On the Impact of Expert Count in Mixture of Experts

Dang Hoang Duy[1], **Tak Shing Au Yeung**[2], **Juntao Yang**[2], **Ka Chun Cheung**[2], **Simon See**[2]

[1] National University of Singapore
[2] NVIDIA AI Technology Center

Abstract

Mixture-of-Experts (MoE) layers have achieved notable success across various deep learning applications. However, the impact of the number of experts on MoE performance across different task settings remains poorly understood. In this work, we investigate the impact of expert quantity within the MoE architecture composed of multilayer perceptron (MLP) experts. Concretely, we develop a formal MoE model with MLP experts, derive a range of expert counts that optimizes performance and load balance, and validate it on synthetic data. By systematically varying the number of experts, we demonstrate that balancing specialization and effective expert routing is key to maximizing performance.

Introduction

Mixture-of-Experts (MoE) architectures partition a network into many specialized *experts* and use a learned *gating* or *routing* function to select only a small subset of experts for each input. This conditional computation raises effective capacity while keeping per-example computational cost nearly constant, which is particularly valuable for transformer-based large language models (LLMs) (Kaplan et al. 2020; Hoffmann et al. 2022).

MoE is originally introduced for adaptive and hierarchical mixtures (Jacobs et al. 1991; Jordan and Jacobs 1994). Early developments broadened the choice of expert families and learning settings. Examples include MoE built on support vector machines (Collobert, Bengio, and Bengio 2001), Gaussian processes (Rasmussen and Ghahramani 2001), and nonlinear Bayesian formulations (Shahbaba and Neal 2009). With the rise of deep learning, researchers embedded stacked routers and experts into deep networks to handle complex perception tasks (Eigen, Ranzato, and Sutskever 2013) and explored connections to generative modeling and distributed learning (Theis and Bethge 2015; Deisenroth and Ng 2015). Practical adoption accelerated with sparse top-k routing that activates only a few experts per token, enabling large-capacity layers at manageable compute (Shazeer et al. 2017). Functional specialization across experts has also been studied in continual and task-incremental settings (Aljundi, Chakravarty, and Tuytelaars 2017).

These ideas now underpin many state-of-the-art LLM systems. Models such as MIXTRAL-8X7B (Jiang et al. 2024), GROK-1 (xAI 2024), DBRX (Databricks 2024), ARCTIC (S.A.R.Team 2024), and DEEPSEEK-V2 (Liu et al. 2024) expand capacity without a proportional increase in serving cost by activating only selected experts for each token.

Despite great empirical success and growing popularity of MoE across various domains, its theoretical understanding remains scarce. Prior work has shown that MoE encourages experts to specialize in different functions and data regions (Chen et al. 2022). However, it is unclear whether this behavior changes under different expert counts. Furthermore, the work considers MoE architecture with convolutional neural networks (CNNs) as experts, while in reality, in transformer-based LLMs, expert networks are usually feed-forward networks.

We address these gaps by analyzing how expert count affects performance and learning dynamics when experts are MLPs, which aligns the theory with the feed-forward structure used in transformer architectures. Our contributions are as follows: (i) we present a formal MoE model with MLP experts that theoretically aligns closer to the architecture of large language models. (ii) We give a theoretical analysis that identifies a range of expert counts where model accuracy and routing balance are optimal. (iii) Experiments on synthetic datasets show that empirical results align with our analysis. These insights aim to guide MoE scaling choices in modern transformer systems.

Related Work

Mixture of Experts. The mixture-of-experts (MoE) paradigm has a long history in machine learning (Jacobs et al. 1991; Jordan and Jacobs 1994). Early MoE realizations employed diverse base learners, including support vector machines (Collobert, Bengio, and Bengio 2001), Gaussian processes (Tresp 2000), and hidden Markov models (Jordan, Ghahramani, and Saul 1996). To expand capacity for vision and speech tasks, Eigen, Ranzato, and Sutskever (2013) embedded MoE within deep neural networks by stacking routers and experts. Subsequent work introduced sparse gating that activates only a small subset of experts per example, improving training stability while lowering computation (Shazeer et al. 2017). Building on these ideas, MoE layers combined with different neural backbones achieved strong

results in language modeling (Shazeer et al. 2017; Dauphin et al. 2017; Vaswani et al. 2017). More recently, routing each input to a single expert rather than to K experts was shown to further reduce routing cost while preserving model quality (Fedus, Zoph, and Shazeer 2022).

Theory of MoE Models. Most theory on MoE models has focused on simple MoE settings that study approximation and statistical guarantees. For hierarchical MoE, Jiang and Tanner (1999a) analyzes convergence when experts are exponential family regressors, and Jiang and Tanner (1999b) establishes rates for hierarchical MoE with generalized linear experts. From a function-approximation view, Zeevi, Meir, and Maiorov (2002) derives error bounds in Sobolev spaces, while Nguyen, Lloyd-Jones, and McLachlan (2016) gives a universal approximation theorem for softmax gating with linear experts without assuming target differentiability. For polynomial experts under softmax gating, Mendes and Jiang (2012) characterize how the MLE convergence rate scales with the number of experts and the polynomial order.

Preliminaries

Data distribution

Following Chen et al. (2022), we study binary classification on inputs composed of P patches, each living in $\mathbb{R}^d$. A labeled example is $(\mathbf{x}, y)$, where $\mathbf{x} = (\mathbf{x}^{(1)}, \ldots, \mathbf{x}^{(P)}) \in (\mathbb{R}^d)^P$ and $y \in \{\pm 1\}$. The data are organized into K clusters. Each cluster $k \in [K]$ is associated with a unit label signal vector $\mathbf{v}_k$ and a unit cluster-center vector $\mathbf{c}_k$. For simplicity, we assume orthogonality among all signals: $\langle \mathbf{v}_k, \mathbf{v}_{k'} \rangle = \langle \mathbf{c}_k, \mathbf{c}_{k'} \rangle = \langle \mathbf{v}_k, \mathbf{c}_{k'} \rangle = 0$ for all $k \neq k'$.

Definition 1. *An example $(\mathbf{x}, y)$ is drawn from a distribution $\mathcal{D}$ by the following procedure:*

1. *Sample a distinct pair (k, k') uniformly from $[K] \times [K]$ with $k \neq k'$.*
2. *Sample the label $y \in \{\pm 1\}$ uniformly and an independent Rademacher variable $\varepsilon \in \{\pm 1\}$.*
3. *Draw independent amplitudes α, β, γ from $\mathcal{D}_\alpha, \mathcal{D}_\beta, \mathcal{D}_\gamma$. We assume there exist absolute constants $C_1, C_2 > 0$ such that almost surely $C_1 \leq \alpha, \beta, \gamma \leq C_2$.*
4. *Construct $\mathbf{x} = (\mathbf{x}^{(1)}, \ldots, \mathbf{x}^{(P)})$ so that exactly three patches carry structured signal and the remaining $P - 3$ patches are i.i.d. Gaussian noise:*
   - ***Feature signal:*** *one patch equals $y\,\alpha\,\mathbf{v}_k$.*
   - ***Cluster-center signal:*** *one patch equals $\beta\,\mathbf{c}_k$.*
   - ***Feature noise:*** *one patch equals $\varepsilon\,\gamma\,\mathbf{v}_{k'}$.*
   - ***Random noise:*** *each of the remaining $P - 3$ patches is drawn i.i.d. from $N\big(0, (\sigma_p^2/d)\,\mathbf{I}_d\big)$, where σ_p is an absolute constant.*

We divide examples into K clusters $\cup_{k \in [K]} \Omega_k$ based on the cluster-center signal, that is, an example $(\mathbf{x}, y) \in \Omega_k$ if and only if a patch of $\mathbf{x}$ aligns with $\mathbf{c}_k$.

New Architecture for MoE with MLP Experts

We study a mixture-of-experts (MoE) layer with M expert networks $f_1, \ldots, f_M$ and a linear gating network (Shazeer et al. 2017; Fedus, Zoph, and Shazeer 2022). Our design removes weight sharing across both experts and gate, and makes each expert patch-aware.

For the m-th expert, with input $\mathbf{x} = (\mathbf{x}^{(1)}, \ldots, \mathbf{x}^{(P)})$, we use a patch-aware MLP

$$f_m(\mathbf{x}; \mathbf{W}) = \sum_{j \in [J]} \sum_{p \in [P]} \sigma\Big(\big\langle \mathbf{w}_{m,j,p},\, \mathbf{x}^{(p)} \big\rangle\Big),$$

where J is the number of hidden neurons per patch, and $\mathbf{w}_{m,j,p} \in \mathbb{R}^d$ is the weight vector of unit j for patch p in expert m. For each expert m, patch p has its own parameter block $\mathbf{W}_{m,p} = \{\mathbf{w}_{m,j,p} : j \in [J]\} \in \mathbb{R}^{d \times J}$.

The gate produces scores $\mathbf{h}(\mathbf{x}; \mathbf{\Theta}) = (h_1, \ldots, h_M) \in \mathbb{R}^M$ via a linear, patch-aggregating map

$$\mathbf{h}(\mathbf{x}; \mathbf{\Theta}) \;=\; \sum_{p \in [P]} \mathbf{\Theta}_p^\top \mathbf{x}^{(p)}, \qquad \mathbf{\Theta}_p \in \mathbb{R}^{d \times M}.$$

We convert scores to routing probabilities with a softmax,

$$\pi_m(\mathbf{x}; \mathbf{\Theta}) = \frac{\exp\big(h_m(\mathbf{x}; \mathbf{\Theta})\big)}{\sum_{m'=1}^{M} \exp\big(h_{m'}(\mathbf{x}; \mathbf{\Theta})\big)}, \quad m \in [M].$$

Given a selection set $\mathcal{T}_\mathbf{x} \subseteq [M]$, the MoE layer output is

$$F(\mathbf{x}; \mathbf{\Theta}, \mathbf{W}) = \sum_{m \in \mathcal{T}_\mathbf{x}} \pi_m(\mathbf{x}; \mathbf{\Theta})\, f_m(\mathbf{x}; \mathbf{W}).$$

Top-1 routing model. Soft routing with $\mathcal{T}_\mathbf{x} = [M]$ (Jordan and Jacobs 1994) is often costly in deep settings. Following the switch strategy of Fedus, Zoph, and Shazeer (2022), we adopt sparse top-1 routing: for each $\mathbf{x}$, activate a single expert,

$$\mathcal{T}_\mathbf{x} \;=\; \arg\max_{m \in [M]} \big\{\, h_m(\mathbf{x}; \mathbf{\Theta}) \,\big\}, \qquad |\mathcal{T}_\mathbf{x}| = 1. \quad (1)$$

Equal weight initialization across patches. We initialize expert weights so that, within each expert and hidden neuron, all patches share the same initial vector, while different neurons remain independent. Concretely, for every (m, j) sample $\tilde{\mathbf{w}}_{m,j} \sim \mathcal{N}(\mathbf{0}, \sigma_0^2 \mathbf{I}_d)$ i.i.d., and set

$$\mathbf{w}_{m,j,p}^{(0)} \;=\; \tilde{\mathbf{w}}_{m,j} \quad \text{for all } p \in [P]. \quad (2)$$

This symmetric initialization across patches is a necessary condition to enable expert specialization guarantees, as shown in Lemma 2 and Lemma 3.

Training Algorithm

Given the training set $S = \{(\mathbf{x}_i, y_i)\}_{i=1}^n$, we train F via gradient descent to minimize the empirical loss

$$\mathcal{L}(\mathbf{\Theta}, \mathbf{W}) = \frac{1}{n} \sum_{i=1}^{n} \ell\big(y_i F(\mathbf{x}_i; \mathbf{\Theta}, \mathbf{W})\big), \quad (3)$$

where ℓ is the logistic loss defined as $\ell(z) = \log(1 + \exp(-z))$. We initialize the router parameters at $\mathbf{\Theta}^{(0)} = \mathbf{0}$, which follows common practice in MoE training. As discussed by Shazeer et al. (2017), a zero-initialized router helps avoid out-of-memory issues and promotes an approximately uniform initial expert load.

Algorithm 1: Gradient descent with random initialization

Require: Number of iterations T, expert learning rate η, router learning rate η_r, initialization scale σ_0, training set $S = \{(x_i, y_i)\}_{i=1}^n$

1: Generate each entry of $\mathbf{W}^{(0)}$ as in (2).
2: Initialize each entry of $\mathbf{\Theta}^{(0)}$ to zero.
3: **for** $t = 0, 1, \ldots, T-1$ **do**
4: Generate each entry of $\mathbf{r}^{(t)}$ independently from $\mathrm{Unif}[0, 1]$.
5: Update $\mathbf{W}^{(t+1)}$ as in (5)
6: Update $\mathbf{\Theta}^{(t+1)}$ as in (6).
7: **end for**
8: **return** $(\mathbf{\Theta}^{(T)}, \mathbf{W}^{(T)})$

Rather than routing with the raw gating scores at iteration t, we add a small uniform perturbation to encourage exploration across experts and to stabilize training. Concretely, example $\mathbf{x}_i$ is routed to $\arg\max_m\{h_m(\mathbf{x}_i; \mathbf{\Theta}^{(t)}) + r_{m,i}^{(t)}\}$, where the noises $\{r_{m,i}^{(t)}\}_{m\in[M], i\in[n]}$ are drawn i.i.d. from $\mathrm{Unif}[0, 1]$ and collected as $\mathbf{r}^{(t)}$. This uniform perturbation is a standard technique for sparsely gated MoE layers (Shazeer et al. 2017; Fedus, Zoph, and Shazeer 2022). Intuitively, the noise smooths the induced routing probabilities, which both mitigates early collapse and helps distribute gradients more evenly during the initial training phase. The following lemma formalizes this smoothing effect.

Lemma 1. *Let $\mathbf{h}, \widehat{\mathbf{h}} \in \mathbb{R}^M$ be the output of the gating network and $\{r_m\}_{m=1}^M$ be i.i.d. noise from $\mathrm{Unif}[0, 1]$. Denote $\mathbf{p}, \widehat{\mathbf{p}} \in \mathbb{R}^M$ as the routing probabilities, i.e., $p_m = \mathbb{P}\big(\arg\max_{m'\in[M]}\{h_{m'} + r_{m'}\} = m\big)$ and $\widehat{p}_m = \mathbb{P}\Big(\arg\max_{m'\in[M]}\{\widehat{h}_{m'} + r_{m'}\} = m\Big)$. Then $\|\mathbf{p} - \widehat{\mathbf{p}}\|_\infty \le M^2 \|\mathbf{h} - \widehat{\mathbf{h}}\|_\infty$.*

At iteration t, the corresponding perturbed objective becomes

$$\mathcal{L}^{(t)}(\mathbf{\Theta}^{(t)}, \mathbf{W}^{(t)}) = \frac{1}{n}\sum_{i=1}^n \ell\big(y_i \pi_{m_{i,t}}(\mathbf{x}_i; \mathbf{\Theta}^{(t)}) f_{m_{i,t}}(\mathbf{x}_i; \mathbf{W}^{(t)})\big), \tag{4}$$

where $m_{i,t} = \arg\max_m\{h_m(\mathbf{x}_i; \mathbf{\Theta}^{(t)}) + r_{m,i}^{(t)}\}$. The factor $\pi_{m_{i,t}}$ captures the router's contribution to the logit, while $f_{m_{i,t}}$ is the selected expert's prediction on $\mathbf{x}_i$.

We then apply separate updates to experts and router. The expert update uses a normalized gradient step on the perturbed loss:

$$\mathbf{W}_m^{(t+1)} = \mathbf{W}_m^{(t)} - \eta \cdot \nabla_{\mathbf{W}_m}\mathcal{L}^{(t)}(\mathbf{\Theta}^{(t)}, \mathbf{W}^{(t)}) / \|\nabla_{\mathbf{W}_m}\mathcal{L}^{(t)}(\mathbf{\Theta}^{(t)}, \mathbf{W}^{(t)})\|_F, \tag{5}$$

where $\eta > 0$ is the expert learning rate.

The router is updated with a standard gradient step on the same perturbed objective:

$$\theta_m^{(t+1)} = \theta_m^{(t)} - \eta_r \cdot \nabla_{\theta_m}\mathcal{L}^{(t)}(\mathbf{\Theta}^{(t)}, \mathbf{W}^{(t)}),\ \forall m \in [M], \tag{6}$$

where $\eta_r > 0$ is the router learning rate. The full training loop is summarized in Algorithm 1.

Main Results

We study how the number of experts influences model behavior by decomposing training into two phases: an exploration stage and a router learning stage. The exploration stage begins at iteration $t = 0$ and ends at $t = T_1 = \lfloor \eta^{-1}\sigma_0^{1/2} \rfloor$. In this phase, we analyze how the expert count affects expert specialization and coverage. Concretely, we characterize whether individual experts specialize to distinct clusters and whether all clusters are covered by at least one expert by the end of the stage. The router learning stage starts at $t = T_1$ and continues until $t = \lfloor \eta^{-1}M^{-2} \rfloor$. In this phase, we show that if experts are already well specialized at the beginning of the stage, then the router converges to accurate routing by its end.

Under suitable initialization of the model weights, our analyses of these two stages yield a learning guarantee for the MoE model with explicit dependence on the number of experts.

Equal Weights Initialization Across Patches Guarantees Specialization

Recall the initialization of each entry of $\mathbf{W}^{(0)}$ in (2). Since the ordering of patches within an input is uniformly random, initializing the per-patch weights of an expert to the same value avoids introducing spurious asymmetries among inputs that belong to the same cluster. Prior analyses (Chen et al. 2022) also indicate that an expert's eventual specialization arises from its initial weights. The next lemma formalizes this intuition.

Lemma 2. *Suppose the weights $\mathbf{W}^{(0)}$ are initialized as in (2). Then, at the end of the exploration stage, with probability at least $1 - o(1)$, the following equations hold for all experts $m \in \mathcal{M}_k$ and any $k, k' \in [K]$ with $k' \neq k$,*

$$\mathbb{P}_{(\mathbf{x},y)\sim\mathcal{D}}\left(y f_m(\mathbf{x}; \mathbf{W}^{(T_1)}) \le 0 \mid (\mathbf{x}, y) \in \Omega_k\right) = o(1),$$
$$\mathbb{P}_{(\mathbf{x},y)\sim\mathcal{D}}\left(y f_m(\mathbf{x}; \mathbf{W}^{(T_1)}) \le 0 \mid (\mathbf{x}, y) \in \Omega_{k'}\right) = \Omega(1/K)$$

In contrast, if an expert's weights are initialized independently across patches, the expert fails to specialize to any single cluster, as shown below.

Lemma 3. *Suppose the weights $\mathbf{W}^{(0)}$ are initialized independently from $N(0, \sigma_0^2)$. Then, at the end of the exploration stage, with probability at least $1 - o(1)$, the following equation holds for all experts $m \in [M]$ and clusters $k \in [K]$,*

$$\mathbb{P}_{(\mathbf{x},y)\sim\mathcal{D}}\left(y f_m(\mathbf{x}; \mathbf{W}^{(T_1)}) \le 0 \mid (\mathbf{x}, y) \in \Omega_k\right) = \Omega(1/K)$$

Taken together, these results imply that equal weights initialization across patches is both necessary and sufficient to guarantee expert specialization. Detailed proofs are deferred to Appendix **Generalization Results**.

Expert Count-dependent Learning Guarantees

Prior analyses do not explicitly quantify how learning guarantees scale with the expert count M. Here we address this gap by providing an explicit range of values of M for which analogous guarantees hold. In particular, when M satisfies $\Theta(K \log K \log\log d) \leq M \leq \Theta(\sigma_0^{-0.075})$, there exists a choice of hyperparameters J, η, η_r such that running Algorithm 1 on the MoE layer in Section **New Architecture for MoE with MLP Experts** achieves the results stated below.

Theorem 1. *Suppose the training data size is $n = \Omega(d)$. Choose experts number $M \in [\Theta(K \log K \log\log d), \Theta(\sigma_0^{-0.075})]$, filter size $J = \Theta(\log M \log\log d)$, initialization scale $\sigma_0 \in [d^{-1/3}, d^{-0.01}]$, learning rate $\eta = \widetilde{O}(\sigma_0)$, $\eta_r = \Theta(M^2)\eta$. Then with probability at least $1-o(1)$, Algorithm 1 is able to output $(\Theta^{(T)}, \mathbf{W}^{(T)})$ within $T = \widetilde{O}(\eta^{-1})$ iterations such that the non-linear MoE defined in Section **New Architecture for MoE with MLP Experts** satisfies*

- *Training error is zero, i.e., $y_i F(x_i; \Theta^{(T)}, \mathbf{W}^{(T)}) > 0,\ \forall i \in [n]$.*
- *Test error is nearly zero, i.e., $\mathbb{P}_{(x,y)\sim\mathcal{D}}\big(yF(x; \Theta^{(T)}, \mathbf{W}^{(T)}) \leq 0\big) = o(1)$.*

More importantly, the experts can be divided into a disjoint union of K non-empty sets $[M] = \bigsqcup_{k\in[K]} \mathcal{M}_k$ and

- *(Each expert is good on one cluster) Each expert $m \in \mathcal{M}_k$ performs well on the cluster Ω_k, $\mathbb{P}_{(x,y)\sim\mathcal{D}}\big(yf_m(x; \mathbf{W}^{(T)}) \leq 0 \,\big|\, (x,y) \in \Omega_k\big) = o(1)$.*
- *(Router only distributes example to good expert) With probability at least $1 - o(1)$, an example $x \in \Omega_k$ will be routed to one of the experts in $\mathcal{M}_k$.*

The lower bound on M ensures adequate coverage of the K clusters, while the upper bound guarantees that experts specialize into distinct clusters *before* the router begins to learn. From a practical perspective, this result provides principled guidance for selecting the number of experts in cluster-structured classification tasks. When the number of underlying clusters K is known, or can be reasonably estimated from domain knowledge or unsupervised analysis, the theorem yields a constrained range of feasible expert counts. This significantly narrows the hyperparameter search space compared to conventional approaches, which leads to faster model selection and reduces computational cost.

Numerical Experiments

Experiment Setup We generate synthetic data following Definition 1 with $K = 4$ clusters, feature dimension $d = 50$, and number of patches $P \in \{4, 8\}$. Unless otherwise noted, scalar parameters are drawn independently as $\alpha \sim \mathrm{Uniform}(0.5, 2)$, $\beta \sim \mathrm{Uniform}(1, 2)$, and $\gamma \sim \mathrm{Uniform}(0.5, 3)$. We set $\sigma_p = 1$.

