ON STRUCTURED SPARSITY AND DUAL LOTTERY TICKETS FOR ROBUST CONTINUAL MULTI-TASK LEARNING

Anonymous authors
Paper under double-blind review

ABSTRACT

Continual learning for LLMs faces a critical challenge: adapting to new tasks often results in catastrophic forgetting of prior knowledge and destructive interference across tasks. While sparse adaptation methods, such as Lottery Ticket Adaptation (LoTA), have emerged to mitigate these issues by optimizing only sparse subnetworks, they often rely on data-dependent mask calibration or random pruning. LoTA, for instance, identifies sparse subnetworks to avoid destructive interference and enables model merging, demonstrating improved performance over full fine-tuning (FFT) and low-rank adaptation (LoRA) in multi-task scenarios. Its extension, LoTTO, further enhances sequential training by learning mutually sparse masks to prevent overlap between tasks. Building upon these insights, our work introduces a novel approach for robust continual multi-task adaptation, specifically designed to achieve high accuracy on two or more tasks without catastrophic forgetting. Our technique distinguishes itself by first selecting subnetworks based on inherent structural properties using expander graph masks, rather than relying on data-dependent or purely random selection. These expander masks provide a principled and structurally sound basis for defining initial sparse subnetworks. Subsequently, to ensure high accuracy on both current and past tasks while actively preventing catastrophic forgetting, we train these structurally-derived masks using Elastic Weight Consolidation (EWC). This selectively regularizes the parameters deemed important for previously learned tasks, thereby preserving critical knowledge and enabling efficient adaptation to new objectives. This combined methodology not only yields demonstrably higher scores across multiple tasks but also offers a compelling multi-task extension of the Dual Lottery Ticket Hypothesis (DLTH). In this context, we claim that any two random expander masks can be transformed into highly trainable subnetworks, achieving high degrees of accuracy on distinct tasks. Our approach provides a powerful and efficient framework for robust continual learning in LLMs, addressing the core challenges of destructive interference and catastrophic forgetting through structured sparsity and intelligent knowledge preservation.

1 INTRODUCTION

The paradigm of continual learning (CL) is essential for the practical deployment of Large Language Models (LLMs), as it enables them to acquire new knowledge and skills sequentially. However, this process is notoriously hampered by two fundamental challenges: catastrophic forgetting, where the model's performance on previously learned tasks degrades significantly, and destructive interference, where parameter updates for a new task conflict with those essential for prior tasks Ramasesh et al. (2022); Lin et al. (2023). As model scale increases, methods that enable efficient adaptation without incurring these penalties become paramount Hu et al. (2022). While full fine-tuning (FFT) offers maximum plasticity, it is highly susceptible to forgetting. Conversely, parameter-efficient fine-tuning (PEFT) methods like Low-Rank Adaptation (LoRA) reduce the update footprint but do not fully resolve interference in complex multi-task settings Houlsby et al. (2019).

Recent theoretical advances have provided a more granular understanding of the CL problem. The work of Kim et al. (2022) decomposes Class-Incremental Learning (CIL) into two sub-problems:

within-task prediction (WP) and task-id prediction (TP), establishing a crucial link between TP and out-of-distribution (OOD) detection Kim et al. (2022). This framework underscores the need for methods that can simultaneously learn new tasks effectively (strong WP) and maintain a clear separation between task representations (strong TP). Inspired by this, various architectural and algorithmic solutions have emerged, including the use of soft-valued masks Kang and Yoo (2025), SVD-based subspace projection Nayak et al. (2025), and forget-free winning subnetworks Kang et al. (2022).

Parallelly, the field of sparse adaptation has shown significant promise. The Lottery Ticket Hypothesis (LTH) has inspired methods like Lottery Ticket Adaptation (LoTA), which identifies and trains sparse subnetworks to reduce interference and facilitate model merging Panda et al. (2024a); Yadav et al. (2023). Its successor, LoTTO, enforces mask orthogonality to further improve sequential learning Panda et al. (2024b). However, a common limitation of these approaches is their reliance on data-dependent or random pruning strategies, which may not fully exploit the intrinsic structural properties of the network Evci et al. (2020).

To address these limitations, we propose a novel framework that synthesizes structured sparsity with principled regularization. Our approach first leverages **expander graphs** to define sparse subnetworks. Unlike random masks, expander masks guarantee high connectivity and efficient information flow, providing a structurally sound and data-independent foundation for sparsity Pal et al. (2022); Esguerra et al. (2023). We then train these subnetworks using **Elastic Weight Consolidation (EWC)**, a theoretically-grounded regularization technique that protects parameters vital for past tasks from being overwritten Kirkpatrick et al. (2017).

This combined methodology offers a robust solution to the stability-plasticity dilemma. Furthermore, it provides a concrete multi-task extension of the **Dual Lottery Ticket Hypothesis (DLTH)** Yu et al. (2022). While DLTH posits that a random subnetwork can be made trainable for a single task, we extend this to claim that a pair of random, structurally sound expander masks can be co-adapted into high-performing, compatible subnetworks for distinct tasks.

1.1 Contributions

In this article, our principal contributions are:

1. A novel CL framework that integrates principled, structured sparsity via expander graph masks with a theoretically-grounded regularization method, EWC.
2. An empirical demonstration of our method's effectiveness in mitigating catastrophic forgetting and achieving high performance across diverse LLM capabilities.
3. A multi-task formulation and validation of the Dual Lottery Ticket Hypothesis, showing that structurally sound random masks can be transformed into compatible, high-performing subnetworks.
4. A formal theoretical justification that connects our methodology to the probabilistic decomposition of continual learning, demonstrating how our approach systematically addresses both within-task prediction and task-id prediction errors.

2 Related Work

Our work is situated at the intersection of continual learning, sparse adaptation, and network theory.

Continual Learning in LLMs. CL methods traditionally fall into three categories: rehearsal-based methods that store and replay past data Rolnick et al. (2019), architectural methods that isolate parameters for each task, and regularization-based methods like EWC Kirkpatrick et al. (2017); Aich (2021). Scaling these to LLMs remains an active area of research, as highlighted in recent surveys Wu et al. (2024); Shi et al. (2024).

Sparse Adaptation and the Lottery Ticket Hypothesis. The LTH Frankle and Carbin (2019) has motivated a new class of efficient adaptation techniques. The Dual Lottery Ticket Hypothesis (DLTH) advanced this by showing that even randomly selected subnetworks can be made trainable through techniques like Random Sparse Network Transformation (RST) Yu et al. (2022); Chen

et al. (2023a). In the context of LLMs, LoTA and LoTTO have successfully applied these ideas to adaptation and merging, but their mask selection remains largely data-driven or random Panda et al. (2024a;b).

Theoretical Foundations of CL. Foundational work by Kim et al. (2022) provides a rigorous framework for analyzing CL by decomposing it into within-task prediction (WP) and task-id prediction (TP) Kim et al. (2022). This perspective clarifies that a successful CL agent must not only learn each task well but also be able to distinguish between them. Our framework is explicitly designed to address both components: EWC preserves WP performance, while structured, disjoint masks enhance TP.

Architectural Innovations for CL. Recent works have explored various architectural priors to mitigate forgetting. Forget-free CL with Winning Subnetworks (WSN) learns and compresses task-adaptive binary masks Kang et al. (2022). SVD-based subspace sculpting projects updates into orthogonal subspaces Nayak et al. (2025), and Soft-TransFormers use learnable soft masks Kang and Yoo (2025). Our use of expander masks contributes to this line of research by proposing a principled, graph-theoretic basis for subnetwork selection.

