always fall into a certain cluster. Therefore, for sufficiently large $t$ we ignore the impact of $\|w'_{n_\tau} - w_{n_\tau}\| = \mathcal{O}(\sigma_0^{1.5})$. Based on this, we have

$$\mathbb{E}[G_t] = \frac{1}{t}\sum_{\tau=1}^{t}\frac{\|X_\tau^\top w_\tau^{(m_\tau)} - y_\tau\|}{\|X_\tau\|} \leqslant \frac{1}{t}\sum_{\tau=1}^{t}\|w_\tau^{(m_\tau)} - w_{n_\tau}\| \tag{84}$$

$$\leqslant \frac{1}{t}\sum_{\tau=1}^{t}(\|w_\tau^{(m_\tau)} - w'_{n_\tau}\| + \|w'_{n_\tau} - w_{n_\tau}\|) \tag{85}$$

$$\begin{aligned}
&= \frac{1}{t}\sum_{\tau=1}^{t} r^{L_t^{(m_\tau)}}\mathbb{E}[\|w_{n_i}\|^2] \\
&+ \frac{1-p_j}{t}\sum_{j=0}^{i}\sum_{\tau=t_{(j)}}^{t_{(j+1)}}\Bigg\{\sum_{l=1}^{L_{t_\tau}^{(m_\tau)}}(1-r)r^{L_t^{(m_\tau)}-l}\mathbb{E}[\|w_{\tau^{(m_\tau)}(l)} - w_{n_\tau}\|^2] \\
&+ \sum_{l=L_{t_\tau}^{(m_\tau)}+1}^{L_t^{(m_\tau)}}(1-r)r^{L_t^{(m_\tau)}-l}\mathbb{E}[\|w_{\tau^{(m_\tau)}(l)} - w_{n_\tau}\|^2]\Bigg\} \\
&+ \frac{p_j}{t}\sum_{j=0}^{i}\sum_{\tau=t_{(j)}}^{t_{(j+1)}}\Bigg\{\sum_{l=1}^{L_{t_\tau}^{(m_\tau)}}(1-r)r^{L_t^{(m_\tau)}-l}\mathbb{E}[\|w_{\tau^{(m_\tau)}(l)} - w_{n_\tau}\|^2] \\
&+ \sum_{L_{t_\tau}^{(m_\tau)}+1}^{L_t^{(m_\tau)}}(1-r)r^{L_t^{(m_\tau)}-l}\mathbb{E}[\|w_{\tau^{(m_\tau)}(l)} - w_{n_\tau}\|^2]\Bigg\}
\end{aligned} \tag{86}$$

Substituting the conclusion of Lemma 4, we obtain

$$\begin{aligned}
\mathbb{E}[G_t] < \frac{1}{t}\sum_{j=1}^{i}\Bigg\{&\Psi\Big((1-p_j)\sum_{\tau=t_{(j-1)}}^{t_{(j)}}(1-r^{L_t^{(m_\tau)}-L_{t_{(j)}}^{(m_\tau)}}) \\
&+ p_j\sum_{\tau=t_{(j)}}^{t_{(j+1)}} r^{L_t^{(m_\tau)}-L_{t_{(j)}}^{(m_\tau)}}(1-r^{L_{t_{(j)}}^{(m_\tau)}})\Big) \\
&+ \sum_{\tau=t_{(j)}}^{t_{(j+1)}}\frac{r^{L_t^{(m_\tau)}}}{N_t}\sum_{n=1}^{N_t}\|w_n\|^2 + \psi p_j\sum_{\tau=t_{(j)}}^{t}(1-r^{L_t^{(m_\tau)}-L_{t_{(j)}}^{(m_\tau)}}) \\
&+ (1-p_j)\sum_{\tau=t_{(j)}}^{t-1}\frac{r^{L_t^{(m_\tau)}-L_{t_{(j)}}^{(m_\tau)}}(1-r^{L_{t_{(j)}}^{(m_\tau)}})}{N_t^2}\sum_{n=1}^{N_t}\|w_{n'} - w_n\|^2\Bigg\}
\end{aligned} \tag{87}$$

□