For evaluation, we report two metrics. Model accuracy measures predictive performance. To assess routing quality, we use dispatch entropy. Let $n_{k,m}$ denote the number of samples from cluster $k \in \{1, \ldots, K\}$ that are routed to expert $m \in \{1, \ldots, M\}$. Define $n_m = \sum_{k=1}^{K} n_{k,m}$ as the load of expert m and $n = \sum_{k=1}^{K}\sum_{m=1}^{M} n_{k,m}$ as the total number of samples. The dispatch entropy is the load-weighted average of the cluster mixture entropy at each expert:

$$\text{entropy} = -\sum_{\substack{m=1 \\ n_m \neq 0}}^{M} \frac{n_m}{n} \sum_{k=1}^{K} \frac{n_{k,m}}{n_m} \log\left(\frac{n_{k,m}}{n_m}\right). \quad (7)$$

This quantity equals zero when every active expert receives samples from at most one cluster. It is maximized when the router dispatches uniformly at random.

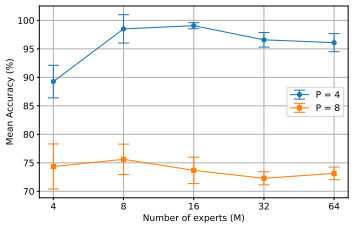
Model Configurations and Training We evaluate both MoE models with MLP experts and with CNN experts. For each architecture, we vary the number of experts $M \in \{4, 8, 16, 32, 64\}$ and fix the number of neurons to $J = 8$. This means each MLP expert has 8 neurons per input patch, while each CNN expert has 8 neurons overall. Training follows Algorithm 1. Experts are optimized with normalized gradient descent at learning rate $\eta = 0.001$. The gating network is trained with gradient descent using learning rate finetuned for each M. In particular, $\eta_r = 0.1$ for $M \in \{4, 8, 16\}$, $\eta_r = 0.25$ for $M = 32$ and $\eta_r = 0.4$ for $M = 64$. For each M, we report model accuracy and dispatch entropy, summarized in Figure 1 and Figure 2.

Results and Observations. As shown in Figure 1, when $P = 4$ the MLP–MoE attains its highest accuracy at $M = 16$ while the CNN–MoE peaks at $M = 8$. Increasing M beyond these settings does not yield further gains and even reduce accuracy. It is worth noting that while $M = 64$ yields an increase in mean accuracy for the CNN–MoE from $M = 32$, it also introduces substantially greater variance. We also observe a clear effect of the patch count P. The CNN–MoE maintains strong performance for both $P = 4$ and $P = 8$, whereas the MLP–MoE exhibits a moderate decline at $P = 8$. This difference is consistent with the update mechanics of the experts: MLP experts update their weights on a per-patch basis, as formalized in Lemma 12, while CNN experts share weights across patches and therefore aggregate information across all patches during training. This shared-weight aggregation stabilizes learning for CNN experts and also explains the higher variance of the MLP–MoE across values of M.

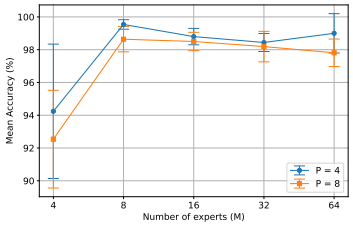
Figure 2 shows a parallel trend for dispatch entropy, indicating that changes in M affect both routing and expert learning in similar ways. This observation aligns with Theorem 1, which suggests the existence of an optimal range of expert counts that jointly supports generalization and accurate router dispatch. In particular, once M moves beyond this range, routing becomes less informative and expert specialization degrades, which mirrors the accuracy drop and the rise in dispatch entropy in Figure 1 and Figure 2.

Conclusion and Discussion

We addressed a key theoretical gap in MoE architecture by analyzing how the expert count M shapes performance and learning dynamics when experts are implemented as MLPs, aligning the theoretical model more closely with

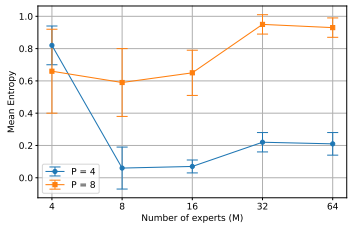


(a) MoE model with MLP experts

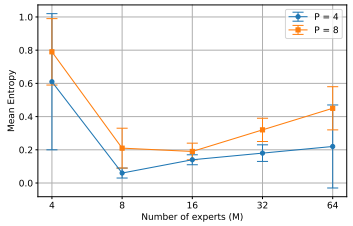


(b) MoE model with CNN experts

Figure 1: Mean accuracies of MoE models with MLP vs with CNN experts. Error bars represent standard deviations across five independent runs.



(a) MoE model with MLP experts



(b) MoE model with CNN experts

Figure 2: Mean dispatch entropies of MoE models with MLP vs with CNN experts. Error bars represent standard deviations across five independent runs.

modern LLM-style architectures. Building on the framework of Chen et al. (2022), we derived a range of expert counts that yields optimal performance and linked this range to the specialization and routing behavior of the model. These findings provide a principled basis for common empirical design choices in large-scale MoE systems and offer practical guidance on when adding experts promotes useful specialization versus when it produces diminishing returns.

Future Work A central next step is to tighten the optimality range by establishing necessity in addition to sufficiency. In particular, we aim to prove that when M falls outside the identified interval, performance must deteriorate. We formalize this as the proposition below, which we plan to prove under the setting of Section **Preliminaries**.

Theorem 2 (Upper bound on expert counts)**.** *Under the problem setting of Section* ***Preliminaries****, there exists an upper bound B (that may depend on distributional and optimization parameters) such that, for all expert counts $M \geq B$, the test error is high. Concretely,*

$$\Pr_{(x,y)\sim\mathcal{D}}\Big(y\, f_m\big(x; \mathbf{W}^{(T)}\big) \leq 0 \,\Big|\, (x,y) \in \Omega_k\Big) \;\geq\; \Omega(1)\,,$$

where $\mathbf{W}^{(T)}$ are the parameters after T training steps and Ω_k denotes cluster k. In words, beyond B experts, additional capacity does not reduce error below a fixed threshold.

Establishing Theorem 2 would yield tight bounds for M. A complementary direction is to extend the analysis to other attention-based experts, to non-uniform cluster sizes and heavy-tailed features, and to alternative routers such as top-k gating.

References

Aljundi, R.; Chakravarty, P.; and Tuytelaars, T. 2017. Expert gate: Lifelong learning with a network of experts. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 3366–3375.

Allen-Zhu, Z.; and Li, Y. 2020. Towards understanding ensemble, knowledge distillation and self-distillation in deep learning. *arXiv preprint arXiv:2012.09816*.

Chen, Z.; Deng, Y.; Wu, Y.; Gu, Q.; and Li, Y. 2022. Towards understanding the mixture-of-experts layer in deep learning. *Advances in neural information processing systems*, 35: 23049–23062.

Collobert, R.; Bengio, S.; and Bengio, Y. 2001. A parallel mixture of SVMs for very large scale problems. *Advances in Neural Information Processing Systems*, 14.

Databricks. 2024. Introducing DBRX: A New State-of-the-Art Open LLM. [Online].

Dauphin, Y. N.; Fan, A.; Auli, M.; and Grangier, D. 2017. Language modeling with gated convolutional networks. In *International conference on machine learning*, 933–941. PMLR.

Deisenroth, M.; and Ng, J. W. 2015. Distributed gaussian processes. In *International conference on machine learning*, 1481–1490. PMLR.

Eigen, D.; Ranzato, M.; and Sutskever, I. 2013. Learning factored representations in a deep mixture of experts. *arXiv preprint arXiv:1312.4314*.

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

Hoffmann, J.; Borgeaud, S.; Mensch, A.; Buchatskaya, E.; Cai, T.; Rutherford, E.; Casas, D. d. L.; Hendricks, L. A.; Welbl, J.; Clark, A.; et al. 2022. Training compute-optimal large language models. *arXiv preprint arXiv:2203.15556*.

Jacobs, R. A.; Jordan, M. I.; Nowlan, S. J.; and Hinton, G. E. 1991. Adaptive mixtures of local experts. *Neural computation*, 3(1): 79–87.

Jiang, A. Q.; Sablayrolles, A.; Roux, A.; Mensch, A.; Savary, B.; Bamford, C.; Chaplot, D. S.; Casas, D. d. l.; Hanna, E. B.; Bressand, F.; et al. 2024. Mixtral of experts. *arXiv preprint arXiv:2401.04088*.

Jiang, W.; and Tanner, M. A. 1999a. Hierarchical mixtures-of-experts for exponential family regression models: approximation and maximum likelihood estimation. *Annals of Statistics*, 987–1011.

Jiang, W.; and Tanner, M. A. 1999b. On the approximation rate of hierarchical mixtures-of-experts for generalized linear models. *Neural computation*, 11(5): 1183–1198.

Jordan, M.; Ghahramani, Z.; and Saul, L. 1996. Hidden Markov decision trees. *Advances in neural information processing systems*, 9.

Jordan, M. I.; and Jacobs, R. A. 1994. Hierarchical mixtures of experts and the EM algorithm. *Neural computation*, 6(2): 181–214.

Kaplan, J.; McCandlish, S.; Henighan, T.; Brown, T. B.; Chess, B.; Child, R.; Gray, S.; Radford, A.; Wu, J.; and Amodei, D. 2020. Scaling laws for neural language models. *arXiv preprint arXiv:2001.08361*.

Liu, A.; Feng, B.; Wang, B.; Wang, B.; Liu, B.; Zhao, C.; Dengr, C.; Ruan, C.; Dai, D.; Guo, D.; et al. 2024. Deepseek-v2: A strong, economical, and efficient mixture-of-experts language model. *arXiv preprint arXiv:2405.04434*.

Mendes, E. F.; and Jiang, W. 2012. On convergence rates of mixtures of polynomial experts. *Neural computation*, 24(11): 3025–3051.

Nguyen, H. D.; Lloyd-Jones, L. R.; and McLachlan, G. J. 2016. A universal approximation theorem for mixture-of-experts models. *Neural computation*, 28(12): 2585–2593.

Rasmussen, C.; and Ghahramani, Z. 2001. Infinite mixtures of Gaussian process experts. *Advances in neural information processing systems*, 14.

S.A.R.Team. 2024. Snowflake Arctic: The Best LLM for Enterprise AI — Efficiently Intelligent, Truly Open. [Online].

Shahbaba, B.; and Neal, R. 2009. Nonlinear models using Dirichlet process mixtures. *Journal of Machine Learning Research*, 10(8).

Shazeer, N.; Mirhoseini, A.; Maziarz, K.; Davis, A.; Le, Q.; Hinton, G.; and Dean, J. 2017. Outrageously large neural networks: The sparsely-gated mixture-of-experts layer. *arXiv preprint arXiv:1701.06538*.

Theis, L.; and Bethge, M. 2015. Generative image modeling using spatial lstms. *Advances in neural information processing systems*, 28.

Tresp, V. 2000. Mixtures of Gaussian processes. *Advances in neural information processing systems*, 13.

Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, Ł.; and Polosukhin, I. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

xAI. 2024. Grok-1. [Online].

Zeevi, A. J.; Meir, R.; and Maiorov, V. 2002. Error bounds for functional approximation and estimation using mixtures of experts. *IEEE Transactions on Information Theory*, 44(3): 1010–1025.

Proof of Main Results

Notations

Unlike CNN experts, MLP experts are sensitive to different orderings of input patches. For notational convenience, denote Q_i the set of random noise patches for input $\mathbf{x}_i$, that is $Q_i = \{q \in [P] : \mathbf{x}_i^{(q)} = \xi_{i,q}\}$, where $\xi_{i,q}$ are random noise generated as in Definition 1. Consider any $\tau = (p_1, p_2, p_3) \in P^3, p_i \neq p_j$ for $i \neq j$, according to the type of the feature noise and patch positions, we further divide Ω_k into $\Omega_k = \bigcup_{k',\tau} \Omega_{k,k',\tau}$, that is $\mathbf{x} \in \Omega_{k,k',\tau}$ if $\mathbf{x}^{(p_1)}, \mathbf{x}^{(p_2)}, \mathbf{x}^{(p_3)}$ are feature signal, feature noise, and cluster-center signal respectively. To better characterize the router training, we also break down $\Omega_{k,k',\tau}$ into $\Omega^+_{k,k',\tau}$ and $\Omega^-_{k,k',\tau}$ where $\Omega^+_{k,k',\tau} = \{\mathbf{x}_i \in \Omega_{k,k',\tau} | y_i = \varepsilon_i\}$ and $\Omega^-_{k,k',\tau} = \{\mathbf{x}_i \in \Omega_{k,k',\tau} | y_i = -\varepsilon_i\}$. Given an example $\mathbf{x}$, unless otherwise stated, we assume $p_1, p_2, p_3 \in [P]$ are patch positions of feature signal, feature noise, and cluster-center signal, respectively.

Initialization properties

For the new MoE architecture, the initialization properties established for the prior CNN-expert architecture still apply (Chen et al. 2022). For completeness, we restate them here without proof.

Lemma 4. *With probability at least $1 - \delta$, the following properties hold for all $k \in [K]$,*

$$\sum_{i\in\Omega_k} y_i\beta_i^3 = \widetilde{O}(\sqrt{n}), \quad \sum_{i\in\Omega_k} \alpha_i^3 = \mathbb{E}[\alpha^3]\cdot n/K + \widetilde{O}(\sqrt{n}), \quad \sum_{i\in\Omega_k} y_i\varepsilon_i\gamma_i^3 = \widetilde{O}(\sqrt{n}), \tag{8}$$

$$\sum_{i\in\Omega^+_{k,k'}} y_i\alpha_i = \widetilde{O}(\sqrt{n}), \quad \sum_{i\in\Omega^-_{k,k'}} y_i\alpha_i = \widetilde{O}(\sqrt{n}), \quad \sum_{i\in\Omega^+_{k,k'}} \varepsilon_i\gamma_i = \widetilde{O}(\sqrt{n}), \tag{9}$$

$$\sum_{i\in\Omega^-_{k,k'}} \varepsilon_i\gamma_i = \widetilde{O}(\sqrt{n}), \quad \sum_{i\in\Omega_k} \beta_i = \mathbb{E}[\beta]\cdot n/K + \widetilde{O}(\sqrt{n}). \tag{10}$$

Lemma 5. *Suppose that $d = \Omega(\log(4nP/\delta))$. With probability at least $1-\delta$, the following inequalities hold for all $i \in [n], k \in [K], q \in Q_i$,*

- $\|\xi_{i,q}\|_2 = O(1)$,
- $\langle \mathbf{v}_k, \xi_{i,q}\rangle \leq \widetilde{O}(d^{-1/2})$, $\langle \mathbf{c}_k, \xi_{i,q}\rangle \leq \widetilde{O}(d^{-1/2})$, $\langle \xi_{i,q}, \xi_{i,q'}\rangle \leq \widetilde{O}(d^{-1/2}), \quad \forall (i',q') \neq (i,q)$.

MoE Initialization Property. We divide the experts into K sets based on the initialization.

Definition 2. *Fix expert $m \in [M]$, note that for each $j \in [J]$ $\mathbf{w}^{(0)}_{m,j,1} = \mathbf{w}^{(0)}_{m,j,2} = \cdots = \mathbf{w}^{(0)}_{m,j,P}$. Denote $(k^*_m, j^*_m) = \arg\max_{j,k}\langle \mathbf{v}_k, \mathbf{w}^{(0)}_{m,j,1}\rangle$. Fix cluster $k \in [K]$, denote the profession experts set as $\mathcal{M}_k = \{m \mid k^*_m = k\}$.*

Lemma 6. *For $M \geq \Theta(K\log(K/\delta))$, $J \geq \Theta(\log(M/\delta))$, the following inequalities hold with probability at least $1-\delta$.*

- $\max_{(j,k)\neq(j^*_m,k^*_m)}\langle \mathbf{w}^{(0)}_{m,j,p}, \mathbf{v}_k\rangle \leq \left(1 - \frac{\delta}{3MJ^2K^2}\right)\langle \mathbf{w}^{(0)}_{m,j^*_m,p}, \mathbf{v}_{k^*_m}\rangle, \quad \forall p \in [P], m \in [M]$.
- $\langle \mathbf{w}^{(0)}_{m,j^*_m,p}, \mathbf{v}_{k^*_m}\rangle \geq 0.01\sigma_0 \quad \forall m \in [M], p \in [P]$.
- $|\mathcal{M}_k| \geq 1 \quad \forall k \in [K]$.

Lemma 7. *Suppose the conclusions in Lemma5 hold, then with probability at least $1-\delta$ we have that $|\langle \mathbf{w}^{(0)}_{m,j,p}, \mathbf{v}\rangle| \leq \widetilde{O}(\sigma_0)$ for all $\mathbf{v} \in \{\mathbf{v}_k\}_{k\in[K]} \cup \{\mathbf{c}_k\}_{k\in[K]} \cup \{\xi_{i,q}\}_{i\in[n],q\in Q_i}$, $m \in [M], j \in [J], p \in [P]$.*

Exploration Stage

Let $T_1 := \left\lfloor \eta^{-1}\sigma_0^{1/2} \right\rfloor$. The exploration stage ends at iteration $t = T_1$. Building on Chen et al. (2022), we establish a set of parameter properties during this stage. In particular, our analysis shows that the growth of the inner products between the weights and the patch signals is governed by their initial alignment gap, which in turn depends on the number of experts M. The detailed arguments appear in Lemmas 13, 14, and 15.

Lemma 8. *For all $t \geq 0$ and any patch $p \in [P]$, we have $\sum_{m=1}^M \nabla_{\theta_{m,p}}\mathcal{L}^{(t)} = 0$, and thus $\sum_m \theta^{(t)}_{m,p} = \sum_m \theta^{(0)}_{m,p}$. In particular, if $\mathbf{\Theta}$ is zero-initialized, then $\sum_m \theta^{(t)}_{m,p} = 0$.*

Proof. The router gradient at iteration t is

$$\begin{aligned}\nabla_{\theta_{m,p}}\mathcal{L}^{(t)} &= \frac{1}{n}\sum_i \mathbb{1}(m_{i,t}=m)\,\ell'_{i,t}\,\pi_{m_{i,t}}(\mathbf{x}_i;\mathbf{\Theta}^{(t)})\big(1-\pi_{m_{i,t}}(\mathbf{x}_i;\mathbf{\Theta}^{(t)})\big)\,y_i f_{m_{i,t}}(\mathbf{x}_i;\mathbf{W}^{(t)})\mathbf{x}_i^{(p)} \\ &\quad - \frac{1}{n}\sum_i \mathbb{1}(m_{i,t}\neq m)\,\ell'_{i,t}\,\pi_{m_{i,t}}(\mathbf{x}_i;\mathbf{\Theta}^{(t)})\pi_m(\mathbf{x}_i;\mathbf{\Theta}^{(t)})\,y_i f_{m_{i,t}}(\mathbf{x}_i;\mathbf{W}^{(t)})\mathbf{x}_i^{(p)} \\ &= \frac{1}{n}\sum_i \mathbb{1}(m_{i,t}=m)\,\ell'_{i,t}\,\pi_{m_{i,t}}(\mathbf{x}_i;\mathbf{\Theta}^{(t)})\,y_i f_{m_{i,t}}(\mathbf{x}_i;\mathbf{W}^{(t)})\mathbf{x}_i^{(p)} \\ &\quad - \frac{1}{n}\sum_i \ell'_{i,t}\,\pi_{m_{i,t}}(\mathbf{x}_i;\mathbf{\Theta}^{(t)})\pi_m(\mathbf{x}_i;\mathbf{\Theta}^{(t)})\,y_i f_{m_{i,t}}(\mathbf{x}_i;\mathbf{W}^{(t)})\mathbf{x}_i^{(p)},\end{aligned}$$

where $\mathbb{1}(\cdot)$ denotes the indicator function.

Summing over $m=1,\dots,M$ yields

$$\begin{aligned}\sum_{m=1}^{M}\nabla_{\theta_{m,p}}\mathcal{L}^{(t)} &= \frac{1}{n}\sum_{i\in[n]}\ell'_{i,t}\,\pi_{m_{i,t}}(\mathbf{x}_i;\mathbf{\Theta}^{(t)})\,y_i f_{m_{i,t}}(\mathbf{x}_i;\mathbf{W}^{(t)})\mathbf{x}_i^{(p)} \\ &\quad - \frac{1}{n}\sum_{i\in[n]}\ell'_{i,t}\,\pi_{m_{i,t}}(\mathbf{x}_i;\mathbf{\Theta}^{(t)})\,y_i f_{m_{i,t}}(\mathbf{x}_i;\mathbf{W}^{(t)})\mathbf{x}_i^{(p)} \\ &= 0.\end{aligned}$$

Therefore $\sum_m \nabla_{\theta_{m,p}}\mathcal{L}^{(t)}=0$ for all t, which implies $\sum_m \theta_{m,p}^{(t)}$ is invariant across iterations. The stated consequences follow immediately. □

Lemma 9. *With probability at least $1-1/d$, uniformly over all vectors $\mathbf{v}\in\{\mathbf{v}_k\}_{k\in[K]}\cup\{\mathbf{c}_k\}_{k\in[K]}$ and indices $m\in[M], j\in[J], p\in[P]$, the following bounds hold:*

$$\Big|\langle\nabla_{\theta_{m,p}}\mathcal{L}^{(t)},\mathbf{v}\rangle| - \mathbb{E}[\langle\nabla_{\theta_{m,p}}\mathcal{L}^{(t)},\mathbf{v}\rangle]\Big| = \widetilde{O}\big(n^{-1/2}(\sigma_0+\eta t)^3\big),$$

$$\Big|\langle\nabla_{\mathbf{w}_{m,j,p}}\mathcal{L}^{(t)},\mathbf{v}\rangle - \mathbb{E}[\langle\nabla_{\mathbf{w}_{m,j,p}}\mathcal{L}^{(t)},\mathbf{v}\rangle]\Big| = \widetilde{O}\big(n^{-1/2}(\sigma_0+\eta t)^2\big),$$

for all $t\le d^{100}$.