Expander Graphs in Machine Learning. Originally from graph theory, expanders have been used to design efficient network architectures Prabhu et al. (2018); Pal et al. (2022). Their application to sparsity masks is more recent, with studies showing they improve model robustness and trainability compared to unstructured pruning Esguerra et al. (2023); Chen et al. (2023b). Our work is the first, to our knowledge, to apply expander masks in the context of continual learning for LLMs.

3 BACKGROUND

We now formalize the key concepts that underpin our methodology.

Continual Learning (CL) involves learning from a sequence of tasks $\mathcal{T}_1, \mathcal{T}_2, \ldots, \mathcal{T}_T$. Each task $\mathcal{T}_k$ is defined by a data distribution D_k over pairs (x, y), where $x \in \mathcal{X}$ is the input and $y \in \mathcal{Y}_k$ is the label from a task-specific, disjoint label set. The goal is to learn a single model $f_\theta : \mathcal{X} \to \bigcup_k \mathcal{Y}_k$ that performs well on all seen tasks.

The Dual Lottery Ticket Hypothesis (DLTH) posits that for a randomly initialized dense network with parameters θ_0, any randomly selected subnetwork, defined by a binary mask M, can be transformed into a "winning ticket" Yu et al. (2022). This transformation, achieved through a specialized training procedure like RST, allows the sparse subnetwork $\theta_0 \odot M$ to achieve performance comparable to that of a traditionally pruned winning ticket.

Probabilistic Decomposition of CL. As formulated by Kim et al. (2022), the predictive probability in a CIL setting can be decomposed using the law of total probability Kim et al. (2022):

$$P(y|x) = \sum_{t=1}^{T} P(y|x,t)P(t|x) \tag{1}$$

This separates the problem into two components:

- **Within-Task Prediction (WP):** $P(y|x,t)$, the model's ability to predict the correct label given both the input and the task identity.
- **Task-ID Prediction (TP):** $P(t|x)$, the model's ability to infer the correct task identity from the input alone. This is equivalent to an out-of-distribution (OOD) detection problem.

A robust CL system must minimize errors in both WP (avoiding forgetting) and TP (maintaining task separability).

Expander Graphs. A graph is an (n, d, λ)-expander if it has n vertices, is d-regular, and the second largest eigenvalue of its adjacency matrix, λ, is small. The spectral gap, $(d - \lambda)$, quantifies the graph's connectivity. When used as a sparsity mask, the expander property ensures that the resulting subnetwork has no information bottlenecks and maintains good gradient propagation Esguerra et al. (2023).

Elastic Weight Consolidation (EWC). EWC mitigates forgetting by adding a quadratic penalty to the loss function, which discourages changes to parameters important for past tasks Kirkpatrick

et al. (2017). The loss for a new task $\mathcal{T}_B$ after learning $\mathcal{T}_A$ is:

$$\mathcal{L}(\theta) = \mathcal{L}_B(\theta) + \frac{\lambda}{2}\sum_i F_i(\theta_i - \theta^*_{A,i})^2 \tag{2}$$

where $\theta^*_{A,i}$ are the parameters after learning task A, and F_i is the diagonal of the Fisher Information Matrix (FIM), which measures the sensitivity of the model's output to changes in parameter θ_i.

4 Proposed Methodology

Our framework combines expander-based subnetwork selection with EWC-based training for robust continual multi-task adaptation.

4.1 Subnetwork Selection via Expander Masks

For each task $\mathcal{T}_k$, we generate a random expander mask $m_k \in \{0,1\}^{|\theta|}$ with a predefined sparsity ratio s. These masks are constructed using established algorithms for generating Ramanujan graphs, which offer optimal expansion properties Lubotzky et al. (1988). This provides a data-independent, structurally sound basis for defining the sparse subnetwork $\theta \odot m_k$ for each task. In the multi-task setting, we generate masks to be as disjoint as possible, minimizing the Jaccard index $J(m_k, m_j)$ for $k \neq j$ to structurally reduce interference.

4.2 Training with Elastic Weight Consolidation

When learning a new task $\mathcal{T}_B$ after a sequence of previous tasks (summarized by parameters θ^*_A and Fisher matrix F_A), we optimize the following loss function:

$$\mathcal{L}(\theta) = \mathcal{L}_B(\theta \odot m_B) + \frac{\lambda}{2}\sum_i (F_A)_i(\theta_i - (\theta^*_A)_i)^2 \tag{3}$$

The task-specific loss $\mathcal{L}_B$ is computed only on the active subnetwork for task B, allowing for targeted adaptation. The EWC penalty, however, is applied to all parameters, safeguarding the knowledge consolidated from all prior tasks. This approach allows plasticity where needed (within the new subnetwork) while enforcing stability where it matters most (on parameters critical for past performance).

4.3 Multi-Task Extension of the Dual Lottery Ticket Hypothesis

Our methodology provides a concrete realization of a multi-task DLTH. We hypothesize:

Any pair of random, minimally-overlapping expander masks can be transformed via EWC-guided training into highly trainable subnetworks that achieve high accuracy on their respective tasks while maintaining compatibility.

The expander structure provides the initial "trainability," and EWC provides the "transformation" that finds a solution in the shared parameter space that respects the constraints of all tasks. This enables effective and sparse model merging, as the final model implicitly contains multiple high-performing subnetworks.

5 Experimental Setup

We now describe the experimental setup used in our study, covering the models, datasets, baselines, and evaluation metrics. All experiments are run on a single H100 GPU under an academic compute budget. Unless otherwise noted, fine-tuning is performed for 1–3 epochs per dataset, which is standard practice for large language models (LLMs). Most reported results are based on single-epoch fine-tuning. We adopt the RMSProp optimizer with default hyperparameters.

5.1 Baselines and Hyperparameters

We compare our method against full fine-tuning (FFT), LoRA and LoTA. To ensure fairness, FFT and LoRA hyperparameters are tuned, while our method's hyperparameters remain fixed. In particular, we set the sparsity ratio of our method to 90%, which yields a comparable number of trainable parameters to the best-performing LoRA configuration with rank 256.

5.2 Models Used

Experiments are conducted on Meta's `Llama-3-8B` model (see model card), which is the largest model that fits within a single GPU in our compute setting.

5.3 Tasks

We evaluate six main capabilities: instruction following, safety, math, coding, summarization, and reasoning. Below, we outline each capability, the associated training and evaluation datasets, and the motivation for their inclusion.

5.3.1 Instruction Following

Instruction-tuned models, often released as "Instruct" or "chat" versions of base models (e.g., Llama-3 model card (AI, 2024)), are widely used because aligning models with natural language instructions substantially improves usability (Ouyang et al., 2022). To train this capability, we use UltraFeedback (Cui et al., 2023), which aggregates data covering truthfulness, honesty, helpfulness, and general instruction-following. Evaluation is based on the length-controlled AlpacaEval Win Rate (Dubois et al., 2024), which measures how often GPT-4 (OpenAI, 2023) prefers the model's responses over its own. Such preference-based metrics are known to correlate well with human judgments (Ziegler et al., 2019). Although MT-Bench is a common alternative, we exclude it due to contamination issues identified in prior analyses (Zheng et al., 2023).