Here, $\mathbb{E}[\langle\nabla_{\mathbf{w}_{m,j,p}}\mathcal{L}^{(t)},\mathbf{v}\rangle]$ and $\mathbb{E}[\langle\nabla_{\theta_{m,p}}\mathcal{L}^{(t)},\mathbf{v}\rangle]$ are given by

$$\mathbb{E}[\langle\nabla_{\theta_{m,p}}\mathcal{L}^{(t)},\mathbf{v}\rangle] = \frac{1}{n}\sum_i \mathbb{P}(m_{i,t}=m)\ell'_{i,t}\pi_m(\mathbf{x}_i;\mathbf{\Theta}^{(t)})y_i f_m(\mathbf{x}_i;\mathbf{W}^{(t)})\langle\mathbf{x}_i^{(p)},\mathbf{v}\rangle$$

$$-\frac{1}{n}\sum_{i,m'}\mathbb{P}(m_{i,t}=m')\ell'_{i,t}\pi_{m'}(\mathbf{x}_i;\mathbf{\Theta}^{(t)})\pi_m(\mathbf{x}_i;\mathbf{\Theta}^{(t)})y_i f_{m'}(\mathbf{x}_i;\mathbf{W}^{(t)})\langle\mathbf{x}_i^{(p)},\mathbf{v}\rangle,$$

$$\mathbb{E}[\langle\nabla_{\mathbf{w}_{m,j,p}}\mathcal{L}^{(t)},\mathbf{v}\rangle] = \frac{1}{n}\sum_i \mathbb{P}(m_{i,t}=m)\ell'_{i,t}\pi_m(\mathbf{x}_i;\mathbf{\Theta}^{(t)})y_i\sigma'(\langle\mathbf{w}_{m,j,p}^{(t)},\mathbf{x}_i^{(p)}\rangle)\langle\mathbf{x}_i^{(p)},\mathbf{v}\rangle.$$

Proof. Under normalized gradient descent, we have $\|\mathbf{w}_{m,j,p}^{(t)}-\mathbf{w}_{m,j,p}^{(0)}\|_2\le O(\eta t)$. By Lemma 7, this implies $|\langle\mathbf{w}_{m,j,p}^{(t)},\mathbf{x}_i^{(p)}\rangle|\le\widetilde{O}(\sigma_0+\eta t)$. Hence

$$\langle\nabla_{\mathbf{w}_{m,j,p}}\mathcal{L}^{(t)},\mathbf{v}\rangle = \frac{1}{n}\sum_i A_{i,p},$$

where $A_{i,p}=\mathbb{1}(m_{i,t}=m)\,\ell'_{i,t}\,\pi_m(\mathbf{x}_i;\mathbf{\Theta}^{(t)})\,y_i\,\sigma'(\langle\mathbf{w}_{m,j}^{(t)},\mathbf{x}_i^{(p)}\rangle)\langle\mathbf{x}_i^{(p)},\mathbf{v}\rangle$, $i\in[n], p\in[P]$, are independent and satisfy $|A_i|\le\widetilde{O}((\sigma_0+\eta t)^2)$.

Hoeffding's inequality yields, with probability at least $1-1/(4d^{101}MJK)$,

$$\Big|\langle\nabla_{\mathbf{w}_{m,j,p}}\mathcal{L}^{(t)},\mathbf{v}\rangle - \mathbb{E}[\langle\nabla_{\mathbf{w}_{m,j,p}}\mathcal{L}^{(t)},\mathbf{v}\rangle]\Big| = \widetilde{O}(n^{-1/2}(\sigma_0+\eta t)^2).$$

A union bound then gives, with probability at least $1-1/(2d)$,

$$|\langle\nabla_{\mathbf{w}_{m,j,p}}\mathcal{L}^{(t)},\mathbf{v}\rangle - \mathbb{E}[\langle\nabla_{\mathbf{w}_{m,j,p}}\mathcal{L}^{(t)},\mathbf{v}\rangle]| = \widetilde{O}(n^{-1/2}(\sigma_0+\eta t)^2),\quad \forall m\in[M], j\in[J], p\in[P], t\le d^{100}.$$

An identical argument, accounting for the additional dependence on routing probabilities, shows

$$|\langle \nabla_{\theta_{m,p}} \mathcal{L}^{(t)}, \mathbf{v} \rangle - \mathbb{E}[\langle \nabla_{\theta_{m,p}} \mathcal{L}^{(t)}, \mathbf{v} \rangle]| = \widetilde{O}(n^{-1/2}(\sigma_0 + \eta t)^3),$$

which completes the proof. □

Lemma 10. *For all $t \le T_1$, the following properties hold:*

- $\langle \mathbf{w}_{m,j,p}^{(t)}, \mathbf{v}_k \rangle = O(\sigma_0^{0.5}), \quad \langle \mathbf{w}_{m,j,p}^{(t)}, \mathbf{c}_k \rangle = O(\sigma_0^{0.5}), \quad \langle \mathbf{w}_{m,j,p}^{(t)}, \xi_{i,q} \rangle = \widetilde{O}(\sigma_0^{0.5})$,
- $f_m(\mathbf{x}_i; \mathbf{W}^{(t)}) = \widetilde{O}(\sigma_0^{1.5})$,
- $|\ell'_{i,t} + 1/2| \le \widetilde{O}(\sigma_0^{1.5})$,
- $\|\theta_m^{(t)}\|_2 \le \widetilde{O}(\sigma_0^{1.5})$,
- $\|h(\mathbf{x}_i; \mathbf{\Theta}^{(t)})\|_\infty = \widetilde{O}(\sigma_0^{1.5}), \quad \pi_m(\mathbf{x}_i; \mathbf{\Theta}^{(t)}) = M^{-1} + \widetilde{O}(\sigma_0^{1.5})$,

for all $m \in [M], k \in [K], i \in [n], q \in Q_i$.

Proof. The first property follows since $\|\mathbf{w}_{m,j,p}^{(t)} - \mathbf{w}_{m,j,p}^{(0)}\|_2 \le O(\eta T_1) = O(\sigma_0^{0.5})$, and thus

$$|f_m(\mathbf{x}_i; \mathbf{W}^{(t)})| \le \sum_{p \in [P]} \sum_{j \in [J]} |\sigma(\langle \mathbf{w}_{m,j,p}^{(t)}, \mathbf{x}_i^{(p)} \rangle)| = \widetilde{O}(\sigma_0^{1.5}).$$

We next show that the loss derivative remains close to $-1/2$ during this stage. Let $s = y_i \pi_{m_i,t}(\mathbf{x}_i; \mathbf{\Theta}^{(t)}) f_{m_i,t}(\mathbf{x}_i, \mathbf{W}^{(t)})$. Then $|s| = \widetilde{O}(\sigma_0^{1.5})$ and

$$\left|\ell'_{i,t} + \tfrac{1}{2}\right| = \left| -\frac{1}{e^s + 1} + \tfrac{1}{2} \right| \le^{(i)} |s| = \widetilde{O}(\sigma_0^{1.5}),$$

where (i) is verified by considering the cases $|s| \le 1$ and $|s| > 1$.

We now prove the fourth bullet in Lemma 10. Because $|f_m| = \widetilde{O}(\sigma_0^{1.5})$, we can upper bound the gradient of the gating network by

$$\|\nabla_{\theta_{m,p}} \mathcal{L}^{(t)}\|_2 = \Bigg\| \frac{1}{n} \sum_i \mathbb{1}(m_{i,t} = m) \ell'_{i,t} \pi_{m_i,t}(\mathbf{x}_i; \mathbf{\Theta}^{(t)}) y_i f_{m_i,t}(\mathbf{x}_i; \mathbf{W}^{(t)}) \mathbf{x}_i^{(p)}$$

$$- \frac{1}{n} \sum_i \ell'_{i,t} \pi_{m_i,t}(\mathbf{x}_i; \mathbf{\Theta}^{(t)}) \pi_m(\mathbf{x}_i; \mathbf{\Theta}^{(t)}) y_i f_{m_i,t}(\mathbf{x}_i; \mathbf{W}^{(t)}) \mathbf{x}_i^{(p)} \Bigg\|_2 = \widetilde{O}(\sigma_0^{1.5}),$$

where we used $|\ell'_{i,t}| \le 1$, $\pi_m, \pi_{m_i,t} \in [0,1]$, and $\|\mathbf{x}_i^{(p)}\|_2 = O(1)$.

This further implies

$$\|\theta_{m,p}^{(t)}\|_2 = \|\theta_{m,p}^{(t)} - \theta_{m,p}^{(0)}\|_2 \le \widetilde{O}(\sigma_0^{1.5} t \eta_r) = \widetilde{O}(\sigma_0^{1.5}),$$

where the last inequality uses $\eta_r = \Theta(M^2)\eta$.

Finally, the bounds $\|h(\mathbf{x}_i; \mathbf{\Theta}^{(t)})\|_\infty \le O(\sigma_0^{1.5})$ and $\pi_m(\mathbf{x}_i; \mathbf{\Theta}^{(t)}) = M^{-1} + O(\sigma_0^{1.5})$ follow directly from $\|\theta_{m,p}^{(t)}\|_2 = \widetilde{O}(\sigma_0^{1.5})$. □

Lemma 11. $\max_{m \in [M]} \left|\mathbb{P}(m_{i,t} = m) - 1/M\right| = \widetilde{O}(\sigma_0^{1.5})$ *for all $t \le T_1, i \in [n], m \in [M]$.*

Proof. By Lemma 10 we have $\|h(\mathbf{x}_i; \mathbf{\Theta}^{(t)})\|_\infty \le \widetilde{O}(\sigma_0^{1.5})$. Applying Lemma 1 then yields

$$\max_{m \in [M]} \left|\mathbb{P}(m_{i,t} = m) - \tfrac{1}{M}\right| = \widetilde{O}(\sigma_0^{1.5}).$$

This completes the proof. □

Lemma 12. *We have the following gradient update rules hold for the experts,*

$$\langle \nabla_{\mathbf{w}_{m,j,p}} \mathcal{L}^{(t)}, \mathbf{v}_k \rangle = \frac{-\mathbb{E}[\alpha^3] + \widetilde{O}(\sigma_0^{1.5})}{2KM^2P} \sigma'(\langle \mathbf{w}_{m,j,p}^{(t)}, \mathbf{v}_k \rangle) + \widetilde{O}(\sigma_0^{2.5}),$$

$$\langle \nabla_{\mathbf{w}_{m,j,p}} \mathcal{L}^{(t)}, \mathbf{c}_k \rangle = -\widetilde{O}(\sigma_0^{1.5}) \sigma'(\langle \mathbf{w}_{m,j,p}^{(t)}, \mathbf{c}_k \rangle) + \widetilde{O}(\sigma_0^{2.5}),$$

$$\langle \nabla_{\mathbf{w}_{m,j,p}} \mathcal{L}^{(t)}, \xi_{i,q} \rangle = -\widetilde{O}(\sigma_0^{1.5}) \sigma'(\langle \mathbf{w}_{m,j,p}^{(t)}, \xi_{i,q} \rangle) + \widetilde{O}(\sigma_0^{2.5}),$$

for all $t \le T_1, j \in [J], k \in [K], m \in [M], p \in [P], q \in Q_i, i \in [n]$.

Proof. The experts' gradients can be computed as follows,

$$\nabla_{\mathbf{w}_{m,j,p}}\mathcal{L}^{(t)} = \frac{1}{n}\sum_{i\in[n]} \mathbb{1}(m_{i,t}=m)\ell'_{i,t}\pi_m(\mathbf{x}_i;\mathbf{\Theta}^{(t)})y_i\sigma'\Big(\langle \mathbf{w}_{m,j,p}^{(t)}, \mathbf{x}_i^{(p)}\rangle\Big)\mathbf{x}_i^{(p)}.$$

We first compute the inner product against cluster-center signal $\langle\nabla_{\mathbf{w}_{m,j,p}}\mathcal{L}^{(t)}, \mathbf{c}_k\rangle$. By Lemma 9, we have that $\big|\langle\nabla_{\mathbf{w}_{m,j,p}}\mathcal{L}^{(t)}, \mathbf{c}_k\rangle - \mathbb{E}[\langle\nabla_{\mathbf{w}_{m,j,p}}\mathcal{L}^{(t)}, \mathbf{c}_k\rangle]\big| = \widetilde{O}(n^{-1/2}\sigma_0) \le \widetilde{O}(\sigma_0^{2.5})$. Therefore we can compute $\langle\nabla_{\mathbf{w}_{m,j,p}}\mathcal{L}^{(t)}, \mathbf{c}_k\rangle$ through its expectation

$$\begin{aligned}
&\mathbb{E}[\langle\nabla_{\mathbf{w}_{m,j,p}}\mathcal{L}^{(t)}, \mathbf{c}_k\rangle]\\
&= \frac{1}{n}\sum_{i\in[n]}\mathbb{P}(m_{i,t}=m)\mathbb{E}\Big[\ell'_{i,t}\pi_m(\mathbf{x}_i;\mathbf{\Theta}^{(t)})\sigma'\Big(\langle \mathbf{w}_{m,j,p}^{(t)}, \mathbf{x}_i^{(p)}\rangle\Big)y_i\langle\mathbf{x}_i^{(p)}, \mathbf{c}_k\rangle\Big]\\
&= \frac{1}{nP}\sum_{i\in\Omega_k}\mathbb{P}(m_{i,t}=m)\mathbb{E}\Big[\ell'_{i,t}\pi_m(\mathbf{x}_i;\mathbf{\Theta}^{(t)})\sigma'\Big(\langle \mathbf{w}_{m,j,p}^{(t)}, \mathbf{c}_k\rangle\Big)y_i\beta_i^3\|\mathbf{c}_k\|_2^2 \mid \mathbf{x}_i^{(p)} = \beta_i\mathbf{c}_k\Big]\\
&\quad + \frac{1}{nP}\sum_{i\in[n],q\in Q_i}\mathbb{P}(m_{i,t}=m)\mathbb{E}\Big[\ell'_{i,t}\pi_m(\mathbf{x}_i;\mathbf{\Theta}^{(t)})\sigma'\Big(\langle \mathbf{w}_{m,j,p}^{(t)}, \xi_{i,q}\rangle\Big)y_i\langle\mathbf{c}_k, \xi_{i,q}\rangle \mid \mathbf{x}_i^{(p)} = \xi_{i,q}\Big]\\
&= \left[-\frac{1}{2nMP}\sum_{i\in\Omega_k}y_i\beta_i^3\mathbb{P}(m_{i,t}=m) + \widetilde{O}(\sigma_0^{1.5})\right]\sigma'\Big(\langle \mathbf{w}_{m,j,p}^{(t)}, \mathbf{c}_k\rangle\Big) + \widetilde{O}(\sigma_0^{2.5})\\
&= \widetilde{O}(n^{-1/2}+\sigma_0^{1.5})\sigma'\Big(\langle \mathbf{w}_{m,j,p}^{(t)}, \mathbf{c}_k\rangle\Big) + \widetilde{O}(\sigma_0^{2.5})\\
&= \widetilde{O}(\sigma_0^{1.5})\sigma'\Big(\langle \mathbf{w}_{m,j,p}^{(t)}, \mathbf{c}_k\rangle\Big) + \widetilde{O}(\sigma_0^{2.5}).
\end{aligned}$$

where the second equality is by taking expectation over random patch permutations, conditioned on $\mathbf{x}_i^{(p)}$, the third equality is due to Lemma 10 and Lemma 5, the fourth equality is due to Lemma 11, and the last equality is by the choice of n and σ_0. With the same argument, we can compute the inner product against random noise, for any $q \in Q_i$

$$\langle\nabla_{\mathbf{w}_{m,j,p}}\mathcal{L}^{(t)}, \xi_{i,q}\rangle = \widetilde{O}(\sigma_0^{1.5})\sigma'\Big(\langle \mathbf{w}_{m,j,p}^{(t)}, \xi_{i,q}\rangle\Big) + \widetilde{O}(\sigma_0^{2.5})$$

Next, we compute the inner product against the feature signal $\langle\nabla_{\mathbf{w}_{m,j,p}}\mathcal{L}^{(t)}, \mathbf{v}_k\rangle$

$$\begin{aligned}
&\mathbb{E}[\langle\nabla_{\mathbf{w}_{m,j,p}}\mathcal{L}^{(t)}, \mathbf{v}_k\rangle]\\
&= \frac{1}{n}\sum_{i\in[n]}\mathbb{P}(m_{i,t}=m)\mathbb{E}\Big[\ell'_{i,t}\pi_m(\mathbf{x}_i;\mathbf{\Theta}^{(t)})\sigma'\Big(\langle \mathbf{w}_{m,j,p}^{(t)}, \mathbf{x}_i^{(p)}\rangle\Big)y_i\langle\mathbf{x}_i^{(p)}, \mathbf{v}_k\rangle\Big]\\
&= \frac{1}{nP}\sum_{i\in\Omega_k}\mathbb{P}(m_{i,t}=m)\mathbb{E}\Big[\ell'_{i,t}\pi_m(\mathbf{x}_i;\mathbf{\Theta}^{(t)})\sigma'\Big(\langle \mathbf{w}_{m,j,p}^{(t)}, \mathbf{v}_k\rangle\Big)\alpha_i^3\|\mathbf{v}_k\|_2^2 \mid \mathbf{x}_i^{(p)} = y_i\alpha_i\mathbf{v}_k\Big]\\
&\quad + \frac{1}{nP}\sum_{k'\neq k}\sum_{i\in\Omega_{k',k}}\mathbb{P}(m_{i,t}=m)\mathbb{E}\Big[\ell'_{i,t}\pi_m(\mathbf{x}_i;\mathbf{\Theta}^{(t)})\sigma'\Big(\langle \mathbf{w}_{m,j,p}^{(t)}, \mathbf{v}_k\rangle\Big)\gamma_i^3y_i\varepsilon_i\|\mathbf{v}_k\|_2^2 \mid \mathbf{x}_i^{(p)} = \varepsilon_i\gamma_i\mathbf{v}_k\Big]\\
&\quad + \frac{1}{nP}\sum_{i\in[n],q\in Q_i}\mathbb{P}(m_{i,t}=m)\mathbb{E}\Big[\ell'_{i,t}\pi_m(\mathbf{x}_i;\mathbf{\Theta}^{(t)})\sigma'\Big(\langle \mathbf{w}_{m,j,p}^{(t)}, \xi_{i,q}\rangle\Big)y_i\langle\mathbf{v}_k, \xi_{i,q}\rangle \mid \mathbf{x}_i^{(p)} = \xi_{i,q}\Big]\\
&= \left[-\frac{1}{2nMP}\sum_{i\in\Omega_k}\mathbb{P}(m_{i,t}=m)\alpha_i^3 - \frac{1}{2nMP}\sum_{i\in\Omega_{k',k}}\mathbb{P}(m_{i,t}=m)\gamma_i^3y_i\varepsilon_i + O(\sigma_0^{1.5})\right]\sigma'\Big(\langle \mathbf{w}_{m,j,p}^{(t)}, \mathbf{v}_k\rangle\Big) + \widetilde{O}(\sigma_0^{2.5})\\
&= \left(-\frac{\mathbb{E}[\alpha^3]}{2KM^2P} + \widetilde{O}(n^{-1/2}+\sigma_0^{1.5})\right)\sigma'\Big(\langle \mathbf{w}_{m,j,p}^{(t)}, \mathbf{v}_k\rangle\Big) + \widetilde{O}(\sigma_0^{2.5})\\
&= -\left(\frac{\mathbb{E}[\alpha^3]}{2KM^2P} + \widetilde{O}(\sigma_0^{1.5})\right)\sigma'\Big(\langle \mathbf{w}_{m,j,p}^{(t)}, \mathbf{v}_k\rangle\Big) + \widetilde{O}(\sigma_0^{2.5}).
\end{aligned}$$

where the second equality is by taking expectation over random patch permutations, conditioned on $\mathbf{x}_i^{(p)}$, the third equality is due to Lemma 10 and Lemma 5, the fourth equality is due to Lemma 11, and the last equality is by the choice of n and σ_0, which ends our proof. □

Lemma 13. *For all $m \in [M], p \in [P]$ and $t \leq T_1$, the following inequalities hold:*

$$\begin{aligned}
\langle \mathbf{w}_{m,j_m^*,p}^{(t)}, \mathbf{v}_{k_m^*} \rangle &= O(\sigma_0^{0.5}), \\
\langle \mathbf{w}_{m,j,p}^{(t)}, \mathbf{v}_k \rangle &= \widetilde{O}(M\sigma_0), \quad \forall (j,k) \neq (j_m^*, k_m^*), \\
\langle \mathbf{w}_{m,j,p}^{(t)}, \mathbf{c}_k \rangle &= \widetilde{O}(\sigma_0), \quad \forall j \in [J], k \in [K], \\
\langle \mathbf{w}_{m,j,p}^{(t)}, \xi_{i,q} \rangle &= \widetilde{O}(\sigma_0), \quad \forall j \in [J], i \in [n], q \in Q_i.
\end{aligned}$$

Proof. For $t \leq T_1$, the update rule for every expert can be written as

$$\begin{aligned}
\langle \mathbf{w}_{m,j,p}^{(t+1)}, \mathbf{v}_k \rangle &= \langle \mathbf{w}_{m,j,p}^{(t)}, \mathbf{v}_k \rangle + \frac{\eta}{\|\nabla \mathbf{W}_m \mathcal{L}^{(t)}\|_F} \left[\frac{3\mathbb{E}[\alpha^3] + \widetilde{O}(\sigma_0^{1.5})}{2KM^2P} \langle \mathbf{w}_{m,j,p}^{(t)}, \mathbf{v}_k \rangle^2 + \widetilde{O}(\sigma_0^{2.5}) \right], \\
\langle \mathbf{w}_{m,j,p}^{(t+1)}, \xi_{i,q} \rangle &= \langle \mathbf{w}_{m,j,p}^{(t)}, \xi_{i,q} \rangle + \frac{\eta}{\|\nabla \mathbf{W}_m \mathcal{L}^{(t)}\|_F} \left[\widetilde{O}(\sigma_0^{1.5}) \langle \mathbf{w}_{m,j,p}^{(t)}, \xi_{i,q} \rangle^2 + \widetilde{O}(\sigma_0^{2.5}) \right], \\
\langle \mathbf{w}_{m,j,p}^{(t+1)}, \mathbf{c}_k \rangle &= \langle \mathbf{w}_{m,j,p}^{(t)}, \mathbf{c}_k \rangle + \frac{\eta}{\|\nabla \mathbf{W}_m \mathcal{L}^{(t)}\|_F} \left[\widetilde{O}(\sigma_0^{1.5}) \langle \mathbf{w}_{m,j,p}^{(t)}, \mathbf{c}_k \rangle^2 + \widetilde{O}(\sigma_0^{2.5}) \right].
\end{aligned} \tag{11}$$

We have $\langle \mathbf{w}_{m,j,p}^{(t)}, \mathbf{v}_{k_m}^* \rangle \leq O(\sigma_0^{0.5})$ for all $t \leq T_1$. Comparing the evolution of $\langle \mathbf{w}_{m,j,p}^{(t)}, \mathbf{v}_{k_m}^* \rangle$ to the other inner products in (11) shows that $\langle \mathbf{w}_{m,j,p}^{(t)}, \mathbf{v}_{k_m}^* \rangle$ grows to order $\sigma_0^{0.5}$ while the remaining inner products stay nearly unchanged.

Comparison with $\langle \mathbf{w}_{m,j,p}^{(t)}, \mathbf{v}_k \rangle$. Fix $k \neq k_m^*$. To obtain an upper bound, assume without loss of generality that $\langle \mathbf{w}_{m,j,p}^{(t)}, \mathbf{v}_k \rangle = \Omega(\sigma_0)$. Then

$$\langle \mathbf{w}_{m,j,p}^{(t)}, \mathbf{v}_k \rangle^2 + \widetilde{O}(\sigma_0^{2.5}) = (1 + \widetilde{O}(\sigma_0^{0.5})) \langle \mathbf{w}_{m,j,p}^{(t)}, \mathbf{v}_k \rangle^2.$$

The equality is due to factoring out $\langle \mathbf{w}_{m,j,p}^{(t)}, \mathbf{v}_k \rangle^2$ and absorbing the $\widetilde{O}(\sigma_0^{2.5})$ remainder into a multiplicative $(1 + \widetilde{O}(\sigma_0^{0.5}))$ term under the assumption $\langle \mathbf{w}_{m,j,p}^{(t)}, \mathbf{v}_k \rangle = \Omega(\sigma_0)$.

Therefore,

$$\begin{aligned}
\langle \mathbf{w}_{m,j,p}^{(t+1)}, \mathbf{v}_{k_m^*} \rangle &= \langle \mathbf{w}_{m,j,p}^{(t)}, \mathbf{v}_{k_m^*} \rangle + \frac{\eta}{\|\nabla \mathbf{W}_m \mathcal{L}^{(t)}\|_F} \frac{3\mathbb{E}[\alpha^3] + \widetilde{O}(M^2\sigma_0^{0.5})}{2KM^2P} \langle \mathbf{w}_{m,j,p}^{(t)}, \mathbf{v}_{k_m^*} \rangle^2, \\
\langle \mathbf{w}_{m,j,p}^{(t+1)}, \mathbf{v}_k \rangle &= \langle \mathbf{w}_{m,j,p}^{(t)}, \mathbf{v}_k \rangle + \frac{\eta}{\|\nabla \mathbf{W}_m \mathcal{L}^{(t)}\|_F} \frac{3\mathbb{E}[\alpha^3] + \widetilde{O}(M^2\sigma_0^{0.5})}{2KM^2P} \langle \mathbf{w}_{m,j,p}^{(t)}, \mathbf{v}_k \rangle^2.
\end{aligned}$$

Applying Lemma 27 with $C_t = \big(3\mathbb{E}[\alpha^3] + \widetilde{O}(M^2\sigma_0^{0.5})\big) \big/ \big(2KM^2P \|\nabla \mathbf{W}_m \mathcal{L}^{(t)}\|_F\big)$, $S = 1$, $G = 1/(3\log(d)MJ^2K^2)$, and using the initialization event $\langle \mathbf{w}_m^{(0)}, \mathbf{v}_{k_m}^* \rangle \geq S(1+G)\langle \mathbf{w}_m^{(0)}, \mathbf{v}_k \rangle$, yields

$$\langle \mathbf{w}_{m,j,p}^{(t)}, \mathbf{v}_k \rangle \leq O(G^{-1}\sigma_0) = \widetilde{O}(M\sigma_0).$$

Comparison with $\langle \mathbf{w}_{m,j,p}^{(t)}, \mathbf{c}_k \rangle$. Assume $\langle \mathbf{w}_{m,j,p}^{(t)}, \mathbf{c}_k \rangle = \Omega(\sigma_0)$. Then

$$\begin{aligned}
\langle \mathbf{w}_{m,j,p}^{(t+1)}, \mathbf{v}_{k_m^*} \rangle &= \langle \mathbf{w}_{m,j,p}^{(t)}, \mathbf{v}_{k_m^*} \rangle + \frac{\eta}{\|\nabla \mathbf{W}_m \mathcal{L}^{(t)}\|_F} \cdot \frac{3\mathbb{E}[\alpha^3] + \widetilde{O}(M^2\sigma_0^{0.5})}{2KM^2P} \langle \mathbf{w}_{m,j,p}^{(t)}, \mathbf{v}_{k_m^*} \rangle^2, \\
\langle \mathbf{w}_{m,j,p}^{(t+1)}, \mathbf{c}_k \rangle &\leq \langle \mathbf{w}_{m,j,p}^{(t)}, \mathbf{c}_k \rangle + \frac{\eta}{\|\nabla \mathbf{W}_m \mathcal{L}^{(t)}\|_F} \cdot \widetilde{O}(\sigma_0^{0.5}) \langle \mathbf{w}_{m,j,p}^{(t)}, \mathbf{c}_k \rangle^2.
\end{aligned}$$

Applying Lemma 27 with $C_t = (3\mathbb{E}[\alpha^3] + \widetilde{O}(\sigma_0^{0.5}))/(2KM^2P\|\nabla \mathbf{W}_m \mathcal{L}^{(t)}\|_F)$, $S = \widetilde{O}(M^2\sigma_0^{0.5})$, $G = 2$, and using the initialization event $\langle \mathbf{w}_m^{(0)}, \mathbf{v}_{k_m}^* \rangle \geq S(1+G^{-1})\langle \mathbf{w}_m^{(0)}, \mathbf{c}_k \rangle$, we obtain

$$\langle \mathbf{w}_{m,j,p}^{(t)}, \mathbf{c}_k \rangle \leq O(G^{-1}\sigma_0) = \widetilde{O}(\sigma_0).$$

This conclusion is due to Lemma 27 with the above parameter choices.

Comparison with $\langle \mathbf{w}_{m,j,p}^{(t)}, \xi_{i,q} \rangle$. The same argument as for $\mathbf{c}_k$ applies, yielding $\langle \mathbf{w}_{m,j,p}^{(t)}, \xi_{i,q} \rangle = \widetilde{O}(\sigma_0)$. □

Denote by $T^{(m)}$ the first iteration for which $\|\nabla \mathbf{W}_m \mathcal{L}^{(t)}\|_F \geq M^2 \sigma_0^{1.8}$. The next lemma upper bounds $T^{(m)}$ for all $m \in \mathcal{M}$.

Lemma 14. *For all $m \in [M]$, we have that $T^{(m)} = \widetilde{O}(\eta^{-1} M^4 \sigma_0^{0.8})$ and thus $T^{(m)} \leq T_1$ only when $M = \widetilde{O}(\sigma_0^{-0.075})$. Besides, for all $T_m < t \leq T_1$ we have that*

$$\langle \nabla_{\mathbf{w}_{m,j_m^*,p}} \mathcal{L}^{(t)}, \mathbf{v}_{k_m}^* \rangle \geq (1 - \sigma_0^{0.2}) \|\nabla \mathbf{W}_m \mathcal{L}^{(t)}\|_F.$$

Proof. Let the projection matrix be $B = \mathbf{v}_{k_m}^* (\mathbf{v}_{k_m}^*)^\top \in \mathbb{R}^{d \times d}$. We decompose the gradient into two orthogonal components:

$$\|\nabla_{\mathbf{w}_{m,j_m^*,p}} \mathcal{L}^{(t)}\|_2 = \|B \nabla_{\mathbf{w}_{m,j_m^*,p}} \mathcal{L}^{(t)} + (I - B) \nabla_{\mathbf{w}_{m,j_m^*,p}} \mathcal{L}^{(t)}\|_2 \leq \|B \nabla_{\mathbf{w}_{m,j_m^*,p}} \mathcal{L}^{(t)}\|_2 + \|(I - B) \nabla_{\mathbf{w}_{m,j_m^*,p}} \mathcal{L}^{(t)}\|_2.$$

The first equality is due to adding and subtracting the orthogonal projection on $\mathbf{v}_{k_m}^*$, and the second inequality is due to the triangle inequality.

Recall that

$$\nabla_{\mathbf{w}_{m,j_m^*,p}} \mathcal{L}^{(t)} = \frac{1}{n} \sum_i \mathbb{1}(m_{i,t} = m) \ell'_{i,t} \pi_m(\mathbf{x}_i; \Theta^{(t)}) y_i \sigma'(\langle \mathbf{w}_{m,j_m^*,p}^{(t)}, \mathbf{x}_i^{(p)} \rangle) \mathbf{x}_i^{(p)}.$$

Hence

$$\begin{aligned}
\|(I - B) \nabla_{\mathbf{w}_{m,j_m^*,p}} \mathcal{L}^{(t)}\|_2 &= \left\| \frac{1}{n} \sum_i \mathbb{1}(m_{i,t} = m) \ell'_{i,t} \pi_m(\mathbf{x}_i; \Theta^{(t)}) y_i \sigma'(\langle \mathbf{w}_{m,j_m^*,p}^{(t)}, \mathbf{x}_i^{(p)} \rangle)(I - B) \mathbf{x}_i^{(p)} \right\|_2 \\
&\leq \frac{1}{n} \sum_i \left\| \sigma'(\langle \mathbf{w}_{m,j_m^*,p}^{(t)}, \mathbf{x}_i^{(p)} \rangle)(I - B) \mathbf{x}_i^{(p)} \right\|_2 \\
&\leq \widetilde{O}(M^2 \sigma_0^2).
\end{aligned}$$

The first inequality is due to $|\ell'_{i,t}| \leq 1$ and $\pi_m \in [0, 1]$. The second inequality is due to the following two cases: when $\mathbf{x}_i^{(p)}$ aligns with $\mathbf{v}_{k_m}^*$, then $(I - B)\mathbf{x}_i^{(p)} = 0$; when $\mathbf{x}_i^{(p)}$ does not align with $\mathbf{v}_{k_m}^*$, then $\langle \mathbf{w}_{m,j_m^*,p}^{(t)}, \mathbf{x}_i^{(p)} \rangle = \widetilde{O}(M \sigma_0)$ by Lemma 13.

Therefore

$$\|\nabla_{\mathbf{w}_{m,j_m^*,p}} \mathcal{L}^{(t)}\|_2 \leq \|B \nabla_{\mathbf{w}_{m,j_m^*,p}} \mathcal{L}^{(t)}\|_2 + \widetilde{O}(M^2 \sigma_0^2) = \langle \nabla_{\mathbf{w}_{m,j_m^*,p}} \mathcal{L}^{(t)}, \mathbf{v}_{k_m}^* \rangle + \widetilde{O}(M^2 \sigma_0^2). \tag{12}$$

The first equality is due to the definition of B, and the second equality is due to $B\mathbf{u} = \langle \mathbf{u}, \mathbf{v}_{k_m}^* \rangle \mathbf{v}_{k_m}^*$ and $\|\mathbf{v}_{k_m}^*\|_2 = 1$.

For neurons $j \neq j_m^*$, we have

$$\|\nabla_{\mathbf{w}_{m,j,p}} \mathcal{L}^{(t)}\|_2 = \left\| \frac{1}{n} \sum_i \mathbb{1}(m_{i,t} = m) \ell'_{i,t} \pi_m(\mathbf{x}_i; \Theta^{(t)}) y_i \sigma'(\langle \mathbf{w}_{m,j,p}^{(t)}, \mathbf{x}_i^{(p)} \rangle) \mathbf{x}_i^{(p)} \right\|_2 = \widetilde{O}(M^2 \sigma_0^2), \tag{13}$$

where the equality is due to $\langle \mathbf{w}_{m,j,p}^{(t)}, \mathbf{x}_i^{(p)} \rangle = \widetilde{O}(\sigma_0)$ for all $j \neq j_m^*$ by Lemma 13.

Thus the gradient norm satisfies

$$\|\nabla \mathbf{W}_m \mathcal{L}^{(t)}\|_F \leq \sum_{j \in [J]} \|\nabla_{\mathbf{w}_{m,j,p}} \mathcal{L}^{(t)}\|_2 \leq \|\nabla_{\mathbf{w}_{m,j_m^*,p}} \mathcal{L}^{(t)}\|_2 + \widetilde{O}(M^2 \sigma_0^2). \tag{14}$$

The first inequality is due to the triangle inequality on the Frobenius norm, and the second inequality is due to (13).

Whenever $\|\nabla \mathbf{W}_m \mathcal{L}^{(t)}\|_F \geq M^2 \sigma_0^{1.8}$, we obtain

$$\langle \nabla_{\mathbf{w}_{m,j_m^*,p}} \mathcal{L}^{(t)}, \mathbf{v}_{k_m}^* \rangle \geq \|\nabla_{\mathbf{w}_{m,j_m^*,p}} \mathcal{L}^{(t)}\|_2 - \widetilde{O}(M^2 \sigma_0^2) \geq \|\nabla \mathbf{W}_m \mathcal{L}^{(t)}\|_F - \widetilde{O}(M^2 \sigma_0^2) \geq (1 - \sigma_0^{0.2}) \|\nabla \mathbf{W}_m \mathcal{L}^{(t)}\|_F,$$

where the first inequality is due to (12), the second inequality is due to (14), and the third inequality is due to $M^2 \sigma_0^2 \leq \sigma_0^{0.2} \|\nabla \mathbf{W}_m \mathcal{L}^{(t)}\|_F$.

We now upper bound $T^{(m)}$. While $t \leq T^{(m)}$, we have $\|\nabla \mathbf{W}_m \mathcal{L}^{(t)}\|_F < M^2 \sigma_0^{1.8}$. On the one hand, Lemma 12 gives

$$\|\nabla \mathbf{W}_m \mathcal{L}^{(t)}\|_2 \geq -\langle \nabla_{\mathbf{w}_{m,j,p}} \mathcal{L}^{(t)}, \mathbf{v}_{k_m}^* \rangle - \widetilde{O}(M^2 \sigma_0^2) = \frac{3\mathbb{E}[\alpha^3] - \widetilde{O}(\sigma_0^{1.5})}{2KM^2P} \langle \mathbf{w}_{m,j_m^*,p}^{(t)}, \mathbf{v}_{k_m}^* \rangle^2 - \widetilde{O}(M^2 \sigma_0^2),$$

which implies $\langle \mathbf{w}_{m,j_m^*,p}^{(t)}, \mathbf{v}_{k_m}^* \rangle \leq \widetilde{O}(M^2 \sigma_0^{0.9})$.

On the other hand, by Lemma 13,

$$\begin{aligned}\langle \mathbf{w}_{m,j_m^*,p}^{(t+1)}, \mathbf{v}_{k_m}^* \rangle &\geq \langle \mathbf{w}_{m,j_m^*,p}^{(t)}, \mathbf{v}_{k_m}^* \rangle + \frac{\eta}{\|\nabla \mathbf{W}_m \mathcal{L}^{(t)}\|_F} \Theta\left(\frac{1}{KM^2P}\right) \langle \mathbf{w}_{m,j_m^*,p}^{(t)}, \mathbf{v}_{k_m}^* \rangle^2 \\ &\geq \langle \mathbf{w}_{m,j_m^*,p}^{(t)}, \mathbf{v}_{k_m}^* \rangle + \Theta\left(\frac{\eta}{KM^4P \cdot \sigma_0^{1.8}}\right) \langle \mathbf{w}_{m,j_m^*,p}^{(t)}, \mathbf{v}_{k_m}^* \rangle^2 \\ &\geq \langle \mathbf{w}_{m,j_m^*,p}^{(t)}, \mathbf{v}_{k_m}^* \rangle + \Theta\left(\frac{\eta}{KM^4P \cdot \sigma_0^{0.8}}\right) \langle \mathbf{w}_{m,j_m^*,p}^{(t)}, \mathbf{v}_{k_m}^* \rangle.\end{aligned}$$

The first inequality is due to substituting the expectation bound from Lemma 13. The second equality is due to replacing $\|\nabla \mathbf{W}_m \mathcal{L}^{(t)}\|_F$ by its upper bound $M^2\sigma_0^{1.8}$. The third inequality is due to $\langle \mathbf{w}_{m,j_m^*,p}^{(t)}, \mathbf{v}_{k_m}^* \rangle \geq 0.01\sigma_0$, by Lemma 6.

Hence the inner product grows exponentially and reaches $\widetilde{O}(M^2\sigma_0^{0.9})$ within $\widetilde{O}(\eta^{-1}M^4\sigma_0^{0.8})$ iterations, which proves $T^{(m)} = \widetilde{O}(\eta^{-1}M^4\sigma_0^{0.8})$. Therefore $T^{(m)} \leq T_1$ only when $M = \widetilde{O}(\sigma_0^{-0.075})$. Finally, the alignment inequality

$$\langle \nabla_{\mathbf{w}_{m,j_m^*,p}} \mathcal{L}^{(t)}, \mathbf{v}_{k_m}^* \rangle \geq (1 - \sigma_0^{0.2}) \|\nabla \mathbf{W}_m \mathcal{L}^{(t)}\|_F$$

follows from combining the previous bounds on the projected and residual components. □

Recall that $T_1 = \lfloor \eta^{-1}\sigma_0^{0.5} \rfloor$. It follows that each expert $m \in [M]$ learns to specialize in a single cluster during the first stage.

Lemma 15. *For all $t \leq T_1, m \in [M]$, we have*

$$\langle \mathbf{w}_{m,j_m^*,p}^{(t)}, \mathbf{v}_{k_m}^* \rangle = O(\sigma_0^{0.5}), \qquad \langle \mathbf{w}_{m,j,p}^{(t)}, \mathbf{v}_k \rangle = \widetilde{O}(M\sigma_0),\ \forall (j,k) \neq (j_m^*, k_m^*),$$

$$\langle \mathbf{w}_{m,j,p}^{(t)}, \mathbf{c}_k \rangle = \widetilde{O}(\sigma_0),\ \forall j \in [J], k \in [K],$$

$$\langle \mathbf{w}_{m,j,p}^{(t)}, \xi_{i,q} \rangle = \widetilde{O}(\sigma_0),\ \forall j \in [J], i \in [n], q \in Q_i.$$

Moreover,

$$\langle \mathbf{w}_{m,j_m^*,p}^{(t)}, \mathbf{v}_{k_m}^* \rangle \geq (1 - \sigma_0^{0.2})\eta(t - T^{(m)}), \quad \forall t \geq T^{(m)}.$$

Proof. By Lemma 14, $T^{(m)} = \widetilde{O}(\eta^{-1}M^4\sigma_0^{0.8}) < T_1$. For all $T^{(m)} \leq t \leq T_1$,

$$\langle \nabla_{\mathbf{w}_{m,j_m^*,p}} \mathcal{L}^{(t)}, \mathbf{v}_{k_m}^* \rangle \geq (1 - \sigma_0^{0.2}) \|\nabla \mathbf{W}_m \mathcal{L}^{(t)}\|_F.$$

Using the normalized update gives

$$\langle \mathbf{w}_{m,j_m^*,p}^{(t+1)}, \mathbf{v}_{k_m}^* \rangle \geq \langle \mathbf{w}_{m,j_m^*,p}^{(t)}, \mathbf{v}_{k_m}^* \rangle + (1 - \sigma_0^{0.2})\eta, \quad \forall T^{(m)} \leq t \leq T_1.$$

Summing these increments from $T^{(m)}$ to $t-1$ yields

$$\langle \mathbf{w}_{m,j_m^*,p}^{(t)}, \mathbf{v}_{k_m}^* \rangle \geq (1 - \sigma_0^{0.2})\,\eta\,(t - T^{(m)}), \quad \forall\, t \geq T^{(m)}.$$

Finally, the remaining bounds on $\langle \mathbf{w}_{m,j,p}^{(t)}, \mathbf{v}_k \rangle$, $\langle \mathbf{w}_{m,j,p}^{(t)}, \mathbf{c}_k \rangle$, and $\langle \mathbf{w}_{m,j,p}^{(t)}, \xi_{i,q} \rangle$ follow directly from Lemma 13. □

Router Learning Stage

Denote $T_2 \coloneqq \lfloor \eta^{-1}M^{-2} \rfloor$. The second stage ends at time $t = T_2$. Given an example $\mathbf{x}$, define $\overline{\mathbf{x}}$ as the vector that retains only the cluster–center signal, i.e., $\overline{\mathbf{x}}^{(p)} = \mathbf{x}^{(p)}$ when $p = p_3$ and $\overline{\mathbf{x}}^{(p)} = 0$ otherwise. Similarly, define $\hat{\mathbf{x}}$ as the vector that retains only the feature signal and feature noise, i.e., $\hat{\mathbf{x}}^{(p)} = \mathbf{x}^{(p)}$ when $p \in \{p_1, p_2\}$ and $\hat{\mathbf{x}}^{(p)} = 0$ otherwise.

For all $T_1 \leq t \leq T_2$, we show that the router concentrates on the cluster–center signals while the experts concentrate on the feature signals; specifically, we prove that

$$\left| f_m\big(\mathbf{x}_i;\, \mathbf{W}^{(t)}\big) - f_m\big(\hat{\mathbf{x}}_i;\, \mathbf{W}^{(t)}\big) \right| \quad \text{and} \quad \left\| h\big(\mathbf{x}_i;\, \mathbf{\Theta}^{(t)}\big) - h\big(\overline{\mathbf{x}}_i;\, \mathbf{\Theta}^{(t)}\big) \right\|_\infty$$

are small. In particular, we claim that for all $T_1 \leq t \leq T_2$, the following proposition holds.

Proposition 3. *For all $T_1 \leq t \leq T_2$, the following inequalities hold:*

$$\left| f_m(\mathbf{x}_i; \mathbf{W}^{(t)}) - f_m(\widehat{\mathbf{x}}_i; \mathbf{W}^{(t)}) \right| = \widetilde{O}\big(\sigma_0^3\big), \qquad \forall\, m \in [M],\ i \in [n], \tag{15}$$

$$\left\| h(\mathbf{x}_i; \mathbf{\Theta}^{(t)}) - h(\overline{\mathbf{x}}_i; \mathbf{\Theta}^{(t)}) \right\|_\infty \leq \widetilde{O}\big(\sigma_0^{2.5} M^2 P^2 K^2\big), \qquad \forall\, i \in [n], \tag{16}$$

$$\mathbb{P}(m_{i,t} = m),\ \pi_m(\mathbf{x}_i; \mathbf{\Theta}^{(t)}) = \Omega(1/M), \qquad \forall\, m \in [M],\ i \in \Omega_{k_m^*}. \tag{17}$$

Proposition 3 shows that, throughout the second stage, the experts attend only to the label signal, whereas the router attends only to the cluster–center signal. We prove Proposition 3 by induction on t. Before presenting the main argument, we establish several auxiliary lemmas. In our analysis, the dependence on the number of experts M is made explicit, and we incorporate the constraint $M = \widetilde{O}\big(\sigma_0^{-0.075}\big)$ from Lemma 14 when deriving Proposition 3.