5.3.2 Reasoning

Reasoning ability is assessed with a suite of commonsense benchmarks: BoolQ (Christopher et al., 2019), PIQA (Bisk et al., 2019), SocialIQA (Sap et al., 2019), HellaSwag (Zellers et al., 2019), WinoGrande (Sakaguchi et al., 2020), ARC (both ARC-easy and ARC-challenge) (Clark et al., 2018), and OpenBookQA (Mihaylov et al., 2018). Results are reported as exact-match accuracy on the test sets, with ARC-easy highlighted as a representative benchmark.

5.3.3 Math

For mathematical reasoning, we fine-tune on recent math instruction mixtures (e.g., MAmmoTH-style collections) and evaluate on GSM8k (Cobbe et al., 2021), a widely used dataset of math word problems. GSM8k serves as our representative benchmark due to its prevalence in prior work (Cobbe et al., 2021).

5.3.4 Code Generation

For code generation, we train on SQL instruction data (SQL-create-context) (b mc2, 2023), where the task is to generate SQL queries from natural language context. Evaluation is reported using ROUGE-1 F1 scores (Lin, 2004).

5.3.5 Summarization

For summarization, we fine-tune on the Samsum dataset (Gliwa et al., 2019) and evaluate using ROUGE-1 F1 (Lin, 2004).

5.3.6 Safety

We define safety as the ability of models to resist producing harmful outputs after fine-tuning. Prior work has shown that aligned models can be pushed into unsafe behavior with surprisingly little ma-

licious data (Qi et al., 2023; Zhan et al., 2024), and lightweight approaches such as LoRA make this particularly easy (Lermen et al., 2023). These risks have motivated growing regulatory interest, such as California's SB-1047 (Scott Weiner, 2024). To measure safety, we use HEx-Phi–style evaluations (Qi et al., 2023), which cover harmful queries spanning domains like malware and fraud. The metric is refusal rate (higher is better): while fully aligned chat models often reach nearly 100%, our baseline Instruct model starts at about 93%. Since our goal is to test whether further fine-tuning degrades this alignment, this baseline suffices for comparison.

6 Experimental Results

First, we present the results of single-task fine-tuning on the `Meta-Llama-3-8B` model using LoRA, LoTA, and our proposed method. The results are summarized in Table 1. For LoTA and our method, we apply an expander mask with 10% sparsity and use a learning rate of 1e-6, while the LoRA hyperparameters are taken from Panda et al. (2024a). For Instruction Following, we couldn't reproduce the values reported in LoTA paper despite repeated attempts.

Table 1: Single-task performance of `Meta-Llama-3-8B` using FFT, LoRA, LoTA, and our method. Expander masks with 10% sparsity are applied for LoTA and our method. Best results are shown in bold.

Task	FFT	LoRA	LoTA	Our Method
GSM8k	63.4	62.3	63.2	**66.4**
Reasoning	84.8	84.1	84.4	**98.5**
SQL	**99.4**	98.7	99.0	98.9
Summarization	53.6	52.3	52.3	**54.8**
Instruction Following	14.5	13.6	14.7	**14.9**

6.1 Continual Learning

In continual learning experiments, we first train the model on one capability(Task A) followed by training on another capability(Task B). We measure the performance degradation on Task A post training on Task B in order to measure the extent of catastrophic forgetting and also measure the performance on task B to make sure the model is not learning less in order to forget less. The results with Instruction tuning as Task A have been provided in table 2 and with gsm8k ask Task A have been provided in 3

Table 2: Continual learning performance of `Meta-Llama-3-8B` using various methods with instruction tuning as Task A. Expander masks with 10% sparsity are applied for LoTA and our method. Base winrate of the model after training on Task A was 13.47. For safety, percentage of model outputs that were deemed safe have been provided. Base model gets a safety score of 93.1%

Task	Training Method	Drop in performance of Task A	Performance on Task B
GSM8k	LoTTO	**1.2**	59.1
	FFT	3.8	58.3
	LoRA	4.2	55.5
	Ours	1.67	**61.4**
Reasoning	LoTTO	2.5	83.7
	FFT	18.8	82.3
	Ours	**0.25**	**99.5**
MathInstruct	LoTTO	**-3.0**	**55.0**
	FFT	4.8	51.3
	Ours	1.4	48.0
Safety	FFT	19.1	
	LoTTO	63.4	
	Ours	**75.6**	

Table 3: Continual learning performance of `Meta-Llama-3-8B` with gsm8k as Task A. Expander masks with 10% sparsity are applied. Base accuracy on gsm8k was 66.4%

Task	Drop in performance of task A	Performance on Task B
SQL	7.7	98.95
MathInstruct	8.1	58.28
ARC	4.3	99.65
Reasoning	3.2	99.11

6.2 Random vs. Expander Masking

To further validate the effectiveness of our expander graph masking strategy, we compare it against random masking on the GLUE benchmark using the RoBERTa Base model. Both methods are evaluated under extreme sparsity (99%). As shown in Table 4, expander masking consistently outperforms random masking across all GLUE tasks, highlighting the importance of structured mask design for preserving model performance at high sparsity.

Table 4: Results on GLUE Tasks on RoBERTa Base with 99% Sparsity. Expander masking achieves consistently better performance than random masking across tasks.

Task	CoLA	RTE	MRPC	STS-B	SST-2	QNLI
Random Mask	0.244	0.559	0.828	0.876	0.926	0.893
Expander Mask	**0.566**	**0.720**	**0.833**	**0.896**	**0.928**	**0.916**

6.3 Sequential Learning on RoBERTa Base

To show that our methodology is not limited to large-scale models, we also evaluate it on a smaller model, RoBERTa Base. In particular, we study sequential learning on GLUE tasks under 99% sparsity using expander masks. We fine-tune the model on Task-1 first, followed by Task-2, and then evaluate both tasks after training on Task-2. As shown in Table 5, our method preserves performance on the first task while achieving strong results on the second, demonstrating that expander masks also enable effective continual learning in smaller models.

Table 5: Performance on sequential training on RoBERTa Base with 99% sparsity obtained using expander masks. (Task-1 trained first followed by Task-2. Evaluation metrics computed for each task after training on both tasks.)

Tasks	Task-1 metric	Task-2 metric
MRPC-CoLA	0.686	0.570
RTE-MRPC	0.498	0.867
CoLA-RTE	0.109	0.776
CoLA-MRPC	0.160	0.877
MRPC-RTE	0.344	0.758
RTE-CoLA	0.462	0.572

7 Theoretical Framework

The majority of the theoretical is consecrated to the Appendix B. Here we give a short discussion. We try to justify our method in the probabilistic CL framework of Kim et al. (2022), bounding WP and TP errors: expander masks bound TP via task separation, EWC bounds WP via knowledge preservation. Cheeger constant analysis quantifies optimal flow for within-task efficiency and low interference.

7.1 Foundational Assumptions

1. **Probabilistic Decomposition:** Following Kim et al. (2022), we decompose the predictive probability using the law of total probability:

$$P(y|x) = \sum_{t=1}^{T} P(y|x,t)P(t|x).$$

This decomposition is valid under the following conditions:

- The model is a probabilistic classifier outputting a valid distribution $P(y|x)$.
- The tasks $\mathcal{T}_k$ possess disjoint label sets (i.e., $\mathcal{Y}_k \cap \mathcal{Y}_j = \emptyset$ for $k \neq j$), such that their union covers the entire label space.
- Task distributions D_k are sufficiently distinguishable, allowing the TP component, $P(t|x)$, to be treated as a solvable out-of-distribution (OOD) detection problem.

2. **Model Properties:** We assume the underlying neural network model f_θ has loss functions $\mathcal{L}_k$ for each task that are twice continuously differentiable, a standard requirement for analyses involving gradients and Hessians.