Lemma 16. *For all $T_1 \le t \le T_2$, the neural network parameters satisfy:*

- $|f_m(\mathbf{x}_i; \mathbf{W}^{(t)})| = O(JP),\ \forall m \in [M]$,
- $\pi_{m_i,t}(\mathbf{x}_i; \mathbf{\Theta}^{(t)}) = \Omega(1/M),\ \forall i \in [n]$.

Proof. Because we use normalized gradient descent, the first claim follows directly:

$$|f_m(\mathbf{x}_i, \mathbf{W}^{(t)})| = \sum_{j\in[J]} \sum_{p\in[P]} \sigma\big(\langle \mathbf{w}_{m,j,p}^{(t)}, \mathbf{x}_i^{(p)} \rangle\big) \overset{(i)}{=} O(JP),$$

where the displayed equality is labeled (i). The equality (i) is due to the fact that $\|\mathbf{w}_{m,j,p}^{(t)} - \mathbf{w}_{m,j,p}^{(0)}\|_2 = O(\eta T_2) = O(M^{-2})$ while each patch vector $\mathbf{x}_i^{(p)}$ has bounded norm, and the activation $\sigma(\cdot)$ is bounded, so the double sum over J filters and P patches is $O(JP)$.

We now prove the second claim. By Lemma 26,

$$h_{m_i,t}(\mathbf{x}; \mathbf{\Theta}) \ge \max_m h_m(\mathbf{x}; \mathbf{\Theta}) - 1,$$

which yields

$$\pi_{m_i,t}(\mathbf{x}_i; \mathbf{\Theta}^{(t)}) = \frac{\exp(h_{m_i,t}(\mathbf{x}_i; \mathbf{\Theta}^{(t)}))}{\sum_m \exp(h_m(\mathbf{x}; \mathbf{\Theta}^{(t)}))} \ge \frac{\exp(h_{m_i,t}(\mathbf{x}_i; \mathbf{\Theta}^{(t)}))}{M \max_m \exp(h_m(\mathbf{x}; \mathbf{\Theta}^{(t)}))} \ge \frac{1}{eM}.$$

The first equality is due to the definition of the softmax routing probability. The second inequality is due to Lemma 26, which gives $\exp(h_{m_i,t}) \ge e^{-1} \max_m \exp(h_m)$. This proves $\pi_{m_i,t}(\mathbf{x}_i; \mathbf{\Theta}^{(t)}) = \Omega(1/M)$. □

Lemma 17. *Denote $\delta_{\mathbf{\Theta}} = \max_i \|h(\overline{\mathbf{x}}_i; \mathbf{\Theta}) - h(\mathbf{x}_i; \mathbf{\Theta})\|_\infty$ and let the random variable $\overline{m}_{i,t}$ be the expert that would be routed if we used the gating output $h(\overline{\mathbf{x}}_i; \mathbf{\Theta}^{(t)})$ instead. Then*

$$\big|\pi_m(\mathbf{x}_i; \mathbf{\Theta}) - \pi_m(\overline{\mathbf{x}}_i; \mathbf{\Theta})\big| = O(\delta_{\mathbf{\Theta}}),\ \forall\, m \in [M],\ i \in [n],$$

and

$$\big|\mathbb{P}(m_{i,t} = m) - \mathbb{P}(\overline{m}_{i,t} = m)\big| = O(M^2 \delta_{\mathbf{\Theta}}),\ \forall\, m \in [M],\ i \in [n].$$

Proof. By the definition of $\delta_{\mathbf{\Theta}}$, we have $\|h(\mathbf{x}_i; \mathbf{\Theta}^{(t)}) - h(\overline{\mathbf{x}}_i; \mathbf{\Theta}^{(t)})\|_\infty \le \delta_{\mathbf{\Theta}}$. Applying Lemma 1 yields $\big|\mathbb{P}(m_{i,t} = m) - \mathbb{P}(\overline{m}_{k,t} = m)\big| = \widetilde{O}(\delta_{\mathbf{\Theta}})$ for all $m \in [M]$ and $i \in [n]$, which completes the proof for the second equation.

We now prove the first equation. For all $i \in [n]$,

$$\pi_m(\mathbf{x}_i; \mathbf{\Theta}) = \frac{\pi_m(\overline{\mathbf{x}}_i; \mathbf{\Theta}) \exp\big(h_m(\mathbf{x}_i; \mathbf{\Theta}) - h_m(\overline{\mathbf{x}}_i; \mathbf{\Theta})\big)}{\sum_{m'} \pi_{m'}(\overline{\mathbf{x}}_i; \mathbf{\Theta}) \exp\big(h_{m'}(\mathbf{x}_i; \mathbf{\Theta}) - h_{m'}(\overline{\mathbf{x}}_i; \mathbf{\Theta})\big)}.$$

Let $\delta_{m'} \coloneqq \exp\big(h_{m'}(\mathbf{x}_i; \mathbf{\Theta}) - h_{m'}(\overline{\mathbf{x}}_i; \mathbf{\Theta})\big) = 1 + O(\delta_{\mathbf{\Theta}})$. For sufficiently small $\delta_{\mathbf{\Theta}}$, we have $\delta_{m'} \ge 0.5$. Then

$$\begin{aligned}
\big|\pi_m(\mathbf{x}_i; \mathbf{\Theta}^{(t)}) - \pi_m(\overline{\mathbf{x}}_i; \mathbf{\Theta})\big| &= \pi_m(\overline{\mathbf{x}}_i; \mathbf{\Theta}) \left| \frac{\delta_m}{\sum_{m'} \pi_{m'}(\overline{\mathbf{x}}_i; \mathbf{\Theta})\, \delta_{m'}} - 1 \right| \\
&= \pi_m(\overline{\mathbf{x}}_i; \mathbf{\Theta}) \frac{|\sum_{m'} \pi_{m'}(\overline{\mathbf{x}}_i; \mathbf{\Theta})(\delta_m - \delta_{m'})|}{\sum_{m'} \pi_{m'}(\overline{\mathbf{x}}_i; \mathbf{\Theta})\, \delta_{m'}} \\
&\le \pi_m(\overline{\mathbf{x}}_i; \mathbf{\Theta}) \frac{\sum_{m'} \pi_{m'}(\overline{\mathbf{x}}_i; \mathbf{\Theta})\, |\delta_m - \delta_{m'}|}{\sum_{m'} \pi_{m'}(\overline{\mathbf{x}}_i; \mathbf{\Theta})\, \delta_{m'}} \\
&\le O(\delta_{\mathbf{\Theta}}),
\end{aligned}$$

where the last inequality is due to $|\delta_{m'} - \delta_m| \le O(\delta_{\mathbf{\Theta}})$, $\pi_m(\overline{\mathbf{x}}_i; \mathbf{\Theta}) \le 1$, and $\sum_{m'} \pi_{m'}(\overline{\mathbf{x}}_i; \mathbf{\Theta})\, \delta_{m'} \ge \big[\sum_{m'} \pi_{m'}(\overline{\mathbf{x}}_i; \mathbf{\Theta})\big]/2 = 0.5$. □

Lemma 18. *Suppose* (15)*,* (16)*,* (17) *hold for all* $t \in [T_1, T] \subseteq [T_1, T_2 - 1]$*, then we have the following inequalities hold for all* $t \in [T_1, T + 1]$*,*

$$\begin{aligned}
\langle \mathbf{w}_{m,j_m^*,p}^{(t)}, \mathbf{v}_{k_m^*} \rangle &\geq (1 - O(\sigma_0^{0.1}))\eta t, \\
\langle \mathbf{w}_{m,j,p}^{(t)}, \mathbf{v}_k \rangle &= \widetilde{O}(\sigma_0),\ \forall (j,k) \neq (j_m^*, k_m^*), \\
\langle \mathbf{w}_{m,j,p}^{(t)}, \mathbf{c}_k \rangle &= \widetilde{O}(\sigma_0),\ \forall j \in [J], k \in [K], \\
\langle \mathbf{w}_{m,j,p}^{(t)}, \boldsymbol{\xi}_{i,q} \rangle &= \widetilde{O}(\sigma_0),\ \forall j \in [J], k \in [K], i \in [n], q \in Q_i.
\end{aligned}$$

Proof. The proof is similar to Lemma 13. First, we recalculate the bound on $\ell_{i,t}'$ for $t \in [T_1, T_2]$. Let $s = y_i \pi_{m_{i,t}}(\mathbf{x}_i; \Theta^{(t)}) f_{m_{i,t}}(\mathbf{x}_i; \mathbf{W}^{(t)})$. By Lemma 16 we have $s = O(JP)$, hence $\ell_{i,t}' = -\frac{1}{e^s+1} = -\Theta(1)$. Recall that the inner product against the feature signal $\langle \nabla_{\mathbf{w}_{m,j,p}} \mathcal{L}^{(t)}, \mathbf{v}_k \rangle$ is computed as follows

$$\begin{aligned}
&\mathbb{E}[\langle \nabla_{\mathbf{w}_{m,j,p}} \mathcal{L}^{(t)}, \mathbf{v}_k \rangle] \\
&= \frac{1}{n} \sum_{i \in [n]} \mathbb{P}(m_{i,t} = m) \mathbb{E}\left[\ell_{i,t}' \pi_m(\mathbf{x}_i; \boldsymbol{\Theta}^{(t)}) \sigma'\Big(\langle \mathbf{w}_{m,j,p}^{(t)}, \mathbf{x}_i^{(p)} \rangle \Big) y_i \langle \mathbf{x}_i^{(p)}, \mathbf{v}_k \rangle \right] \\
&= \underbrace{\frac{1}{nP} \sum_{i \in \Omega_k} \mathbb{P}(m_{i,t} = m) \mathbb{E}\left[\ell_{i,t}' \pi_m(\mathbf{x}_i; \boldsymbol{\Theta}^{(t)}) \sigma'\Big(\langle \mathbf{w}_{m,j,p}^{(t)}, \mathbf{v}_k \rangle \Big) \alpha_i^3 \|\mathbf{v}_k\|_2^2 \;\middle|\; \mathbf{x}_i^{(p)} = y_i \alpha_i \mathbf{v}_k \right]}_{I_1} \\
&\quad + \underbrace{\frac{1}{nP} \sum_{k' \neq k} \sum_{i \in \Omega_{k',k}} \mathbb{P}(m_{i,t} = m) \mathbb{E}\left[\ell_{i,t}' \pi_m(\mathbf{x}_i; \boldsymbol{\Theta}^{(t)}) \sigma'\Big(\langle \mathbf{w}_{m,j,p}^{(t)}, \mathbf{v}_k \rangle \Big) \gamma_i^3 y_i \varepsilon_i \|\mathbf{v}_k\|_2^2 \;\middle|\; \mathbf{x}_i^{(p)} = \varepsilon_i \gamma_i \mathbf{v}_k \right]}_{I_2} \\
&\quad + \underbrace{\frac{1}{nP} \sum_{i \in [n],\, q \in Q_i} \mathbb{P}(m_{i,t} = m) \mathbb{E}\left[\ell_{i,t}' \pi_m(\mathbf{x}_i; \boldsymbol{\Theta}^{(t)}) \sigma'\Big(\langle \mathbf{w}_{m,j,p}^{(t)}, \xi_{i,q} \rangle \Big) y_i \langle \mathbf{v}_k, \xi_{i,q} \rangle \;\middle|\; \mathbf{x}_i^{(p)} = \xi_{i,q} \right]}_{I_3}.
\end{aligned}$$

where the second equality is by taking expectation over random patch permutations. The new bound on $\ell_{i,t}'$ gives us a bound on I_1.

$$\begin{aligned}
I_1 &= \frac{1}{nP} \sum_{i \in \Omega_k} \mathbb{P}(m_{i,t} = m) \mathbb{E}\left[\ell_{i,t}' \pi_m(\mathbf{x}_i; \boldsymbol{\Theta}^{(t)}) \sigma'\Big(\langle \mathbf{w}_{m,j,p}^{(t)}, \mathbf{v}_k \rangle \Big) \alpha_i^3 \|\mathbf{v}_k\|_2^2 \;\middle|\; \mathbf{x}_i^{(p)} = y_i \alpha_i \mathbf{v}_k \right] \\
&= -O\left(\frac{1}{nP} \right) \sum_{i \in \Omega_k} \mathbb{P}(m_{i,t} = m) \alpha_i^3 \sigma'\Big(\langle \mathbf{w}_{m,j,p}^{(t)}, \mathbf{v}_k \rangle \Big) \\
&= -O\left(\frac{1}{KP} \right) \mathbb{E}[\alpha_i^3] \sigma'\Big(\langle \mathbf{w}_{m,j,p}^{(t)}, \mathbf{v}_k \rangle \Big),
\end{aligned} \tag{18}$$

where the last equality is due to Lemma 4. Similarly, we can bound I_2 as follows

$$\begin{aligned}
I_2 &= -O\left(\frac{1}{nP} \right) \sum_{k' \neq k} \sum_{i \in \Omega_{k',k}} \mathbb{P}(m_{i,t} = m) \gamma_i^3 y_i \epsilon_i \sigma'\Big(\langle \mathbf{w}_{m,j,p}^{(t)}, \mathbf{v}_k \rangle \Big) \\
&= -\widetilde{O}(\sigma_0^{1.5}) \sigma'\Big(\langle \mathbf{w}_{m,j,p}^{(t)}, \mathbf{v}_k \rangle \Big)
\end{aligned} \tag{19}$$

Finally, we can bound I_3 as follows

$$\begin{aligned}
I_3 &= -O\left(\frac{1}{nP} \right) \sum_{i \in [n], q \in Q_i} \mathbb{P}(m_{i,t} = m) y_i \sigma'\Big(\langle \mathbf{w}_{m,j,p}^{(t)}, \xi_{i,q} \rangle \Big) \langle \mathbf{v}_k, \xi_{i,q} \rangle \\
&= -\widetilde{O}\left(\frac{d^{-1/2}}{nP} \right) \sum_{i \in [n], q \in Q_i} \mathbb{P}(m_{i,t} = m) y_i \\
&= -\widetilde{O}(d^{-1/2} n^{-1/2}) \\
&= -\widetilde{O}(\sigma_0^3),
\end{aligned} \tag{20}$$

where the second equality is due to $\langle \mathbf{w}_{m,j,p}^{(t)}, \xi_{i,q} \rangle = O(\eta t) = O(1)$ and Lemma 5, the third inequality is due to Lemma 4, the last equality is due to choice of n, d.

Moreover, when $k = k_m^*$, by (17) we can give an lower bound on I_1 as follows

$$\begin{aligned} I_1 &= -\Omega\left(\frac{1}{nMP}\right) \sum_{i \in \Omega_k} \mathbb{P}(m_{i,t} = m) \alpha_i^3 \sigma'\Big(\langle \mathbf{w}_{m,j,p}^{(t)}, \mathbf{v}_{k_m^*} \rangle\Big) \\ &= -O\left(\frac{1}{KM^2P}\right) \mathbb{E}[\alpha_i^3] \sigma'\Big(\langle \mathbf{w}_{m,j,p}^{(t)}, \mathbf{v}_{k_m^*} \rangle\Big). \end{aligned} \tag{21}$$

Combining the bounds on I_1, I_2, I_3 gives us

$$\begin{aligned} \mathbb{E}[\langle \nabla_{\mathbf{w}_{m,j,p}} \mathcal{L}^{(t)}, \mathbf{v}_k \rangle] &= -O\left(\frac{1}{KP}\right) \Big[\mathbb{E}[\alpha_i^3] + \widetilde{O}(\sigma_0^{1.5})\Big] \sigma'\Big(\langle \mathbf{w}_{m,j,p}^{(t)}, \mathbf{v}_k \rangle\Big) + \widetilde{O}(\sigma_0^3). \\ \mathbb{E}[\langle \nabla_{\mathbf{w}_{m,j,p}} \mathcal{L}^{(t)}, \mathbf{v}_{k_m^*} \rangle] &= -\Omega\left(\frac{1}{KM^2P}\right) \Big[\mathbb{E}[\alpha_i^3] + \widetilde{O}(\sigma_0^{1.5})\Big] \sigma'\Big(\langle \mathbf{w}_{m,j,p}^{(t)}, \mathbf{v}_{k_m^*} \rangle\Big) + \widetilde{O}(\sigma_0^3). \end{aligned} \tag{22}$$

Therefore, the update rule for every expert can be written as follows,

$$\begin{aligned} \langle \mathbf{w}_{m,j,p}^{(t+1)}, \mathbf{v}_k \rangle &= \langle \mathbf{w}_{m,j,p}^{(t)}, \mathbf{v}_k \rangle + \frac{\eta}{\|\nabla \mathbf{W}_m \mathcal{L}^{(t)}\|_F} \left[O\left(\frac{3\mathbb{E}[\alpha^3] + \widetilde{O}(\sigma_0^{1.5})}{KP}\right) \langle \mathbf{w}_{m,j,p}^{(t)}, \mathbf{v}_k \rangle^2 + \widetilde{O}(\sigma_0^3) \right], \\ \langle \mathbf{w}_{m,j,p}^{(t+1)}, \mathbf{v}_{k_m^*} \rangle &= \langle \mathbf{w}_{m,j,p}^{(t)}, \mathbf{v}_{k_m^*} \rangle + \frac{\eta}{\|\nabla \mathbf{W}_m \mathcal{L}^{(t)}\|_F} \left[\Omega\left(\frac{3\mathbb{E}[\alpha^3] + \widetilde{O}(\sigma_0^{1.5})}{KM^2P}\right) \langle \mathbf{w}_{m,j,p}^{(t)}, \mathbf{v}_{k_m^*} \rangle^2 + \widetilde{O}(\sigma_0^3) \right], \end{aligned} \tag{23}$$

Applying Lemma 27 by choosing $C_t = \big(3\mathbb{E}[\alpha^3] + \widetilde{O}(\sigma_0^{1.5})\big) / \big(2KM^2P \|\nabla \mathbf{W}_m \mathcal{L}^{(t)}\|_F\big)$, $S = \Theta(M^2)$, $G = 1$, and verifying that $\langle \mathbf{w}_m^{(T_1)}, \mathbf{v}_{k_m}^* \rangle \geq S(1+G) \langle \mathbf{w}_m^{(T_1)}, \mathbf{v}_k \rangle$ (which holds due to Lemma 15 and the choice of M, we have that

$$\langle \mathbf{w}_{m,j,p}^{(t)}, \mathbf{v}_k \rangle \leq O(G^{-1} M \sigma_0) = \widetilde{O}(M \sigma_0).$$

By the same arguments, we can compute the update rule w.r.t cluster-center signal and random noise as follows,

$$\begin{aligned} \langle \mathbf{w}_{m,j,p}^{(t+1)}, \xi_{i,q} \rangle &= \langle \mathbf{w}_{m,j,p}^{(t)}, \xi_{i,q} \rangle + \frac{\eta}{\|\nabla \mathbf{W}_m \mathcal{L}^{(t)}\|_F} \Big[\widetilde{O}(\sigma_0^{1.5}) \langle \mathbf{w}_{m,j,p}^{(t)}, \xi_{i,q} \rangle^2 + \widetilde{O}(\sigma_0^3) \Big], \\ \langle \mathbf{w}_{m,j,p}^{(t+1)}, \mathbf{c}_k \rangle &= \langle \mathbf{w}_{m,j,p}^{(t)}, \mathbf{c}_k \rangle + \frac{\eta}{\|\nabla \mathbf{W}_m \mathcal{L}^{(t)}\|_F} \Big[\widetilde{O}(\sigma_0^{1.5}) \langle \mathbf{w}_{m,j,p}^{(t)}, \mathbf{c}_k \rangle^2 + \widetilde{O}(\sigma_0^3) \Big]. \end{aligned} \tag{24}$$

Again, applying Lemma 27 by choosing $C_t = \big(3\mathbb{E}[\alpha^3] + \widetilde{O}(\sigma_0^{1.5})\big) / \big(2KM^2P \|\nabla \mathbf{W}_m \mathcal{L}^{(t)}\|_F\big)$, $S = \Theta(1)$, $G = 1$, and verifying that $\langle \mathbf{w}_m^{(T_1)}, \mathbf{v}_{k_m}^* \rangle \geq S(1+G) \langle \mathbf{w}_m^{(T_1)}, \mathbf{v}_k \rangle$ (which holds due to Lemma 15 and the choice of M, we have that

$$\begin{aligned} \langle \mathbf{w}_{m,j,p}^{(t)}, \mathbf{c}_k \rangle &\leq O(G^{-1} \sigma_0) = \widetilde{O}(\sigma_0). \\ \langle \mathbf{w}_{m,j,p}^{(t)}, \xi_{i,q} \rangle &\leq O(G^{-1} \sigma_0) = \widetilde{O}(\sigma_0). \end{aligned} \tag{25}$$

Following the same arguments in Lemma 14 and Lemma 15 completes our proof. □

Lemma 19. *Suppose* (15)*,* (16)*, and* (17) *hold for all* $t \in [T_1, T] \subseteq [T_1, T_2 - 1]$*. Then*

$$\big| f_m(\mathbf{x}_i; \mathbf{W}^{(t)}) - f_m(\widehat{\mathbf{x}}_i; \mathbf{W}^{(t)}) \big| = \widetilde{O}(JP\sigma_0^3), \quad \forall\, m \in [M],\, i \in [n],\, t \in [T_1, T+1].$$

Moreover, we can write

$$y_i f_m(\widehat{\mathbf{x}}_i; \mathbf{W}^{(t)}) = \sum_{j \in [J]} \Big[\alpha_i^3\, \sigma\big(\langle \mathbf{w}_{m,j,p_1}^{(t)}, \mathbf{v}_k \rangle\big) + \gamma_i^3\, \sigma\big(\langle \mathbf{w}_{m,j,p_2}^{(t)}, \mathbf{v}_{k'} \rangle\big) \Big], \quad \forall\, i \in \Omega_{k,k'}^+,\, m \in [M],$$

and

$$y_i f_m(\widehat{\mathbf{x}}_i; \mathbf{W}^{(t)}) = \sum_{j \in [J]} \Big[\alpha_i^3\, \sigma\big(\langle \mathbf{w}_{m,j,p_1}^{(t)}, \mathbf{v}_k \rangle\big) - \gamma_i^3\, \sigma\big(\langle \mathbf{w}_{m,j,p_2}^{(t)}, \mathbf{v}_{k'} \rangle\big) \Big], \quad \forall\, i \in \Omega_{k,k'}^-,\, m \in [M].$$