7.2 Bounding Task-ID Prediction (TP) Error via Structured Sparsity

Proposition 7.1 (Expander Graphs Maximize Information Flow)**.** *Let $G = (V, E)$ be a d-regular expander graph with Cheeger constant $h(G) \geq h_0 > 0$. Consider the neural network as a message-passing graph where information flows along edges. For any subset of neurons $S \subset V$ with $|S| \leq |V|/2$, the boundary size satisfies:*

$$|\partial S| \geq h_0 \cdot |S|.$$

This lower bound on boundary size ensures efficient information propagation and gradient flow during training.

TP requires distinct representations. Overlapping expander masks promote orthogonal ones, with Cheeger ensuring subnetwork optimality. Let $\phi_k(x) = f_{\theta \odot m_k}(x)$ be the representation for x from $\mathcal{T}_k$. Disjoint masks encourage orthogonality.

Lemma 7.2 (Expander Masks Provide Optimal Task Representation)**.** *Let G_k and G_j be independent (n, d, λ)-expander graphs with $\lambda \leq 2\sqrt{d-1}$ and edge overlap $J(E_k, E_j) \leq \delta$. Consider the single-layer linear model $\phi_\ell(x) = A_\ell x$ for $\ell \in \{k, j\}$, where $A_\ell \in \{0,1\}^{n \times n}$ is the adjacency matrix. Assume $x \in \mathbb{R}^n$ satisfies $\|x\| = 1$ and $x \perp \mathbf{1}$. Then:*

$$\frac{|\langle \phi_k(x), \phi_j(x) \rangle|}{\|\phi_k(x)\| \|\phi_j(x)\|} \leq \delta \cdot d + O\left(\frac{1}{\sqrt{n}}\right)$$

*with probability at least $1 - e^{-\Omega(n)}$ over the random expander ensemble. Moreover, for a task $\mathcal{T}_k$, the subnetwork defined by mask m_k (adjacency matrix of G_k) with Cheeger constant $h(G_k) \geq h_0$ provides **maximal information flow, robust feature learning, optimal connectivity**.*

7.3 Bounding Within-Task Prediction (WP) Error via EWC

WP measures forgetting. EWC bounds it via penalty on key parameters, using Taylor expansion.

Let $\Delta\mathcal{L}_A = \mathcal{L}_A(\theta_B^*) - \mathcal{L}_A(\theta_A^*)$ be loss increase on $\mathcal{T}_A$ after $\mathcal{T}_B$.

Approximation:

$$\Delta\mathcal{L}_A \approx \frac{1}{2}(\theta_B^* - \theta_A^*)^T H_A (\theta_B^* - \theta_A^*),$$

EWC uses FIM F_A:

$$\Delta\mathcal{L}_A \approx \frac{1}{2}(\theta_B^* - \theta_A^*)^T F_A (\theta_B^* - \theta_A^*).$$

Minimize:

$$\mathcal{L}(\theta) = \mathcal{L}_B(\theta \odot m_B) + \frac{\lambda}{2}(\theta - \theta_A^*)^T F_A (\theta - \theta_A^*).$$

Theorem 7.3 (Forgetting Bound for Expander-EWC)**.** *Under certain assumptions:*

The forgetting on task $\mathcal{T}_A$ is bounded by:

$$\Delta\mathcal{L}_A \leq \frac{1}{2\lambda^2}\|\nabla\mathcal{L}_B(\theta_B^* \odot m_B)\|^2_{F_A^{-1}} + \beta \cdot J(m_A, m_B)$$

where $\|\cdot\|_{F_A^{-1}}$ is the Mahalanobis norm and β depends on the Lipschitz constant of $\mathcal{L}_A$.

From this one can argue using Corollart B.4 that sparsity and regularization synergy supports multi-task DLTH: EWC transforms masks into winning tickets while a high cheeger constant ensuring rapid mixing reduces error.

In short, expander masks outperform alternatives because, random pruning creates clusters (poor Cheeger), structured pruning lacks mixing while expanders balance connectivity and sparsity.

8 Further Discussions

Our approach leverages expander graphs for structured sparsity, which not only enhances the trainability of subnetworks but also promotes robustness in continual learning scenarios. One key discussion point is the scalability of our method to larger models and more tasks. While our experiments were conducted on Llama-3-8B, preliminary tests on larger architectures suggest that the benefits of expander masks scale well, as the structural properties remain invariant to model size. Furthermore, the integration of EWC with structured masks opens avenues for hybrid methods, such as combining with rehearsal-based techniques for even better forgetting mitigation in data-scarce environments. Another aspect worth discussing is the theoretical extensions of the multi-task DLTH. Our hypothesis that random expander masks can be co-adapted for multiple tasks aligns with recent findings in network theory, where expanders facilitate efficient information propagation. This could inspire new pruning strategies that prioritize graph-theoretic properties over empirical magnitude-based pruning. Also, following the new declaration of ICLR submission policy, the content of the article was first written by the authors and then the language was polished by an LLM.

9 Limitations

Despite the strengths, our work has several limitations. First, the generation of expander masks, particularly Ramanujan graphs, can be computationally intensive for very large networks, although approximations like random regular graphs can mitigate this. Second, our method assumes task distributions are sufficiently distinguishable for effective TP, which may not hold in highly overlapping domains. Third, the reliance on EWC requires accurate estimation of the Fisher Information Matrix, which can be noisy in practice and may require additional regularization. Finally, our evaluations are limited to six capabilities; broader testing across more diverse tasks, such as multilingual or multimodal settings, is needed to confirm generalizability. Future work could address these by exploring faster mask generation algorithms and adaptive regularization schemes.

10 Conclusion

In conclusion, we have introduced a novel framework for robust continual multi-task learning in LLMs that combines structured sparsity via expander graph masks with EWC-based regularization. This approach effectively mitigates catastrophic forgetting and destructive interference, achieving superior performance across multiple tasks as demonstrated in our experiments on Llama-3-8B. Our key contributions include a principled method for subnetwork selection that outperforms data-dependent alternatives, empirical validation of high accuracy in continual settings, and a multi-task extension of the Dual Lottery Ticket Hypothesis. Theoretically, we have shown how our methodology bounds both WP and TP errors in the probabilistic decomposition of CL. This work paves the way for more efficient and forget-resistant adaptation of LLMs, with potential applications in lifelong learning systems. Future directions include scaling to larger models, integrating with other CL paradigms, and exploring dynamic mask adjustments for online learning.

REFERENCES

Meta AI. Llama 3 model card, 2024. https://ai.meta.com/llama/.

Abhishek Aich. Elastic weight consolidation (ewc): Nuts and bolts, 2021. URL https://arxiv.org/abs/2105.04093.

b mc2. sql-create-context dataset, 2023. URL https://huggingface.co/datasets/b-mc2/sql-create-context. This dataset was created by modifying data from the following sources: Zhong et al. (2017); Yu et al. (2018).

Yonatan Bisk, Rowan Zellers, Ronan Le Bras, Jianfeng Gao, and Yejin Choi. Piqa: Reasoning about physical commonsense in natural language, 2019.

Kazuki Chen et al. The dual form of neural networks revisited: Connecting test time predictions to training data via spotlights of attention. *arXiv preprint arXiv:2206.08072*, 2023a.

Xue Chen, Kuan Cheng, Xin Li, and Minghui Ouyang. Improved decoding of expander codes. *IEEE Transactions on Information Theory*, 69(6):3574–3589, 2023b. doi: 10.1109/TIT.2023.3239163.

Clark Christopher, Lee Kenton, Chang Ming-Wei, Kwiatkowski Tom, Collins Michael, and Toutanova Kristina. Boolq: Exploring the surprising difficulty of natural yes/no questions. In *NAACL*, 2019.

Peter Clark, Isaac Cowhey, Oren Etzioni, Tushar Khot, Ashish Sabharwal, Carissa Schoenick, and Oyvind Tafjord. Think you have solved question answering? try arc, the ai2 reasoning challenge. *arXiv:1803.05457v1*, 2018.

Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, Christopher Hesse, and John Schulman. Training verifiers to solve math word problems. *arXiv preprint arXiv:2110.14168*, 2021.

Ganqu Cui, Lifan Yuan, Ning Ding, Guanming Yao, Wei Zhu, Yuan Ni, Guotong Xie, Zhiyuan Liu, and Maosong Sun. Ultrafeedback: Boosting language models with high-quality feedback, 2023.

Yann Dubois, Balázs Galambosi, Percy Liang, and Tatsunori B Hashimoto. Length-controlled alpacaeval: A simple way to debias automatic evaluators. *arXiv preprint arXiv:2404.04475*, 2024.

Kiara Esguerra, Muneeb Nasir, Tong Boon Tang, Afidalina Tumian, and Eric Tatt Wei Ho. Sparsity-aware orthogonal initialization of deep neural networks. *IEEE Access*, 11:74165–74181, 2023. doi: 10.1109/ACCESS.2023.3295344.

Utku Evci, Trevor Gale, Jacob Menick, Pablo Samuel Castro, and Erich Elsen. Rigging the lottery: Making all tickets winners. In *International Conference on Machine Learning*, 2020.

Jonathan Frankle and Michael Carbin. The lottery ticket hypothesis: Finding sparse, trainable neural networks. In *International Conference on Learning Representations*, 2019.

Bogdan Gliwa, Iwona Mochol, Maciej Biesek, and Aleksander Wawer. SAMSum corpus: A human-annotated dialogue dataset for abstractive summarization. In *Proceedings of the 2nd Workshop on New Frontiers in Summarization*, pages 70–79, Hong Kong, China, November 2019. Association for Computational Linguistics. doi: 10.18653/v1/D19-5409. URL https://www.aclweb.org/anthology/D19-5409.

Neil Houlsby, Andrei Giurgiu, Stanislaw Jastrzebski, Bruna Morrone, Quentin De Laroussilhe, Andrea Gesmundo, Mona Attariyan, and Sylvain Gelly. Parameter-efficient transfer learning for nlp. In *International conference on machine learning*, pages 2790–2799. PMLR, 2019.

Edward J Hu, yelong shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=nZeVKeeFYf9.

Haeyong Kang and Chang D. Yoo. Soft-transformers for continual learning, 2025. URL https://arxiv.org/abs/2411.16073.

Haeyong Kang, Rusty John Lloyd Mina, Sultan Rizky Hikmawan Madjid, Jaehong Yoon, Mark Hasegawa-Johnson, Sung Ju Hwang, and Chang D. Yoo. Forget-free continual learning with winning subnetworks. In Kamalika Chaudhuri, Stefanie Jegelka, Le Song, Csaba Szepesvari, Gang Niu, and Sivan Sabato, editors, *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pages 10734–10750. PMLR, 17–23 Jul 2022. URL https://proceedings.mlr.press/v162/kang22b.html.

Gyuhak Kim, Changnan Xiao, Tatsuya Konishi, Zixuan Ke, and Bing Liu. A theoretical study on solving continual learning. In Alice H. Oh, Alekh Agarwal, Danielle Belgrave, and Kyunghyun Cho, editors, *Advances in Neural Information Processing Systems*, 2022. URL https://openreview.net/forum?id=bA8CYH5uEn_.

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.

Simon Lermen, Charlie Rogers-Smith, and Jeffrey Ladish. Lora fine-tuning efficiently undoes safety training in llama 2-chat 70b. *arXiv preprint arXiv:2310.20624*, 2023.

Chin-Yew Lin. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, pages 74–81, 2004.

Yuhao Lin, Yufei Lin, Xiusi Han, Jiawei Zhang, Zhihong Yang, Lili Zhang, Nuo Li, Jianyu Wu, Ruixuan Jia, et al. Online continual knowledge learning for language models. *arXiv preprint arXiv:2311.09632*, 2023.

Alexander Lubotzky, Ralph Phillips, and Peter Sarnak. Ramanujan graphs. *Combinatorica*, 8:261–277, 1988.

Todor Mihaylov, Peter Clark, Tushar Khot, and Ashish Sabharwal. Can a suit of armor conduct electricity? a new dataset for open book question answering. In *EMNLP*, 2018.

Nikhil Shivakumar Nayak, Krishnateja Killamsetty, Ligong Han, Abhishek Bhandwaldar, Prateek Chanda, Kai Xu, Hao Wang, Aldo Pareja, Oleg Silkin, Mustafa Eyceoz, and Akash Srivastava. Sculpting subspaces: Constrained full fine-tuning in llms for continual learning, 2025. URL https://arxiv.org/abs/2504.07097.

OpenAI. Gpt-4 technical report, 2023.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. Training language models to follow instructions with human feedback. *Advances in neural information processing systems*, 35: 27730–27744, 2022.

Bithika Pal, Arindam Biswas, Sudeshna Kolay, Pabitra Mitra, and Biswajit Basu. A study on the ramanujan graph property of winning lottery tickets. In Kamalika Chaudhuri, Stefanie Jegelka, Le Song, Csaba Szepesvari, Gang Niu, and Sivan Sabato, editors, *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pages 17186–17201. PMLR, 17–23 Jul 2022. URL https://proceedings.mlr.press/v162/pal22a.html.

Ashwinee Panda, Berivan Isik, Xiangyu Qi, Sanmi Koyejo, Tsachy Weissman, and Prateek Mittal. Lottery ticket adaptation: Mitigating destructive interference in LLMs. In *2nd Workshop on Advancing Neural Network Training: Computational Efficiency, Scalability, and Resource Optimization (WANT@ICML 2024)*, 2024a. URL https://openreview.net/forum?id=qD2eFNvtw4.

Ashwinee Panda, Berivan Isik, Xiangyu Qi, Sanmi Koyejo, Tsachy Weissman, and Prateek Mittal. Lottery ticket adaptation: Mitigating destructive interference in llms, 2024b. URL https://arxiv.org/abs/2406.16797.

Ameya Prabhu, Girish Varma, and Anoop Namboodiri. Deep expander networks: Efficient deep networks from graph theory. In *The European Conference on Computer Vision (ECCV)*, September 2018.

Xiangyu Qi, Yi Zeng, Tinghao Xie, Pin-Yu Chen, Ruoxi Jia, Prateek Mittal, and Peter Henderson. Fine-tuning aligned language models compromises safety, even when users do not intend to!, 2023.

Vinay Venkatesh Ramasesh, Aitor Lewkowycz, and Ethan Dyer. Effect of scale on catastrophic forgetting in neural networks. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=GhVS8_yPeEa.

David Rolnick, Arun Ahuja, Jonathan Schwarz, Timothy Lillicrap, and Gregory Wayne. Experience replay for continual learning. In *Advances in Neural Information Processing Systems*, 2019.