Proof. For all $i \in \Omega_k$, we have

$$\big|f_m(\mathbf{x}_i; \mathbf{W}^{(t)}) - f_m(\widehat{\mathbf{x}}_i; \mathbf{W}^{(t)})\big| \le \Big|\sum_{j\in[J]} \sigma\big(\langle \mathbf{w}_{m,j,p_3}^{(t)}, \mathbf{c}_k\rangle\big)\Big| + \Big|\sum_{j\in[J],\, q\in Q_i} \sigma\big(\langle \mathbf{w}_{m,j,q}^{(t)}, \boldsymbol{\xi}_{i,q}\rangle\big)\Big|. \tag{26}$$

By bounding the magnitude of each term, we obtain

$$\begin{aligned}\big|f_m(\mathbf{x}_i; \mathbf{W}^{(t)}) - f_m(\widehat{\mathbf{x}}_i; \mathbf{W}^{(t)})\big| &\le \Big|\sum_{j\in[J]} \sigma\big(\langle \mathbf{w}_{m,j,p_3}^{(t)}, \mathbf{c}_k\rangle\big)\Big| + \Big|\sum_{j\in[J],\, q\in Q_i} \sigma\big(\langle \mathbf{w}_{m,j,q}^{(t)}, \xi_{i,q}\rangle\big)\Big| \\ &\le O(J)\max_{k,j} \sigma\big(\langle \mathbf{w}_{m,j,p_3}^{(t)}, \mathbf{c}_k\rangle\big) + O(JP)\max_{i,j,q}\big|\sigma\big(\langle \mathbf{w}_{m,j,q}^{(t)}, \xi_{i,q}\rangle\big)\big| \\ &= \widetilde{O}(JP\sigma_0^3).\end{aligned} \tag{27}$$

where the first inequality is by the triangle inequality and the last equality follows from Lemma 18. □

Lemma 20. *Suppose* (15)*,* (16)*,* (17) *hold for all* $t \in [T_1, T] \subseteq [T_1, T_2 - 1]$*, then we have that*

$$\|\mathbf{h}(\overline{\mathbf{x}}_i; \boldsymbol{\Theta}^{(t)}) - \mathbf{h}(\mathbf{x}_i; \boldsymbol{\Theta}^{(t)})\|_\infty = \widetilde{O}(\sigma_0^{2.5} M^2 P^2 K^2)$$

Besides, we have that

$$\max_{m,k,p} |\langle \boldsymbol{\theta}_{m,p}^{(t)}, \mathbf{v}_k\rangle|,\ \max_{m,i,p,q} |\langle \boldsymbol{\theta}_{m,p}^{(t)}, \boldsymbol{\xi}_{i,q}\rangle| = \widetilde{O}(\sigma_0^{2.5} M^2 P^2 K^2).$$

Proof. Recall the definition of δ_Θ in Lemma17, we need to show that $\delta_\Theta^{(t)} = \widetilde{O}(d^{-0.005})$ for all $t \in [T_1, T+1]$. We first prove the following router parameter update rules:

$$\begin{aligned}\langle \nabla_{\boldsymbol{\theta}_{m,p}} \mathcal{L}^{(t)}, \mathbf{v}_k\rangle &= -\widetilde{O}\left(\sigma_0^{1.5}\big(P^2K^2 + J\delta_{\Theta_{(t)}}\big)\right), \\ \langle \nabla_{\boldsymbol{\theta}_{m,p}} \mathcal{L}^{(t)}, \boldsymbol{\xi}_{i,q}\rangle &= -\widetilde{O}(\sigma_0^{1.5}), \quad \forall T_1 \le t \le T.\end{aligned} \tag{28}$$

Consider the inner product of the router gradient and the feature vector:

$$\begin{aligned}&\mathbb{E}\big[\langle \nabla_{\boldsymbol{\theta}_{m,p}} \mathcal{L}^{(t)}, \mathbf{v}_k\rangle\big] \\ &= \underbrace{\frac{1}{nP}\sum_{i\in\Omega_k} \mathbb{P}(m_{i,t} = m)\mathbb{E}\left[\ell_{i,t}' y_i \pi_m(\mathbf{x}_i; \boldsymbol{\Theta}^{(t)}) f_m(\mathbf{x}_i; \mathbf{W}^{(t)}) y_i \alpha_i \mid \mathbf{x}_i^{(p)} = \alpha_i y_i \mathbf{v}_k\right]}_{I_1} \\ &\quad + \underbrace{\frac{1}{nP}\sum_{i\in\Omega_{k',k}} \mathbb{P}(m_{i,t} = m)\mathbb{E}\left[\ell_{i,t}' y_i \pi_m(\mathbf{x}_i; \boldsymbol{\Theta}^{(t)}) f_m(\mathbf{x}_i; \mathbf{W}^{(t)}) \varepsilon_i \gamma_i \mid \mathbf{x}_i^{(p)} = \gamma_i \varepsilon_i \mathbf{v}_k\right]}_{I_2} \\ &\quad - \underbrace{\frac{1}{nP}\sum_{i\in\Omega_k, m'\in[M]} \mathbb{P}(m_{i,t} = m')\mathbb{E}\left[\ell_{i,t}' y_i \pi_{m'}(\mathbf{x}_i; \boldsymbol{\Theta}^{(t)}) \pi_m(\mathbf{x}_i; \boldsymbol{\Theta}^{(t)}) f_{m'}(\mathbf{x}_i; \mathbf{W}^{(t)}) y_i \alpha_i \mid \mathbf{x}_i^{(p)} = \alpha_i y_i \mathbf{v}_k\right]}_{I_3} \\ &\quad - \underbrace{\frac{1}{nP}\sum_{i\in\Omega_{k',k}, m'\in[M]} \mathbb{P}(m_{i,t} = m')\mathbb{E}\left[\ell_{i,t}' y_i \pi_{m'}(\mathbf{x}_i; \boldsymbol{\Theta}^{(t)}) \pi_m(\mathbf{x}_i; \boldsymbol{\Theta}^{(t)}) f_{m'}(\mathbf{x}_i; \mathbf{W}^{(t)}) \varepsilon_i \gamma_i \mid \mathbf{x}_i^{(p)} = \gamma_i \varepsilon_i \mathbf{v}_k\right]}_{I_4} \\ &\quad + \underbrace{\frac{1}{nP}\sum_{i\in[n], q\in Q_i} \mathbb{P}(m_{i,t} = m)\mathbb{E}\left[\ell_{i,t}' y_i \pi_m(\mathbf{x}_i; \boldsymbol{\Theta}^{(t)}) f_m(\mathbf{x}_i; \mathbf{W}^{(t)}) \langle \mathbf{x}_i^{(q)}, \mathbf{v}_k\rangle \mid \mathbf{x}_i^{(q)} = \xi_{i,q}\right]}_{I_5} \\ &\quad - \underbrace{\frac{1}{nP}\sum_{i\in[n], q\in Q_i, m'\in[M]} \mathbb{P}(m_{i,t} = m')\mathbb{E}\left[\ell_{i,t}' y_i \pi_{m'}(\mathbf{x}_i; \boldsymbol{\Theta}^{(t)}) \pi_m(\mathbf{x}_i; \boldsymbol{\Theta}^{(t)}) f_{m'}(\mathbf{x}_i; \mathbf{W}^{(t)}) \langle \mathbf{x}_i^{(q)}, \mathbf{v}_k\rangle \mid \mathbf{x}_i^{(q)} = \xi_{i,q}\right]}_{I_6}.\end{aligned} \tag{29}$$

We will bound I_k individually. Denote $y_i\pi_m(\widehat{\mathbf{x}}_i;\mathbf{\Theta}^{(t)})f_m(\widehat{\mathbf{x}}_i;\mathbf{W}^{(t)})$, $\forall i \in \Omega^+_{k,k',\tau}$ by $\overline{F}^+_{k,k',\tau}$. We show that the output of the MoE multiplied by label: $y_i\pi_m(\mathbf{x}_i;\mathbf{\Theta}^{(t)})f_m(\mathbf{x}_i;\mathbf{W}),\ \forall i \in \Omega^+_{k,k',\tau}$ can be approximated by $\overline{F}^+_{k,k',\tau}$. Indeed,

$$\begin{aligned}
&\big|\pi_m(\overline{\mathbf{x}}_i;\mathbf{\Theta}^{(t)})f_m(\mathbf{x}_i;\mathbf{W}^{(t)}) - \pi_m(\overline{\mathbf{x}}_i;\mathbf{\Theta}^{(t)})f_m(\widehat{\mathbf{x}}_i;\mathbf{W}^{(t)})\big| \\
&\quad\le \big|\pi_m(\overline{\mathbf{x}}_i;\mathbf{\Theta}^{(t)}) - \pi_m(\overline{\mathbf{x}}_i;\mathbf{\Theta}^{(t)})\big|f_m(\mathbf{x}_i;\mathbf{W}^{(t)}) + \pi_m(\overline{\mathbf{x}}_i;\mathbf{\Theta}^{(t)})\big|f_m(\mathbf{x}_i;\mathbf{W}^{(t)}) - f_m(\widehat{\mathbf{x}}_i;\mathbf{W}^{(t)})\big| \\
&\quad\le O\Big(JP(\eta T_2)^3\big|\pi_m(\overline{\mathbf{x}}_i;\mathbf{\Theta}^{(t)}) - \pi_m(\overline{\mathbf{x}}_i;\mathbf{\Theta}^{(t)})\big|\Big) + \big|f_m(\mathbf{x}_i;\mathbf{W}^{(t)}) - f_m(\widehat{\mathbf{x}}_i;\mathbf{W}^{(t)})\big| \\
&\quad\le \widetilde{O}\big(JP(\eta T_2)^3\delta_{\Theta^{(t)}} + JP\sigma_0^3\big),
\end{aligned} \tag{30}$$

where the first inequality is by triangle inequality, the second inequality is by $\pi_m(\widehat{\mathbf{x}}_i;\mathbf{\Theta}^{(t)}) \le 1$ and $|f_m(\mathbf{x}_i;\mathbf{W}^{(t)})| = O\big(JP(\eta T_2)^3\big)$ in Lemma 16, and the third inequality is by Lemma 17 and Lemma 19.

Similarly, denote $y_i\pi_m(\widehat{\mathbf{x}}_i;\mathbf{\Theta}^{(t)})f_m(\widehat{\mathbf{x}}_i;\mathbf{W}^{(t)})$, $i \in \Omega^-_{k,k',\tau}$ by $\overline{F}^-_{k,k',\tau}$, and we can show that the value $y_i\pi_m(\mathbf{x}_i;\mathbf{\Theta}^{(t)})f_m(\mathbf{x}_i;\mathbf{W}^{(t)})$, $\forall i \in \Omega^-_{k,k',\tau}$ can be approximated by $\overline{F}^-_{k,k',\tau}$. Let $g(z) = z\,\ell'(z) = \frac{z}{e^z+1}$, since $g'(z) = \frac{e^z+1-e^z z}{(e^z+1)^2} \le 1, \forall z \ge 0$, we have that

$$|g(z) - g(z')| = g'(a)|z - z'| \le |z - z'|, \forall\ z, z' \ge 0$$

Applying the above inequality with $z = y_i\pi_m(\widehat{\mathbf{x}}_i;\mathbf{\Theta}^{(t)})f_m(\widehat{\mathbf{x}}_i;\mathbf{W}^{(t)})$, $i \in \Omega^+_{k,k',\tau}$, $z' = \overline{F}^+_{k,k',\tau}$ gives us

$$\big|\ell'_{i,t}y_i\pi_m(\mathbf{x}_i;\mathbf{\Theta}^{(t)})f_m(\mathbf{x}_i;\mathbf{W}^{(t)}) - \ell'(\overline{F}^+_{k,k',\tau})\,\overline{F}^+_{k,k',\tau}\big| = \widetilde{O}\big(JP(\eta T_2)^3\delta_{\Theta^{(t)}} + JP\sigma_0^3\big)$$

Similarly, for $y_i\pi_m(\widehat{\mathbf{x}}_i;\mathbf{\Theta}^{(t)})f_m(\widehat{\mathbf{x}}_i;\mathbf{W}^{(t)})$, $i \in \Omega^-_{k,k',\tau}$. Now we can bound I_1 as follows:

$$\begin{aligned}
I_1 &= \frac{1}{nP}\sum_{i\in\Omega_k}\mathbb{P}(m_{i,t}=m)\mathbb{E}\Big[\ell'_{i,t}y_i\pi_m(\mathbf{x}_i;\mathbf{\Theta}^{(t)})f_m(\mathbf{x}_i;\mathbf{W}^{(t)})y_i\alpha_i \mid \mathbf{x}_i^{(p)} = \alpha_i y_i\mathbf{v}_k\Big] \\
&= \sum_\tau\sum_{k'\ne k}\sum_{i\in\Omega^+_{k,k',\tau}}\mathbb{P}(m_{i,t}=m)\mathbb{E}\left[\frac{\ell'(\overline{F}^+_{k,k',\tau})\,\overline{F}^+_{k,k',\tau} + O\big(JP(\eta T_2)^3\delta_{\Theta^{(t)}} + JP\sigma_0^3\big)}{nP}y_i\alpha_i\right] \\
&\quad + \sum_\tau\sum_{k'\ne k}\sum_{i\in\Omega^-_{k,k',\tau}}\mathbb{P}(m_{i,t}=m)\mathbb{E}\left[\frac{\ell'(\overline{F}^-_{k,k',\tau})\,\overline{F}^-_{k,k',\tau} + O\big(JP(\eta T_2)^3\delta_{\Theta^{(t)}} + JP\sigma_0^3\big)}{nP}y_i\alpha_i\right] \\
&= \sum_\tau\sum_{k'\ne k}\mathbb{E}\left[\frac{\ell'(\overline{F}^+_{k,k',\tau})\,\overline{F}^+_{k,k',\tau} + O\big(JP(\eta T_2)^3\delta_{\Theta^{(t)}} + JP\sigma_0^3\big)}{nP}\sum_{i\in\Omega^+_{k,k',\tau}}\mathbb{1}(m_{i,t}=m)y_i\alpha_i\right] \\
&\quad + \sum_\tau\sum_{k'\ne k}\mathbb{E}\left[\frac{\ell'(\overline{F}^-_{k,k',\tau})\,\overline{F}^-_{k,k',\tau} + O\big(JP(\eta T_2)^3\delta_{\Theta^{(t)}} + JP\sigma_0^3\big)}{nP}\sum_{i\in\Omega^-_{k,k',\tau}}\mathbb{1}(m_{i,t}=m)y_i\alpha_i\right] \\
&= -\widetilde{O}\left(\frac{P^2K^2(\eta T_2)^3 + J(\eta T_2)^3\delta_{\Theta_{(t)}} + J\sigma_0^3}{n^{1/2}}\right) \\
&= -\widetilde{O}\left(\sigma_0^{1.5}\big(P^2K^2 + J\delta_{\Theta_{(t)}}\big)\right)
\end{aligned} \tag{31}$$

Where the second to last equality is due to $\sum_{i\in\Omega^+_{k,k',\tau}}\mathbb{1}(m_{i,t}=m)y_i\alpha_i = \widetilde{O}(n^{1/2})$ by Hoeffding's inequality with $\mathbb{E}\Big[\mathbb{1}\big((\mathbf{x}_i,y_i)\in\Omega^+_{k,k',\tau}\big)\mathbb{1}(m_{i,t}=m)y_i\alpha_i\Big] = 0$. The last equality is due to the choice of η, T_2. Similarly, we can prove that $I_2, I_3, I_4 = -\widetilde{O}\left(\sigma_0^{1.5}\big(P^2K^2 + J\delta_{\Theta_{(t)}}\big)\right)$. Since $\langle\mathbf{x}_i^{(p)},\mathbf{v}_i\rangle = \widetilde{O}(d^{-1/2})$, $\forall q \in Q_i$, $\pi_m, \pi_{m_i,t} \le 1$ and $f_{m_i,t} = O\big(JP(\eta T_2)^3\big)$, we can upper bound I_5, I_6 by $\widetilde{O}\big(JP(\eta T_2)^3\sigma_0^{1.5}\big)$. Plugging those bounds into the gradient computation (29) gives:

$$\mathbb{E}\big[\langle\nabla_{\boldsymbol{\theta}_{m,p}}\mathcal{L}^{(t)},\mathbf{v}_k\rangle\big] = -\widetilde{O}\left(\sigma_0^{1.5}\big(P^2K^2 + J\delta_{\Theta_{(t)}}\big)\right).$$

We finally consider the alignment between router gradient and noise

$$\begin{aligned}\left\langle \nabla_{\boldsymbol{\theta}_{m,q}} \mathcal{L}^{(t)}, \boldsymbol{\xi}_{i',q'} \right\rangle &= \frac{1}{n} \sum_{i\in[n]} \mathbb{1}(m_{i,t}=m)\, \ell'_{i,t} y_i \pi_{m_{i,t}}(\mathbf{x}_i;\boldsymbol{\Theta}^{(t)}) f_{m_{i,t}}(\mathbf{x}_i;\mathbf{W}^{(t)}) \langle \mathbf{x}_i^{(q)}, \boldsymbol{\xi}_{i',q'}\rangle \\ &\quad - \frac{1}{n} \sum_{i\in[n]} \ell'_{i,t} y_i \pi_{m_{i,t}}(\mathbf{x}_i;\boldsymbol{\Theta}^{(t)}) \pi_m(\mathbf{x}_i;\boldsymbol{\Theta}^{(t)}) f_{m_{i,t}}(\mathbf{x}_i;\mathbf{W}^{(t)}) \langle \mathbf{x}_i^{(q)}, \boldsymbol{\xi}_{i',q'}\rangle \\ &\overset{(i)}{=} \widetilde{O}\left(\frac{1}{n}\right) + \widetilde{O}(d^{-1/2}) \\ &\overset{(ii)}{=} \widetilde{O}(\sigma_0^{1.5}),\end{aligned}$$

where (i) is by considering the cases $\xi_{i',q'} = \xi_{i,q}$ and $\xi_{i',q'} \neq \xi_{i,q}$ respectively, and (ii) is due to our choice of n. Now, we have completed the proof of 28.

Plugging the gradient estimation (28) into the gradient update rule for the gating network gives

$$\begin{aligned}\langle \boldsymbol{\theta}_{m,p}^{(t+1)}, \mathbf{v}_k \rangle &= \langle \boldsymbol{\theta}_{m,p}^{(t)}, \mathbf{v}_k \rangle + \widetilde{O}\left(\eta_r \sigma_0^{1.5}\big(P^2K^2 + J\delta_{\Theta_{(t)}}\big)\right) \\ \langle \boldsymbol{\theta}_{m,p}^{(t+1)}, \xi_{i,q} \rangle &= \langle \boldsymbol{\theta}_{m,p}^{(t)}, \xi_{i,q} \rangle + \widetilde{O}(\eta_r \sigma_0^{1.5})\end{aligned} \tag{32}$$

From (32) we have

$$\begin{aligned}\delta_{\Theta^{(t+1)}} &\le \delta_{\Theta^{(t)}} + \widetilde{O}\left(\eta_r \sigma_0^{1.5}\big(P^2K^2 + J\delta_{\Theta_{(t)}}\big)\right) \\ &\le \delta_{\Theta^{(t)}} \left[1 + \widetilde{O}\left(\eta_r \sigma_0^{1.5} J\right)\right] + \widetilde{O}\left(\eta_r \sigma_0^{1.5} P^2K^2\right) \\ &\le \exp\left(\widetilde{O}\left(\eta_r \sigma_0^{1.5} J\right)\right)\left(\delta_{\Theta^{(t)}} + \widetilde{O}\left(\eta_r \sigma_0^{1.5} P^2K^2\right)\right),\end{aligned} \tag{33}$$

where the last inequality is due to $1 + z \le \exp(z)$ for all $z \in \mathbb{R}$. This implies that

$$\delta_{\Theta^{(t)}} \le \exp\left(\widetilde{O}\left(t\eta_r \sigma_0^{1.5} J\right)\right)\left(\delta_{\Theta^{(0)}} + \widetilde{O}\left(\eta_r \sigma_0^{1.5} P^2K^2\right)\right) = \widetilde{O}(\sigma_0^{2.5} M^2 P^2 K^2),$$

where the last equality is by $\widetilde{O}\left(t\eta_r \sigma_0^{1.5} J\right) = \widetilde{O}(t\eta M^2 \sigma^{1.5} J) = \widetilde{O}(1)$ and $\widetilde{O}\left(\eta_r \sigma_0^{1.5} P^2K^2\right) = \widetilde{O}(\sigma_0^{2.5} M^2 P^2 K^2)$ due to the choice of M, η_r. □

Define

$$\Delta_{\boldsymbol{\Theta}} \coloneqq \max_{k\in[K]} \max_{m,m'\in\mathcal{M}_k} \max_{(\mathbf{x}_i,y_i)\in\Omega_k} \left|h_m(\mathbf{x}_i;\boldsymbol{\Theta}) - h_{m'}(\mathbf{x}_i;\boldsymbol{\Theta})\right|,$$

which measures the router's bias toward different experts within the same $\mathcal{M}_k$. The following lemma shows that when $\Delta_{\boldsymbol{\Theta}}$ is small, the router treats professional experts approximately equally.