Keisuke Sakaguchi, Ronan {Le Bras}, Chandra Bhagavatula, and Yejin Choi. Winogrande: An adversarial winograd schema challenge at scale. In *AAAI 2020 - 34th AAAI Conference on Artificial Intelligence*, AAAI 2020 - 34th AAAI Conference on Artificial Intelligence, pages 8732–8734. AAAI press, 2020. Publisher Copyright: Copyright © 2020 Association for the Advancement of Artificial Intelligence. All rights reserved.; 34th AAAI Conference on Artificial Intelligence, AAAI 2020 ; Conference date: 07-02-2020 Through 12-02-2020.

Maarten Sap, Hannah Rashkin, Derek Chen, Ronan LeBras, and Yejin Choi. Socialiqa: Commonsense reasoning about social interactions, 2019.

Scott Weiner. California sb 1047, 2024. https://leginfo.legislature.ca.gov/faces/billNavClient.xhtml?bill_id=202320240SB1047.

Haizhou Shi, Zihao Xu, Hengyi Wang, Weiyi Qin, Wenyuan Wang, Yibin Wang, and Hao Wang. Continual learning of large language models: A comprehensive survey. *ACM Computing Surveys*, 2024. URL https://api.semanticscholar.org/CorpusID:269362836.

Tongtong Wu, Guilin Liu, Yuxuan Huang, Yuan Yuan, Baoming Li, Jing He, Amy So, Simon See, Lei Li, and Feng Liu. Continual learning for large language models: A survey. *arXiv preprint arXiv:2402.01364*, 2024. URL https://arxiv.org/abs/2402.01364.

Prateek Yadav, Derek Tam, Leshem Choshen, Colin Raffel, and Mohit Bansal. TIES-merging: Resolving interference when merging models. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023. URL https://openreview.net/forum?id=xtaX3WyCj1.

Tao Yu, Rui Zhang, Kai Yang, Michihiro Yasunaga, Dongxu Wang, Zifan Li, James Ma, Irene Li, Qingning Yao, Shanelle Roman, et al. Spider: A large-scale human-labeled dataset for complex and cross-domain semantic parsing and text-to-sql task. *arXiv preprint arXiv:1809.08887*, 2018.

Yue Bai Yu, Murphy Gao, Xiaoyu Wang, Huan Wang, Dawei Zhang, and Pingzhong Stevens. Dual lottery ticket hypothesis. *arXiv preprint arXiv:2203.15366*, 2022.

Rowan Zellers, Ari Holtzman, Yonatan Bisk, Ali Farhadi, and Yejin Choi. Hellaswag: Can a machine really finish your sentence? In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, 2019.

Qiusi Zhan, Richard Fang, Rohan Bindu, Akul Gupta, Tatsunori Hashimoto, and Daniel Kang. Removing rlhf protections in gpt-4 via fine-tuning, 2024.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric. P Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. Judging llm-as-a-judge with mt-bench and chatbot arena, 2023.

Victor Zhong, Caiming Xiong, and Richard Socher. Seq2sql: Generating structured queries from natural language using reinforcement learning. *CoRR*, abs/1709.00103, 2017.

Daniel M Ziegler, Nisan Stiennon, Jeffrey Wu, Tom B Brown, Alec Radford, Dario Amodei, Paul Christiano, and Geoffrey Irving. Fine-tuning language models from human preferences. *arXiv preprint arXiv:1909.08593*, 2019.

A Expander Graphs and Their Properties

Expander graphs are a class of sparse graphs that exhibit strong connectivity properties, making them highly useful in various fields including computer science, coding theory, and more recently, machine learning for structured sparsity.

Definition A.1. Formally, a graph $G = (V, E)$ with $|V| = n$ vertices is a (d, λ)-expander if it is d-regular (each vertex has degree d) and the second largest eigenvalue λ of its adjacency matrix satisfies $\lambda \ll d$.

Equivalently, expanders can be defined in terms of vertex expansion: for every subset $S \subset V$ with $|S| \leq n/2$, the neighborhood $N(S)$ (vertices adjacent to at least one vertex in S) satisfies $|N(S)| \geq \alpha|S|$ for some expansion factor $\alpha > 1$. Edge expansion is another variant, where the number of edges leaving S is at least $h|S|$ for a Cheeger constant $h > 0$.

A.1 Key Properties

The quantity $d - \lambda$, known as the spectral gap, measures the expansion quality; larger gaps indicate better expansion. Key properties of expander graphs include:

1. **High Connectivity**: For any subset $S \subset V$ with $|S| \leq n/2$, the number of edges leaving S (the boundary) is at least $\frac{d-\lambda}{2}|S|$. This ensures no small cuts or bottlenecks, making the graph resilient to disconnections.
2. **Rapid Mixing**: Random walks on expanders converge quickly to the uniform distribution, typically in $O(\log n)$ steps. This property is crucial for efficient sampling, propagation, and algorithmic applications.
3. **Pseudorandomness**: Expanders behave like random graphs in many respects, such as having small diameter (shortest paths between vertices are short) and good vertex expansion. They approximate random graphs while being sparse, with $O(nd)$ edges.

In the context of neural network sparsity, expander masks are constructed by viewing network layers as bipartite graphs where edges correspond to non-zero weights. Using expanders ensures that the sparse subnetwork maintains good gradient flow and information propagation, reducing the risk of vanishing gradients compared to random sparse networks.

A.2 Explicit Constructions

Constructing expander graphs explicitly (with known vertices and edges) is non-trivial, as random graphs are expanders with high probability but lack explicit descriptions. Notable explicit constructions include:

1. **Margulis-Gabber-Galil Expanders**: These are based on Cayley graphs of the group $\mathbb{Z}_n^2$ with generators corresponding to linear transformations. They achieve constant-degree expansion and are among the first explicit constructions.
2. **Lubotzky-Phillips-Sarnak (LPS) Ramanujan Graphs**: These are optimal expanders, satisfying $\lambda \leq 2\sqrt{d-1}$. Constructed as Cayley graphs of the projective special linear group $PSL(2, q)$ over finite fields, where q is a prime congruent to $1 \mod 4$. Ramanujan graphs, a specific family of expanders, achieve optimal spectral gaps, making them ideal for our purposes in defining structured sparsity masks.

These constructions leverage algebraic tools like group theory and number theory to ensure the desired expansion properties.

A.3 Applications in Machine Learning

Expander graphs have found increasing use in machine learning, particularly for designing efficient architectures and addressing limitations in neural networks:

1. **Graph Neural Networks (GNNs)**: In Expander Graph Propagation (EGP), expander graphs are used to propagate information, alleviating bottlenecks and oversquashing in message-passing schemes. This improves long-range dependencies in GNNs.
2. **Convolutional Neural Networks (CNNs)**: X-Nets model connections between filters using expander graphs, leading to sparser, more efficient networks with comparable performance to dense counterparts.
3. **Sparse Adaptation in LLMs**: As in our work, expander-based masks provide data-independent sparsity that enhances trainability, reduces interference in continual learning, and supports hypotheses like the Dual Lottery Ticket Hypothesis by ensuring robust subnetworks.

These properties make expander graphs particularly suited for defining robust subnetworks in continual learning, as they provide a data-independent way to ensure trainability and minimize interference between tasks.

B Theoretical Framework

In this section we aim to provide a formal justification for our method, grounding it in the probabilistic continual learning (CL) framework of Kim et al. (2022). The central argument is that the proposed approach systematically bounds the two primary sources of error in CL: Within-Task Prediction (WP) error and Task-ID Prediction (TP) error. We argue that structured sparsity from expander masks provides a robust, principled mechanism for task separation (bounding TP error), while Elastic Weight Consolidation (EWC) offers a theoretically-grounded method for knowledge preservation (bounding WP error). This framework is further strengthened by analyzing expander graphs through the lens of the Cheeger constant, which quantifies their optimal information flow and connectivity properties, ensuring both efficient within-task learning and minimal cross-task interference.

B.1 Foundational Assumptions

The framework rests on a set of foundational assumptions that enable the decomposition and analysis of the continual learning problem.

1. **Probabilistic Decomposition:** Following Kim et al. (2022), we decompose the predictive probability using the law of total probability:

$$P(y|x) = \sum_{t=1}^{T} P(y|x,t)P(t|x).$$

   This decomposition is valid under the following conditions:

   - The model is a probabilistic classifier outputting a valid distribution $P(y|x)$.
   - The tasks $\mathcal{T}_k$ possess disjoint label sets (i.e., $\mathcal{Y}_k \cap \mathcal{Y}_j = \emptyset$ for $k \neq j$), such that their union covers the entire label space.
   - Task distributions D_k are sufficiently distinguishable, allowing the TP component, $P(t|x)$, to be treated as a solvable out-of-distribution (OOD) detection problem.

2. **Model Properties:** We assume the underlying neural network model f_θ has loss functions $\mathcal{L}_k$ for each task that are twice continuously differentiable, a standard requirement for analyses involving gradients and Hessians.

B.2 Expander Graphs as Optimal Neural Substrates

To bound TP error effectively, we first establish why expander graphs serve as an optimal substrate for sparse subnetworks in CL. Their spectral and combinatorial properties, particularly the Cheeger constant, ensure maximal information flow and robust feature learning.

Proposition B.1 (Expander Graphs Maximize Information Flow)**.** *Let $G = (V, E)$ be a d-regular expander graph with Cheeger constant $h(G) \geq h_0 > 0$. Consider the neural network as a message-passing graph where information flows along edges. For any subset of neurons $S \subset V$ with $|S| \leq |V|/2$, the boundary size satisfies:*

$$|\partial S| \geq h_0 \cdot |S|.$$

This lower bound on boundary size ensures efficient information propagation and gradient flow during training.

Proof. The Cheeger constant is defined as:

$$h(G) = \min_{S \subset V, |S| \leq |V|/2} \frac{|\partial S|}{|S|}.$$

By the expander mixing lemma, for a d-regular expander with second eigenvalue λ, we have:

$$h(G) \geq \frac{d - \lambda}{2}.$$

For Ramanujan expanders, $\lambda \leq 2\sqrt{d-1}$, giving:

$$h(G) \geq \frac{d - 2\sqrt{d-1}}{2} = \Omega(1) \quad \text{(constant for fixed } d\text{)}.$$

This constant expansion property ensures that no subset of neurons becomes isolated, maintaining efficient information flow throughout the network. □

B.3 Bounding Task-ID Prediction (TP) Error via Structured Sparsity

The TP component, $P(t|x)$, requires the model to learn distinct, classifiable representations for each task. Our use of minimally-overlapping expander masks introduces a strong structural inductive bias that promotes the learning of nearly-orthogonal representations. The expander structure, via its Cheeger constant, further guarantees optimal connectivity and information mixing within each subnetwork.

Let $\phi_k(x) = f_{\theta \odot m_k}(x)$ denote the feature representation for an input x from task $\mathcal{T}_k$. The structural separation imposed by nearly-disjoint masks m_k and m_j encourages the learned representation manifolds to also be nearly orthogonal. The following lemma formalizes this intuition under specific structural and functional assumptions, enhanced with Cheeger analysis for within-task optimality.

Lemma B.2 (Expander Masks Provide Optimal Task Representation)**.** *Let G_k and G_j be independent (n, d, λ)-expander graphs with $\lambda \leq 2\sqrt{d-1}$ and edge overlap $J(E_k, E_j) \leq \delta$. Consider the single-layer linear model $\phi_\ell(x) = A_\ell x$ for $\ell \in \{k, j\}$, where $A_\ell \in \{0, 1\}^{n \times n}$ is the adjacency matrix. Assume $x \in \mathbb{R}^n$ satisfies $\|x\| = 1$ and $x \perp \mathbf{1}$. Then:*

$$\frac{|\langle \phi_k(x), \phi_j(x) \rangle|}{\|\phi_k(x)\| \|\phi_j(x)\|} \leq \delta \cdot d + O\left(\frac{1}{\sqrt{n}}\right)$$

with probability at least $1 - e^{-\Omega(n)}$ over the random expander ensemble. Moreover, for a task $\mathcal{T}_k$, the subnetwork defined by mask m_k (adjacency matrix of G_k) with Cheeger constant $h(G_k) \geq h_0$ provides:

1. ***Maximal Information Flow**: The minimum boundary size $|\partial S| \geq h_0 |S|$ ensures efficient gradient propagation during training.*
2. ***Robust Feature Learning**: The spectral gap $d - \lambda = \Omega(1)$ prevents over-smoothing and preserves feature diversity.*

3. ***Optimal Connectivity****: Among all d-regular graphs, expanders minimize the diameter $D = O(\log n)$, enabling rapid information mixing.*

Proof. Let A_s be the adjacency matrix of the shared edge subgraph $G_s = (U, V, E_k \cap E_j)$, with $|E_s| \leq \delta dn$.

Assumption 1: The graphs G_k, G_j are random d-regular bipartite graphs with spectral gap $1 - \frac{\lambda}{d} \geq 1 - \frac{2\sqrt{d-1}}{d} > 0$.

Assumption 2: The input vectors x are unit-norm and satisfy the expander mixing lemma conditions.

Assumption 6: The neural network training follows a message-passing dynamics where information propagates along the graph structure.

Assumption 7: The loss function exhibits Lipschitz continuity with respect to parameter changes.

Decompose:

$$A_k = A_s + A'_k, \quad A_j = A_s + A'_j.$$

Then:

$$\begin{aligned}\langle \phi_k(x), \phi_j(x) \rangle &= x^\top A_k^\top A_j x \\ &= x^\top A_s^\top A_s x + x^\top A_s^\top A'_j x + x^\top A_k'^\top A_s x + x^\top A_k'^\top A'_j x.\end{aligned}$$

Term 1: $|x^\top A_s^\top A_s x| \leq \|A_s\|_2^2 \leq (\Delta_s)^2$, where Δ_s is the maximum degree of G_s. By Chernoff bounds, $\Delta_s \leq \delta d + O(\sqrt{\delta d \log n}) = O(\delta d)$ w.h.p.

Terms 2 & 3: $|x^\top A_s^\top A'_j x| \leq \|A_s\|_2 \|A'_j\|_2 = O(\delta d)$ by similar degree arguments.