Lemma 21. *For all $t \ge 0$, the following inequalities hold:*

$$\begin{aligned}\max_{k\in[K]} \max_{m,m'\in\mathcal{M}_k} \max_{(\mathbf{x}_i,y_i)\in\Omega_k} \left|\pi_{m'}(\mathbf{x}_i;\boldsymbol{\Theta}^{(t)}) - \pi_m(\mathbf{x}_i;\boldsymbol{\Theta}^{(t)})\right| &\le 2\,\Delta_{\boldsymbol{\Theta}^{(t)}}, \\ \max_{k\in[K]} \max_{m,m'\in\mathcal{M}_k} \max_{(\mathbf{x}_i,y_i)\in\Omega_k} \left|\mathbb{P}(m_{i,t}=m) - \mathbb{P}(m_{i,t}=m')\right| &= O(M^2)\,\Delta_{\boldsymbol{\Theta}^{(t)}}.\end{aligned}$$

Proof. By Lemma 1, we have

$$\left|\mathbb{P}(m_{i,t}=m) - \mathbb{P}(m_{i,t}=m')\right| \le O(M^2)\left|h_m(\mathbf{x}_i;\boldsymbol{\Theta}^{(t)}) - h_{m'}(\mathbf{x}_i;\boldsymbol{\Theta}^{(t)})\right|.$$

Next, we show that

$$\left|\pi_{m'}(\mathbf{x}_i;\boldsymbol{\Theta}) - \pi_m(\mathbf{x}_i;\boldsymbol{\Theta})\right| \le 2\left|h_m(\mathbf{x}_i;\boldsymbol{\Theta}^{(t)}) - h_{m'}(\mathbf{x}_i;\boldsymbol{\Theta}^{(t)})\right|. \tag{34}$$

When $\left|h_m(\mathbf{x}_i;\boldsymbol{\Theta}^{(t)}) - h_{m'}(\mathbf{x}_i;\boldsymbol{\Theta}^{(t)})\right| \ge 1$, the inequality (34) is trivial. When $\left|h_m(\mathbf{x}_i;\boldsymbol{\Theta}^{(t)}) - h_{m'}(\mathbf{x}_i;\boldsymbol{\Theta}^{(t)})\right| \le 1$, we have

$$\begin{aligned}\left|\pi_{m'}(\mathbf{x}_i;\boldsymbol{\Theta}) - \pi_m(\mathbf{x}_i;\boldsymbol{\Theta})\right| &= \left|\frac{\exp\big(h_m(\mathbf{x}_i;\boldsymbol{\Theta}^{(t)})\big) - \exp\big(h_{m'}(\mathbf{x}_i;\boldsymbol{\Theta}^{(t)})\big)}{\sum_{m''}\exp\big(h_{m''}(\mathbf{x}_i;\boldsymbol{\Theta}^{(t)})\big)}\right| \\ &= \left|\frac{\exp\big(h_{m'}(\mathbf{x}_i;\boldsymbol{\Theta}^{(t)})\big)}{\sum_{m''}\exp\big(h_{m''}(\mathbf{x}_i;\boldsymbol{\Theta}^{(t)})\big)}\right| \cdot \left|\exp\big(h_m(\mathbf{x}_i;\boldsymbol{\Theta}^{(t)}) - h_{m'}(\mathbf{x}_i;\boldsymbol{\Theta}^{(t)})\big) - 1\right| \\ &\le 2\left|h_m(\mathbf{x}_i;\boldsymbol{\Theta}^{(t)}) - h_{m'}(\mathbf{x}_i;\boldsymbol{\Theta}^{(t)})\right|.\end{aligned}$$

This completes the proof for (34).

□

Notice that the gating network is initialized to be zero, so we have $\Delta_{\Theta^{(t)}} = 0$ at initialization. We can further show that $\Delta_\Theta = \widetilde{O}(\sigma_0^{1.5})$ during the training up to time $T = T_2$.

Lemma 22. *Suppose* (15)*,* (16)*,* (17) *hold for all $t \in [T_1, T] \subseteq [T_1, T_2 - 1]$, then we have that $\Delta_{\Theta^{(t)}} \leq \widetilde{O}(\sigma_0^{1.5})$ holds for all $t \in [T_1, T + 1]$.*

Proof. One of the key observations is the similarity of the m-th and the m'-th expert in the same expert class $\mathcal{M}_k$. Lemma 18 and Lemma 19 imply that

$$
\begin{aligned}
&\big|f_m(x_i, \mathbf{W}^{(t)}) - f_{m'}(x_i, \mathbf{W}^{(t)})\big| \\
&= \big|f_m(\widehat{x}_i, \mathbf{W}^{(t)}) - f_{m'}(\widehat{x}_i, \mathbf{W}^{(t)})\big| + \widetilde{O}(JP\sigma_0^3) \\
&\leq \big|y_i\alpha_i^3 \sum_{j\in[J]} \Big[\sigma(\langle \mathbf{w}_{m,j,p_1}^{(t)}, \mathbf{v}_k\rangle) - \sigma(\langle \mathbf{w}_{m,j,p_1}^{(t)}, \mathbf{v}_k\rangle)\Big]\big| + \big|\epsilon_i\gamma_i^3 \sum_{j\in[J]} \Big[\sigma(\langle \mathbf{w}_{m,j,p_3}^{(t)}, \mathbf{v}_{k'}\rangle) - \sigma(\langle \mathbf{w}_{m,j,p_3}^{(t)}, \mathbf{v}_{k'}\rangle)\Big]\big| + \widetilde{O}(JP\sigma_0^3) \\
&\leq \alpha_i^3\big|\sigma(\langle \mathbf{w}_{m,j_m^*,p_1}^{(t)}, \mathbf{v}_k\rangle) - \sigma(\langle \mathbf{w}_{m,j_{m'}^*,p_1}^{(t)}, \mathbf{v}_k\rangle)\big| + (JP\sigma_0^3) \\
&\leq \alpha_i^3 O(\sigma_0^{0.2}) + (JP\sigma_0^3) = O(\sigma_0^{0.2}).
\end{aligned}
\tag{35}
$$

Another key observation is that we only need to focus on the k-th cluster-center signal. Lemma20 implies that,

$$
\begin{aligned}
\Delta_{\Theta^{(t)}} &= \max_{k\in[K]} \max_{m,m'\in\mathcal{M}_k} \max_{(x_i,y_i)\in\Omega_k} \big|h_m(x_i;\Theta) - h_{m'}(x_i;\Theta^{(t)})\big| \\
&\leq \max_{k\in[K]} \max_{m,m'\in\mathcal{M}_k} \max_{(x_i,y_i)\in\Omega_k} \big|h_m(\overline{x}_i;\Theta^{(t)}) - h_{m'}(\overline{x}_i;\Theta^{(t)})\big| + 2\delta_{\Theta^{(t)}} \\
&= \max_{k\in[K]} \max_{m,m'\in\mathcal{M}_k} \max_{p\in[P]} \big|\langle \theta_{m,p}^{(t)} - \theta_{m',p}^{(t)}, \beta_i c_k\rangle\big| + 2\delta_{\Theta^{(t)}} \\
&= C_2 \max_{k\in[K]} \max_{m,m'\in\mathcal{M}_k} \big|\langle \theta_{m,p}^{(t)} - \theta_{m',p}^{(t)}, c_k\rangle\big| + 2\delta_{\Theta^{(t)}},
\end{aligned}
$$

where the first inequality is by definition of $\delta_{\Theta^{(t)}}$ and the last equality is by $\max_i \beta_i = C_2$. With a similar argument, we have the corresponding lower bound

$$
\Delta_{\Theta^{(t)}} \geq C_2 \max_{k\in[K]} \max_{m,m'\in\mathcal{M}_k} \big|\langle \theta_{m,p}^{(t)} - \theta_{m',p}^{(t)}, c_k\rangle\big| - 2\delta_{\Theta^{(t)}}. \tag{36}
$$

We now prove that the following gradient difference is small

$$
\begin{aligned}
&\mathbb{E}\Big[\langle \nabla_{\theta_{m,p}}\mathcal{L}^{(t)} - \nabla_{\theta_{m',p}}\mathcal{L}^{(t)},\, c_k\rangle\Big] \\
&= \frac{1}{n}\sum_{i\in\Omega_k} \mathbb{P}(m_{i,t} = m)\, \ell_{i,t}'\, \pi_m(x_i;\Theta^{(t)})\, y_i f_m(x_i;\mathbf{W}^{(t)})\langle x_i^{(p)}, c_k\rangle \\
&- \frac{1}{n}\sum_{i\in\Omega_k} \mathbb{P}(m_{i,t} = m')\, \ell_{i,t}'\, \pi_{m'}(x_i;\Theta^{(t)})\, y_i f_{m'}(x_i;\mathbf{W}^{(t)})\langle x_i^{(p)}, c_k\rangle \\
&- \frac{1}{n}\sum_{i\in\Omega_k, m''\in[M]} \big[\pi_m(x_i;\Theta^{(t)}) - \pi_{m'}(x_i;\Theta^{(t)})\big]\mathbb{P}(m_{i,t} = m'')\, \ell_{i,t}'\, \pi_{m''}(x_i;\Theta^{(t)}) y_i f_{m''}(\mathbf{x}_i;\mathbf{W}^{(t)}\langle \mathbf{x}_i^{(p)}, \mathbf{c}_k\rangle.
\end{aligned}
\tag{37}
$$

First, we bound the last term in (37) as follows.

$$
\begin{aligned}
&\frac{1}{n}\sum_{i\in\Omega_k,m''\in[M]}\big[\pi_m(x_i;\Theta^{(t)})-\pi_{m'}(x_i;\Theta^{(t)})\big]\mathbb{P}(m_{i,t}=m'')\,\ell'_{i,t}\,\pi_{m''}(x_i;\Theta^{(t)})y_i f_{m''}(\mathbf{x}_i;\mathbf{W}^{(t)}\langle\mathbf{x}_i^{(p)},\mathbf{c}_k\rangle\\
&=-\frac{1}{2n}\sum_{i\in\Omega_k}\left[\sum_{m''\in[M]}\mathbb{P}(m_{i,t}=m'')\pi_{m''}(\mathbf{x}_i;\boldsymbol{\Theta}^{(t)})\right]\beta_i y_i O(JP)+O(\sigma_0^{1.5})\\
&\le-\frac{1}{2n}\sum_{i\in\Omega_k}\beta_i y_i O(JP)+O(\sigma_0^{1.5})\\
&=\widetilde{O}(JP\sigma_0^{1.5}),
\end{aligned}
\tag{38}
$$

where the first equality is due to Lemma 10 and Lemma 16, the inequality is due to Lemma 4. To bound the first two terms, we make use of (35).

$$
\begin{aligned}
&\frac{1}{n}\sum_{i\in\Omega_k}\mathbb{P}(m_{i,t}=m)\,\ell'_{i,t}\,\pi_m(x_i;\Theta^{(t)})\,y_i f_m(x_i;\mathbf{W}^{(t)})\langle x_i^{(p)},c_k\rangle\\
&-\frac{1}{n}\sum_{i\in\Omega_k}\mathbb{P}(m_{i,t}=m')\,\ell'_{i,t}\,\pi_{m'}(x_i;\Theta^{(t)})\,y_i f_{m'}(x_i;\mathbf{W}^{(t)})\langle x_i^{(p)},c_k\rangle\\
&=\frac{1}{n}\sum_{i\in\Omega_k}\beta_i y_i\ell'_{i,t}\Big[\mathbb{P}(m_{i,t}=m)\,\pi_m(x_i;\Theta^{(t)})\,f_m(x_i;\mathbf{W}^{(t)})-\mathbb{P}(m_{i,t}=m')\,\pi_{m'}(x_i;\Theta^{(t)})\,f_{m'}(x_i;\mathbf{W}^{(t)})\Big]\\
&\le\frac{1}{n}(-1/2+\widetilde{O}(\sigma_0^{1.5}))\sum_{i\in\Omega_k}\beta_i y_i(O(M^2)\Delta_{\boldsymbol{\Theta}^{(t)}}+\widetilde{O}(\sigma_0^{0.2}))\\
&\le(-1/2+\widetilde{O}(\sigma_0^{1.5}))(O(M^2\sigma_0^{1.5})\Delta_{\boldsymbol{\Theta}^{(t)}}+\widetilde{O}(\sigma_0^{1.7})),
\end{aligned}
\tag{39}
$$

where the first inequality is due to (35) and Lemma 21, the second inequality is due to Lemma 4. Combining equations (38) and (39) we have

$$
\mathbb{E}\Big[\langle\nabla_{\theta_{m,p}}\mathcal{L}^{(t)}-\nabla_{\theta_{m',p}}\mathcal{L}^{(t)},\,c_k\rangle\Big]=-\Big(\widetilde{O}(JP\sigma_0^{1.5})+O(M^2\sigma_0^{1.5})\Delta_{\boldsymbol{\Theta}^{(t)}}\Big)
\tag{40}
$$

By Lemma 9, this implies that $\langle\nabla_{\theta_{m,p}}\mathcal{L}^{(t)}-\nabla_{\theta_{m',p}}\mathcal{L}^{(t)},\,c_k\rangle=-\Big(\widetilde{O}(JP\sigma_0^{1.5})+O(M^2\sigma_0^{1.5})\Delta_{\boldsymbol{\Theta}^{(t)}}\Big)$. Therefore, we can write the gradient updates of $\langle\theta_{m,p}^{(t)}-\theta_{m',p}^{(t)},c_k\rangle$ as follows

$$
\langle\theta_{m,p}^{(t+1)}-\theta_{m',p}^{(t+1)},c_k\rangle=\langle\theta_{m,p}^{(t)}-\theta_{m',p}^{(t)},c_k\rangle+\eta_r\Big(\widetilde{O}(JP\sigma_0^{1.5})+O(M^2\sigma_0^{1.5})\Delta_{\boldsymbol{\Theta}^{(t)}}\Big)
\tag{41}
$$

This allows us to bound $\Delta_{\boldsymbol{\Theta}^{(t)}}$ as follows.

$$
\begin{aligned}
\Delta_{\boldsymbol{\Theta}^{(t+1)}}&\le C_2\max_{k\in[K]}\max_{m,m'\in\mathcal{M}_k}\big|\langle\theta_{m,p}^{(t+1)}-\theta_{m',p}^{(t+1)},c_k\rangle\big|+2\delta_{\Theta^{(t+1)}}\\
&=C_2\max_{k\in[K]}\max_{m,m'\in\mathcal{M}_k}\big|\langle\theta_{m,p}^{(t)}-\theta_{m',p}^{(t)},c_k\rangle\big|+\eta_r\Big(\widetilde{O}(JP\sigma_0^{1.5})+O(M^2\sigma_0^{1.5})\Delta_{\boldsymbol{\Theta}^{(t)}}\Big)+2\delta_{\Theta^{(t+1)}}\\
&\le\Delta_{\boldsymbol{\Theta}^{(t)}}+2\delta_{\Theta^{(t)}}+\eta_r\Big(\widetilde{O}(JP\sigma_0^{1.5})+O(M^2\sigma_0^{1.5})\Delta_{\boldsymbol{\Theta}^{(t)}}\Big)+2\delta_{\Theta^{(t+1)}}\\
&\le\Delta_{\boldsymbol{\Theta}^{(t)}}(1+O(\eta_r M^2\sigma_0^{1.5}))+O(\eta_r JP\sigma_0^{1.5})+\widetilde{O}(P^2\sigma_0^{1.5})
\end{aligned}
\tag{42}
$$

Following the same argument in Lemma 20 for $\delta_{\boldsymbol{\Theta}^{(t)}}$, we have that

$$
\Delta_{\boldsymbol{\Theta}^{(t)}}=\widetilde{O}(\eta_r\exp(t\eta_r M^2\sigma_0^{1.5})\sigma_0^{1.5})=\widetilde{O}(\sigma_0^{1.5})
$$

Note that the last equality is due to the choice of M. □

Lemma 23. *Suppose* (15), (16), (17) *hold for all* $t\in[T_1,T]\subseteq[T_1,T_2-1]$. *Then for* $m\notin\mathcal{M}_k$ *and* $t\in[T_1,T]$, *if* $\langle\theta_{m,p}^{(t)},c_k\rangle\ge\max_{m'}\langle\theta_{m',p}^{(t)},c_k\rangle-1$, *we have that*

$$
\big\langle\nabla_{\theta_{m,p}}\mathcal{L}^{(t)},c_k\big\rangle\ge\Omega\bigg(\frac{\eta^3t^3}{KM^3}\bigg)-\widetilde{O}(\sigma_0^{1.5}).
$$

Moreover, even without the above condition, we can still show that

$$\left\langle \nabla_{\theta_{m,p}} \mathcal{L}^{(t)}, c_k \right\rangle \geq -\widetilde{O}(\sigma_0^{1.5}).$$

Proof. The expectation of the inner product $\langle \nabla_{\theta_{m,p}} \mathcal{L}^{(t)}, c_k \rangle$ can be computed as follows,

$$\begin{aligned}
&\mathbb{E}\Big[\left\langle \nabla_{\theta_{m,p}} \mathcal{L}^{(t)}, c_k \right\rangle\Big] \\
&= \frac{1}{n} \sum_{i\in[n]} \mathbb{P}(m_{i,t} = m)\, \ell'_{i,t}\, y_i \mathbb{E}\left[\pi_m(x_i; \Theta^{(t)}) f_m(x_i; \mathbf{W}^{(t)}) \left\langle x_i^{(p)}, c_k \right\rangle\right] \\
&\quad - \frac{1}{n} \sum_{i\in[n], m'\in[M]} \mathbb{P}(m_{i,t} = m')\, \ell'_{i,t}\, y_i \mathbb{E}\left[\pi_{m'}(x_i; \Theta^{(t)})\, \pi_m(x_i; \Theta^{(t)})\, f_{m'}(x_i; \mathbf{W}^{(t)}) \left\langle x_i^{(p)}, c_k \right\rangle\right] \\
&= \frac{1}{nP} \sum_{i\in\Omega_k} \mathbb{P}(m_{i,t} = m)\, \ell'_{i,t}\, y_i \mathbb{E}\left[\pi_m(x_i; \Theta^{(t)})\, \beta_i f_m(x_i; \mathbf{W}^{(t)}) \mid \mathbf{x}_i^{(p)} = \beta_i \mathbf{c}_k\right] + \widetilde{O}\big(P\sigma_0^{1.5}\big) \\
&\quad - \frac{1}{nP} \sum_{i\in\Omega_k} \sum_{m'\in[M]} \mathbb{P}(m_{i,t} = m')\, \ell'_{i,t}\, y_i \mathbb{E}\left[\pi_{m'}(x_i; \Theta^{(t)})\, \pi_m(x_i; \Theta^{(t)})\, \beta_i f_{m'}(x_i; \mathbf{W}^{(t)}) \mid \mathbf{x}_i^{(p)} = \beta_i \mathbf{c}_k\right].
\end{aligned} \tag{43}$$

where the second equality is due to $\langle \mathbf{x}_i^{(p)}, \mathbf{c}_k \rangle = \beta_i$ if $\mathbf{x}_i^{(p)} = \beta_i \mathbf{c}_k$, $\langle \mathbf{x}_i^{(p)}, \mathbf{c}_k \rangle = \widetilde{O}(\sigma_0^{1.5})$ if $\mathbf{x}_i^{(p)} = \xi_{i,q}, q \in Q_i$ by Lemma 5, and $f_m(x_i; \mathbf{W}^{(t)}) = O(P\sigma_0^{1.5})$ by Lemma 16.

We can rewrite the inner product as follows,

$$\begin{aligned}
&\mathbb{E}\Big[\left\langle \nabla_{\theta_{m,p}} \mathcal{L}^{(t)}, c_k \right\rangle\Big] \\
&= \underbrace{\frac{1}{nP} \sum_{i\in\Omega_k} \mathbb{P}(m_{i,t} = m)\, \ell'_{i,t}\, y_i\, \mathbb{E}\Big[\pi_m(x_i; \Theta^{(t)})\, \beta_i f_m(x_i; \mathbf{W}^{(t)}) \;\Big|\; \mathbf{x}_i^{(p)} = \beta_i \mathbf{c}_k\Big]}_{I_1} + \widetilde{O}\big(P\sigma_0^{1.5}\big) \\
&\quad \underbrace{- \frac{1}{nP} \sum_{i\in\Omega_k} \sum_{m'\in\mathcal{M}_k} \mathbb{P}(m_{i,t} = m')\, \ell'_{i,t}\, y_i\, \mathbb{E}\Big[\pi_{m'}(x_i; \Theta^{(t)})\, \pi_m(x_i; \Theta^{(t)})\, \beta_i f_{m'}(x_i; \mathbf{W}^{(t)}) \;\Big|\; \mathbf{x}_i^{(p)} = \beta_i \mathbf{c}_k\Big]}_{I_2} \\
&\quad \underbrace{- \frac{1}{nP} \sum_{i\in\Omega_k} \sum_{m'\notin\mathcal{M}_k} \mathbb{P}(m_{i,t} = m')\, \ell'_{i,t}\, y_i\, \mathbb{E}\Big[\pi_{m'}(x_i; \Theta^{(t)})\, \pi_m(x_i; \Theta^{(t)})\, \beta_i f_{m'}(x_i; \mathbf{W}^{(t)}) \;\Big|\; \mathbf{x}_i^{(p)} = \beta_i \mathbf{c}_k\Big]}_{I_3}.
\end{aligned} \tag{44}$$

We first give a lower bound for I_2. Since $\langle \theta_{m,p}^{(t)}, c_k \rangle \geq \max_{m'} \langle \theta_{m',p}^{(t)}, c_k \rangle - 1$, we have

$$\mathbb{E}\left[\pi_m(\overline{\mathbf{x}}_i; \Theta^{(t)}) \mid \mathbf{x}_i^{(p)} = \beta_i \mathbf{c}_k\right] = \Omega(1/M).$$

By Lemma 17 and Lemma 20 we also have

$$|\pi_m(\mathbf{x}_i; \Theta^{(t)}) - \pi_m(\overline{\mathbf{x}}_i; \Theta^{(t)})| = O(\delta_{\Theta_t}) = \widetilde{O}(\sigma_0^{2.5} M^2 P^2 K^2).$$

Therefore,

$$\mathbb{E}\left[\pi_m(\mathbf{x}_i; \Theta^{(t)}) \mid \mathbf{x}_i^{(p)} = \beta_i \mathbf{c}_k\right] = \Omega(1/M - M^2P^2K^2\sigma_0^{2.5}) = \Omega(1/M) \tag{45}$$

due to the choice of M. Moreover, for any $m' \in \mathcal{M}_k$, we can lower bound $y_i f_{m'}(x_i; \mathbf{W}^{(t)})$ as follows,

$$\begin{aligned}
y_i f_{m'}(x_i; \mathbf{W}^{(t)}) &= f_{m'}(x_i; \mathbf{W}^{(t)}) + \widetilde{O}(JP\sigma_0^3) \\
&= \sum_{j\in[J]} \left[\alpha_i^3 \sigma(\langle \mathbf{w}_{m,j,p_1}, \mathbf{v}_k \rangle) + \epsilon_i y_i \gamma_i^3 \sigma(\langle \mathbf{w}_{m,j,p_2}, \mathbf{v}_{k'} \rangle)\right] \\
&= \alpha_i^3 \langle \mathbf{w}_{m,j,p_1}, \mathbf{v}_k \rangle^3 + O(JM^3\sigma_0^3) \\
&\geq \Omega(\eta^3 t^3 (1 - \sigma_0^{0.2})) + O(JM^3\sigma_0^3) \\
&\geq \Omega(\eta^3 t^3),
\end{aligned} \tag{46}$$

where the first equality is due to Lemma 19, the third equality and first inequality are due Lemma 18. This allows us to lower bound I_2 as follows,

$$\begin{aligned} I_2 &= -\frac{1}{nP}\sum_{i\in\Omega_k}\sum_{m'\in\mathcal{M}_k}\mathbb{P}(m_{i,t}=m')\,\ell'_{i,t}\,y_i\mathbb{E}\left[\pi_{m'}(x_i;\Theta^{(t)})\,\pi_m(x_i;\Theta^{(t)})\,\beta_i f_{m'}(x_i;\mathbf{W}^{(t)})\mid \mathbf{x}_i^{(p)}=\beta_i\mathbf{c}_k\right] \\ &\geq \Omega\left(\frac{\eta^3t^3}{PnM^3}\right)\sum_{i\in\Omega_k,m'\in\mathcal{M}_k}\beta_i \\ &\geq \Omega\left(\frac{\eta^3t^3}{PKM^3}\right), \end{aligned} \tag{47}$$

where the first inequality is due to (45) and (46), along with Lemma 10 and Proposition E.9. The last inequality is due to Lemma 4 and Lemma 6.