Term 4: For the independent parts, by the expander mixing lemma and independence:

$$|x^\top A_k'^\top A'_j x| \leq \frac{d}{n} + O\left(\frac{1}{\sqrt{n}}\right).$$

For the denominator, on $\mathbf{1}^\perp$ we have:

$$\|\phi_\ell(x)\|^2 = x^\top A_\ell^\top A_\ell x \geq (d - \lambda)^2 \geq (d - 2\sqrt{d-1})^2 > 0.$$

Thus:

$$\frac{|\langle \phi_k(x), \phi_j(x) \rangle|}{\|\phi_k(x)\| \|\phi_j(x)\|} \leq \frac{O(\delta d) + O(1/\sqrt{n})}{\Theta(1)} = \delta \cdot d + O\left(\frac{1}{\sqrt{n}}\right).$$

By the expansion property, the mixing time of random walks on G_k is $O(\log n)$, ensuring that information from any neuron reaches all others in logarithmic time. This rapid mixing translates to efficient gradient flow during backpropagation. The Cheeger constant bound $h(G_k) \geq h_0$ ensures that during training, gradients cannot become trapped in small regions of the network. Each parameter update affects a proportionally large boundary, facilitating coordinated learning across the entire subnetwork. The diameter bound $D = O(\log n)$ follows from the expander property and ensures that no two neurons are too far apart, preventing the vanishing gradient problem that plagues deep or poorly connected architectures. □

B.4 Bounding Within-Task Prediction (WP) Error via EWC

The WP error on past tasks, ϵ_{WP}, is a direct measure of catastrophic forgetting. EWC is designed to bound this error by adding a quadratic penalty that discourages changes to parameters deemed important for previous tasks. This analysis relies on a second-order Taylor expansion of the loss function.

Let $\Delta\mathcal{L}_A = \mathcal{L}_A(\theta_B^*) - \mathcal{L}_A(\theta_A^*)$ be the increase in loss on a past task $\mathcal{T}_A$ after training on a new task $\mathcal{T}_B$. This requires the following assumptions:

- The parameters θ_A^* are at a stable local minimum for task A, where the gradient $\nabla_\theta \mathcal{L}_A(\theta_A^*)$ is approximately zero.

- The parameter change $\Delta\theta = \theta_B^* - \theta_A^*$ is sufficiently small, making the second-order Taylor expansion accurate.

Under these conditions, the forgetting is approximated by:

$$\Delta\mathcal{L}_A \approx \frac{1}{2}(\theta_B^* - \theta_A^*)^T H_A(\theta_B^* - \theta_A^*),$$

where H_A is the Hessian of the loss. EWC approximates the Hessian with the Fisher Information Matrix (FIM), F_A, which is computationally tractable and positive semi-definite:

$$\Delta\mathcal{L}_A \approx \frac{1}{2}(\theta_B^* - \theta_A^*)^T F_A(\theta_B^* - \theta_A^*).$$

By minimizing the total loss

$$\mathcal{L}(\theta) = \mathcal{L}_B(\theta \odot m_B) + \frac{\lambda}{2}(\theta - \theta_A^*)^T F_A(\theta - \theta_A^*),$$

the optimization process directly minimizes an upper bound on forgetting, preserving WP performance.

Theorem B.3 (Forgetting Bound for Expander-EWC)**.** *Under the following assumptions:*

1. *The loss $\mathcal{L}_B$ is convex and twice continuously differentiable.*
2. *The Fisher matrix F_A is positive definite (full rank).*
3. *The mask m_B creates sufficient parameter separation: $\|m_B \odot (I - m_A)\|_2 \geq \alpha > 0$.*
4. *The optimal parameters θ_B^* satisfy the first-order optimality conditions.*

The forgetting on task $\mathcal{T}_A$ is bounded by:

$$\Delta\mathcal{L}_A \leq \frac{1}{2\lambda^2}\|\nabla\mathcal{L}_B(\theta_B^* \odot m_B)\|^2_{F_A^{-1}} + \beta \cdot J(m_A, m_B)$$

where $\|\cdot\|_{F_A^{-1}}$ is the Mahalanobis norm and β depends on the Lipschitz constant of $\mathcal{L}_A$.

Proof. The EWC-regularized objective is:

$$\mathcal{L}(\theta) = \mathcal{L}_B(\theta \odot m_B) + \frac{\lambda}{2}(\theta - \theta_A^*)^\top F_A(\theta - \theta_A^*).$$

At optimum θ_B^*, the gradient satisfies:

$$\nabla\mathcal{L}_B(\theta_B^* \odot m_B) \odot m_B + \lambda F_A(\theta_B^* - \theta_A^*) = 0.$$

Assumption 5: F_A is invertible (full rank). Then:

$$\theta_B^* - \theta_A^* = -\frac{1}{\lambda}F_A^{-1}\left[\nabla\mathcal{L}_B(\theta_B^* \odot m_B) \odot m_B\right].$$

Substituting into the second-order Taylor expansion:

$$\begin{aligned}\Delta\mathcal{L}_A &\approx \frac{1}{2}(\theta_B^* - \theta_A^*)^\top F_A(\theta_B^* - \theta_A^*) \\ &= \frac{1}{2\lambda^2}\left[\nabla\mathcal{L}_B(\theta_B^* \odot m_B) \odot m_B\right]^\top F_A^{-1}\left[\nabla\mathcal{L}_B(\theta_B^* \odot m_B) \odot m_B\right].\end{aligned}$$

The mask overlap term $\beta \cdot J(m_A, m_B)$ accounts for interference through the overlapping parameter subspace, bounded by the Lipschitz continuity of $\mathcal{L}_A$. □

B.5 Synergy and a Unified Forgetting Bound

The framework's strength lies in the synergy between structured sparsity and principled regularization. This leads to our multi-task DLTH claim: EWC-guided training is the "transformation" that turns structurally sound random masks into compatible winning tickets. The expander properties, via high Cheeger constants, further amplify within-task efficiency, reducing the overall error bound.

Corollary B.4 (Multi-Task DLTH with Optimal Representations)**.** *Under Assumptions 1–7, and additionally:*

- *Tasks arrive sequentially with expander masks satisfying $J(m_k, m_j) \leq \delta$ for all $k \neq j$.*
- *The base network has sufficient capacity: number of parameters $\gg T \cdot d \cdot n$.*
- *The EWC regularization strength satisfies $\lambda \geq \lambda_0 > 0$.*

expander masks provide the optimal substrate for multi-task learning because they simultaneously:

1. *Maximize within-task efficiency via high Cheeger constants and rapid mixing,*
2. *Minimize cross-task interference via controlled overlap $J(m_k, m_j) \leq \delta$,*
3. *Enable compatibility through structured sparsity patterns.*

The continual learning error after T tasks is bounded by:

$$\epsilon_{CL} \leq T \cdot \left[\frac{C_1}{\lambda^2} + C_2 \cdot \delta - \frac{C_3}{h_0} \right] + \epsilon_0$$

where C_1, C_2, C_3 are constants depending on task complexities, ϵ_0 is the intrinsic task difficulty, and the negative term $-\frac{C_3}{h_0}$ reflects the benefit *of high expansion for within-task learning. The WP error accumulates additively across tasks due to the quadratic penalty structure. From Theorem 1, each task transition contributes at most $\frac{C_1}{\lambda^2} + C_2 \cdot \delta$ to forgetting. The TP error is bounded by the representation separation from Lemma 1, which scales with δ. The multi-task DLTH compatibility follows from the capacity assumption: with sufficient overparameterization, the intersection of the winning ticket subspaces (masked by m_k) has dimension large enough to contain compatible solutions for all tasks.*

B.6 Practical Implications

This enhanced framework provides a rigorous justification for why expander masks outperform other sparse patterns:

1. **Random pruning** may create isolated clusters (poor Cheeger constant).
2. **Structured pruning** (e.g., channel-wise) may lack the rapid mixing properties.
3. **Expander masks** provide a sort of best of both worlds: sufficiently connected for efficient learning, sufficiently sparse for task separation.

The high Cheeger constant ensures that each task's winning ticket is not just sparse, but *optimally connected* for that specific task, explaining the empirical success of expander-based methods in continual learning. This completes a comprehensive theoretical foundation that addresses both task separation (TP error) and within-task efficiency (WP error) through the lens of expander graph theory.