It remains to show that I_1, I_3 are upper bounded. For any $m\notin\mathcal{M}_k$, we utilize Lemma 18 to show an upper bound on the following term

$$\frac{1}{nP}\sum_{i\in\Omega_k}\mathbb{P}(m_{i,t}=m)\,\ell'_{i,t}\,y_i\mathbb{E}\left[\pi_m(x_i;\Theta^{(t)})\,\beta_i f_m(x_i;\mathbf{W}^{(t)})\mid \mathbf{x}_i^{(p)}=\beta_i\mathbf{c}_k\right]$$

.

Suppose that $m\in\mathcal{M}_{k'}, k'\neq k$. When $i\notin\Omega_{k,k'}$, we have $y_i f_m(x_i;\mathbf{W}^{(t)})=O(JP\sigma_0^3)$ due to Lemma 18. Therefore,

$$\mathbb{E}\left[\pi_m(x_i;\Theta^{(t)})\,\beta_i f_m(x_i;\mathbf{W}^{(t)})\mid \mathbf{x}_i^{(p)}=\beta_i\mathbf{c}_k, i\notin\Omega_{k,k'}\right]=\widetilde{O}(JP\sigma_0^3) \tag{48}$$

When $i\in\Omega^+_{k,k'}$, we have $y_i f_m(x_i;\mathbf{W}^{(t)})=\gamma_i^3\langle\mathbf{w}_{m,j,p_2},\mathbf{v}_{k'}\rangle^3+O(JP\sigma_0^3)$ due to Lemma 18. Similarly, when $i\in\Omega^-_{k,k'}$, we have $y_i f_m(x_i;\mathbf{W}^{(t)})=-\gamma_i^3\langle\mathbf{w}_{m,j,p_2},\mathbf{v}_{k'}\rangle^3+O(JP\sigma_0^3)$. This implies

$$\begin{aligned} &\mathbb{E}\left[\pi_m(x_i;\Theta^{(t)})\,\beta_i f_m(x_i;\mathbf{W}^{(t)})\mid \mathbf{x}_i^{(p)}=\beta_i\mathbf{c}_k, i\in\Omega_{k,k'}\right] \\ &=\mathbb{E}\left[\pi_m(x_i;\Theta^{(t)})\,\beta_i f_m(x_i;\mathbf{W}^{(t)})\mid \mathbf{x}_i^{(p)}=\beta_i\mathbf{c}_k, i\in\Omega^+_{k,k'}\right] \\ &+\mathbb{E}\left[\pi_m(x_i;\Theta^{(t)})\,\beta_i f_m(x_i;\mathbf{W}^{(t)})\mid \mathbf{x}_i^{(p)}=\beta_i\mathbf{c}_k, i\in\Omega^-_{k,k'}\right] \\ &=\widetilde{O}(JP\sigma_0^3) \end{aligned} \tag{49}$$

Combining (48) and (49) gives us

$$\frac{1}{nP}\sum_{i\in\Omega_k}\mathbb{P}(m_{i,t}=m)\,\ell'_{i,t}\,y_i\mathbb{E}\left[\pi_m(x_i;\Theta^{(t)})\,\beta_i f_m(x_i;\mathbf{W}^{(t)})\mid \mathbf{x}_i^{(p)}=\beta_i\mathbf{c}_k\right]=\widetilde{O}(JP\sigma_0^3) \tag{50}$$

Therefore, it is easy to show that $I_1=O(JP\sigma_0^3$ and $I_3=O(JP\sigma_0^3)$. Plugging the bounds of I_1, I_2, I_3 into (44) gives us

$$\mathbb{E}\Big[\big\langle\nabla_{\theta_{m,p}}\mathcal{L}^{(t)},c_k\big\rangle\Big]\geq\Omega\left(\frac{\eta^3t^3}{PKM^3}\right)+\widetilde{O}(P\sigma_0^{1.5}) \tag{51}$$

By Lemma 9, this implies that

$$\big\langle\nabla_{\theta_{m,p}}\mathcal{L}^{(t)},c_k\big\rangle\geq\Omega\left(\frac{\eta^3t^3}{PKM^3}\right)+\widetilde{O}(P\sigma_0^{1.5})+\widetilde{O}(\sigma_0^{1.5})\geq\Omega\left(\frac{\eta^3t^3}{PKM^3}\right)-\widetilde{O}(P\sigma_0^{1.5}) \tag{52}$$

Moreover, for $m\notin\mathcal{M}_k$ and $t\in[T_1,T]$, even without the condition $\langle\theta^{(t)}_{m,p},c_k\rangle\geq\max_{m'}\langle\theta^{(t)}_{m',p},c_k\rangle-1$,, we can still show that $I_2\geq 0$, hence

$$\big\langle\nabla_{\theta_{m,p}}\mathcal{L}^{(t)},c_k\big\rangle\geq-\widetilde{O}(P\sigma_0^{1.5}) \tag{53}$$

This completes the proof. □

Lemma 24. *For all $T_1\leq t\leq T_2$, Proposition 3 holds. Moreover,*

$$\langle\theta^{(T_2)}_{m,p},c_k\rangle\leq\max_{m'\in[M]}\langle\theta^{(T_2)}_{m',p},c_k\rangle-\Omega\big(P^{-1}K^{-1}M^{-9}\big)\quad\textit{for all } m\notin\mathcal{M}_k..$$

Proof. We first prove Proposition 3 by induction. Note that Proposition 3 holds at the beginning of the second stage $t = T_1$. Suppose (15), (16), and (17) hold for all $t \in [T_1, T] \subseteq [T_1, T_2 - 1]$. We verify that they also hold for $t \in [T_1, T+1]$. Lemma 19 implies that (15) holds for $t \in [T_1, T+1]$, and Lemma 20 implies that (16) holds for $t \in [T_1, T+1]$. Hence it remains to check (17) for $t \in [T_1, T+1]$. For each pair $i \in \Omega_k$ and $m \in \mathcal{M}_k$, we estimate the gap between expert m and the best-performing expert, $h_m(\mathbf{x}_i; \mathbf{\Theta}^{(t)}) - \max_{m'} h_{m'}(\mathbf{x}_i; \mathbf{\Theta}^{(t)})$. By Lemma 23 and Lemma 20, we can deduce that $h_{m'}(\mathbf{x}_i; \mathbf{\Theta}^{(t)})$ for $m' \notin \mathcal{M}_k$ is small therefore cannot be the largest one. Thus,

$$h_m(\mathbf{x}_i; \mathbf{\Theta}^{(t)}) - \max_{m'} h_{m'}(\mathbf{x}_i; \mathbf{\Theta}^{(t)}) \leq \Delta_{\mathbf{\Theta}^{(t)}} = \widetilde{O}\big(\sigma_0^{1.5}\big).$$

Therefore, by Lemma 25, (17) holds. This completes the induction for Proposition 3.

We next characterize $\langle \theta_{m,p}^{(t)}, c_k \rangle$ for $\eta_r \eta^{-1} = \Theta(M^2)$ and $m \notin \mathcal{M}_k$. If there exists $t \leq T_2$ such that $\langle \theta_{m,p}^{(t)}, c_k \rangle \leq \max_{m'} \langle \theta_{m',p'}^{(t)}, c_k \rangle - 1$, then by Lemma 23,

$$\langle \theta_{m,p}^{(T_2)}, c_k \rangle - \max_{m'} \langle \theta_{m',p}^{(t)}, c_k \rangle \leq \langle \theta_{m,p}^{(t)}, c_k \rangle - \max_{m'} \langle \theta_{m',p}^{(t)}, c_k \rangle + \widetilde{O}(\eta_r T_2 P \sigma_0^{1.5}) \leq -1 + \widetilde{O}(P\sigma_0^{1.5}) = -\Omega\big(P^{-1}K^{-1}M^{-9}\big).$$

If instead $\langle \theta_{m,p}^{(t)}, c_k \rangle \geq \max_{m'} \langle \theta_{m',p}^{(t)}, c_k \rangle - 1$ for all $t \leq T_2 - 1$, then by Lemma 23,

$$\langle \theta_m^{(t+1)}, c_k \rangle \leq \langle \theta_m^{(t)}, c_k \rangle - \Theta\bigg(\frac{\eta_r \eta^3 t^3}{PKM^3}\bigg) + \widetilde{O}(\eta_r P \sigma_0^{1.5}). \tag{54}$$

Taking the telescoping sum of (54) from $t = 0$ to $t = T_2 - 1$ yields

$$\begin{aligned}
\langle \theta_{m,p}^{(T_2)}, c_k \rangle &\leq \langle \theta_{m,p}^{(0)}, c_k \rangle - \sum_{s=0}^{T_2-1} \Theta\bigg(\frac{\eta_r \eta^3 s^3}{PKM^3}\bigg) + \widetilde{O}(\eta_r T_2 P \sigma_0^{1.5}) \\
&\overset{(i)}{=} - \sum_{s=0}^{T_2-1} \Theta\bigg(\frac{\eta_r \eta^3 s^3}{PKM^3}\bigg) + \widetilde{O}(P\sigma_0^{1.5}) \\
&\overset{(ii)}{=} -\Theta\bigg(\frac{\eta_r \eta^3 T_2^4}{PKM^3}\bigg) + \widetilde{O}(P\sigma_0^{1.5}) \\
&\leq -\Omega\big(P^{-1}K^{-1}M^{-9}\big),
\end{aligned} \tag{55}$$

where equality (i) is by $\theta_{m,p}^{(0)} = 0$ and the choice of η_r, equality (ii) is by $\sum_{i=0}^{n-1} i^3 = n^2(n-1)^2/4$, and the last inequality is due to $T_2 = \lfloor \eta^{-1} M^{-2} \rfloor$ and $\eta_r = \Theta(M^2)\eta$. Thus $\langle \theta_{m,p}^{(T_2)}, c_k \rangle \leq -\Omega\big(P^{-1}K^{-1}M^{-9}\big)$ for all $m \notin \mathcal{M}_k$. Finally, by Lemma 8,

$$\max_{m' \in [M]} \langle \theta_{m',p'}^{(T_2)}, c_k \rangle \;\geq\; \frac{1}{M} \sum_{m' \in [M]} \langle \theta_{m',p}^{(T_2)}, c_k \rangle \;=\; 0.$$

Therefore,

$$\langle \theta_{m,p}^{(T_2)}, c_k \rangle \leq -\Omega\big(P^{-1}K^{-1}M^{-9}\big) \leq \max_{m' \in [M]} \langle \theta_{m',p}^{(T_2)}, c_k \rangle - \Omega\big(P^{-1}K^{-1}M^{-9}\big),$$

which completes the proof. □

Remark. In prior analyses, T_2 is set to $\lfloor \eta^{-1} M^{-2} \rfloor$ to cancel the factor η_r. In our setting, the explicit bound on M makes this choice unnecessary. Consequently, we establish generalization guarantees up to $T_2 = \widetilde{O}(\eta^{-1})$, rather than only $T_2 = \widetilde{O}(\eta^{-1} M^{-2})$.

Generalization Results

In this section, we will present the detailed proof of Lemma 2, Lemma 3 and Theorem 1 based on the analysis in the previous stages. Given the bound on M in Lemma 14, the proof of Lemma 2 and Theorem 1 are built on Chen et al. (2022).

Lemma 2. *Suppose the weights $\mathbf{W}^{(0)}$ are initialized as in* (2)*. Then, at the end of the exploration stage, with probability at least $1 - o(1)$, the following equations hold for all experts $m \in \mathcal{M}_k$ and any $k, k' \in [K]$ with $k' \neq k$,*

$$\mathbb{P}_{(\mathbf{x},y)\sim\mathcal{D}}\Big(y f_m(\mathbf{x}; \mathbf{W}^{(T_1)}) \leq 0 \mid (\mathbf{x}, y) \in \Omega_k\Big) = o(1),$$

$$\mathbb{P}_{(\mathbf{x},y)\sim\mathcal{D}}\Big(y f_m(\mathbf{x}; \mathbf{W}^{(T_1)}) \leq 0 \mid (\mathbf{x}, y) \in \Omega_{k'}\Big) = \Omega(1/K)$$

Proof. We consider the m-th expert in the MoE layer, suppose that $m \in \mathcal{M}_k$. Then if we draw a new sample $(\mathbf{x}, y) \in \Omega_k$. Suppose p_1, p_2, p_3 are the patches of $\mathbf{x}$ which contain feature signal, feature noise, and cluster-center signal, respectively. Denote Q the set of random noise patches for $\mathbf{x}$, that is $Q = \{q \in [P] : x^{(q)} = \xi_q\}$. By Lemma 15, we have already computed the bound for the inner product between weights and feature signal, cluster-center signal, and feature noise. However, we need to recalculate the bound of the inner product between weights and random noises because we have fresh random noises i.i.d. drawn from $\mathcal{N}(0, (\sigma_q^2/d) \cdot I_d)$. Notice that we use normalized gradient descent for experts with step size η, so we have that

$$\left\|\mathbf{w}_{m,j,p}^{(T_1)} - \mathbf{w}_{m,j,p}^{(0)}\right\|_2 \leq \eta T_1 = O\big(\sigma_0^{0.5}\big)\,.$$

Therefore, by triangle inequality we have that

$$\left\|\mathbf{w}_{m,j,p}^{(T_1)}\right\|_2 \leq \left\|\mathbf{w}_{m,j,p}^{(0)}\right\|_2 + O\big(\sigma_0^{0.5}\big) \leq \widetilde{O}\Big(\sigma_0\sqrt{d}\Big)\,.$$

Because the inner product $\langle \mathbf{w}_{m,j,p}^{(T_1)}, \boldsymbol{\xi}_q \rangle$ follows the distribution $\mathcal{N}\bigg(0,\ (\sigma_q^2/d) \cdot \left\|\mathbf{w}_{m,j,p}^{(T_1)}\right\|_2^2\bigg)$, we have that with probability at least $1 - 1/(dPMJ)$,

$$\big|\langle \mathbf{w}_{m,j,p}^{(T_1)}, \boldsymbol{\xi}_q \rangle\big| = O\Big(\sigma_q d^{-1/2} \left\|\mathbf{w}_{m,j,p}^{(T_1)}\right\|_2 \log(dPMJ)\Big) \leq \widetilde{O}(\sigma_0).$$

Applying Union bound for $m \in [M]$, $j \in [J]$, $q \in Q$ gives that, with probability at least $1 - 1/d$,

$$\big|\langle \mathbf{w}_{m,j,p}^{(T_1)}, \boldsymbol{\xi}_q \rangle\big| = \widetilde{O}(\sigma_0), \quad \forall\, m \in [M],\, j \in [J],\, q \in Q. \tag{56}$$

Now under the event that (56) holds, we have that

$$\begin{aligned}
y f_m(\mathbf{x}, \mathbf{W}^{(T_1)}) &= y \sum_{j \in [J]} \sum_{p \in [P]} \sigma\Big(\big\langle \mathbf{w}_{m,j,p}^{(T_1)}, \mathbf{x}^{(p)} \big\rangle\Big) \\
&= y\, \sigma\Big(\big\langle \mathbf{w}_{m,j_m^*,p_1}^{(T_1)}, \alpha y \mathbf{v}_k \big\rangle\Big) \ + \ y \sum_{(j,p) \neq (j_m^*, p_1)} \sigma\Big(\big\langle \mathbf{w}_{m,j,p}^{(T_1)}, \mathbf{x}^{(p)} \big\rangle\Big) \\
&\geq \alpha^3 \big(1 - \sigma_0^{0.2}\big)^3 \eta^3 (T_1 - T^{(m)})^3 - \widetilde{O}(JPM^3\sigma_0^3) \\
&\geq \alpha^3 \big(1 - 3\sigma_0^{0.2}\big) \widetilde{O}(\sigma_0^{1.5}) - \widetilde{O}(JPM^3\sigma_0^3) \\
&\geq \Omega(\sigma_0^{1.5}),
\end{aligned} \tag{57}$$

where the first inequality is due to Lemma 15, the last inequality is due to the choice of M. Because (56) holds with probability at least $1 - 1/d$, we have shown that

$$\mathbb{P}_{(x,y)\sim\mathcal{D}}\Big(y\, f_m(\mathbf{x};\, \mathbf{W}^{(T_1)}) \leq 0 \,\Big|\, (x,y) \in \Omega_k\Big) \leq \frac{1}{d}.$$

On the other hand, if we draw a new sample $(x, y) \in \Omega_{k'}$, $k' \neq k$. Then we consider the subset $\Omega_{k',k}^- \subseteq \Omega_{k'}$ where the feature noise is $\mathbf{v}_k$ and the sign of the feature noise ϵ is not equal to the label y. Then, under the event that (56) holds, we have that

$$\begin{aligned}
y\, f_m(\mathbf{x}, \mathbf{W}^{(T_1)}) &= y \sum_{j \in [J]} \sum_{p \in [P]} \sigma\Big(\big\langle \mathbf{w}_{m,j,p}^{(T_1)}, \mathbf{x}^{(p)} \big\rangle\Big) \\
&= y\, \sigma\Big(\big\langle \mathbf{w}_{m,j_m^*,p_2}^{(T_1)}, -\gamma y\, \mathbf{v}_k \big\rangle\Big) + y \sum_{(j,p) \neq (j_m^*, p_3)} \sigma\Big(\big\langle \mathbf{w}_{m,j,p}^{(T_1)}, \mathbf{x}^{(p)} \big\rangle\Big) \\
&\leq -\gamma^3 \big(1 - \sigma_0^{0.2}\big)^3 \eta^3 (T_1 - T^{(m)})^3 + \widetilde{O}(JPM^3\sigma_0^3) \\
&\leq -\gamma^3 \big(1 - 3\sigma_0^{0.2}\big)\, \widetilde{O}(\sigma_0^{1.5}) + \widetilde{O}(JPM^3\sigma_0^3) \\
&\leq -\Omega(\sigma_0^{1.5}),
\end{aligned} \tag{58}$$

where the first inequality follows from Lemma 15, the last inequality is due to the choice of M. Because (56) holds with probability at least $1 - 1/d$, we have shown that

$$\mathbb{P}_{(x,y)\sim\mathcal{D}}\Big(y\, f_m(\mathbf{x};\, \mathbf{W}^{(T_1)}) \leq 0 \,\Big|\, (x,y) \in \Omega_{k',k}^-\Big) \geq 1 - \frac{1}{d}.$$

Therefore, we further have that

$$\begin{aligned}
&\mathbb{P}_{(x,y)\sim\mathcal{D}}\Big(y\, f_m(\mathbf{x};\, \mathbf{W}^{(T_1)}) \le 0 \,\Big|\, (x,y)\in\Omega_{k'}\Big) \\
&\ge \mathbb{P}_{(x,y)\sim\mathcal{D}}\Big(y\, f_m(\mathbf{x};\, \mathbf{W}^{(T_1)}) \le 0 \,\Big|\, (x,y)\in\Omega^-_{k',k}\Big)\cdot \mathbb{P}_{(x,y)\sim\mathcal{D}}\Big((x,y)\in\Omega^-_{k',k} \,\Big|\, (x,y)\in\Omega_{k'}\Big) \\
&\ge \Omega\Big(\frac{1}{K}\Big),
\end{aligned} \tag{59}$$

which completes the proof. □

Lemma 3. *Suppose the weights $\mathbf{W}^{(0)}$ are initialized independently from $N(0,\sigma_0^2)$. Then, at the end of the exploration stage, with probability at least $1-o(1)$, the following equation holds for all experts $m\in[M]$ and clusters $k\in[K]$,*

$$\mathbb{P}_{(\mathbf{x},y)\sim\mathcal{D}}\Big(y f_m(\mathbf{x};\mathbf{W}^{(T_1)}) \le 0 \mid (\mathbf{x},y)\in\Omega_k\Big) = \Omega(1/K)$$

Proof. The proof is similar to the proof of Lemma 2. Intuitively, we follow the same analysis in the exploration stage to show that each weight patch within an expert will specialize in different clusters. Given an expert $m\in[M]$ and an input patch $p\in[P]$, denote $(k^*_{m,p}, j^*_{m,p}) = \arg\max_{j,k}\langle \mathbf{v}_k, \mathbf{w}^{(0)}_{m,j,p}\rangle$. Due to independent initialization we have

$$\mathbb{P}(k^*_{m,p}=k) = \frac{1}{K},\ \forall\, k\in[K]. \tag{60}$$

By the same arguments as in Lemma 6 we have the following inequalities

$$\begin{aligned}
&\max_{(j,k)\neq(j^*_{m,p},k^*_{m,p})} \langle \mathbf{w}^{(0)}_{m,j,p}, \mathbf{v}_k\rangle \le \left(1-\frac{\delta}{3MJ^2K^2}\right)\langle \mathbf{w}_{m,j^*_{m,p},p}, \mathbf{v}_{k^*_{m,p}}\rangle \\
&\langle \mathbf{w}_{m,j^*_{m,p},p}, \mathbf{v}_{k^*_{m,p}}\rangle \ge 0.01\sigma_0
\end{aligned} \tag{61}$$

Now, for all $t\le T_1$, following the same arguments as Lemma 13 to Lemma 15 gives us

$$\begin{aligned}
\langle \mathbf{w}^{(t)}_{m,j,p}, \mathbf{v}_k\rangle &= \widetilde{O}(M\sigma_0),\ \forall (j,k)\neq(j^*_m,k^*_m), \\
\langle \mathbf{w}^{(t)}_{m,j,p}, \mathbf{c}_k\rangle &= \widetilde{O}(\sigma_0),\ \forall j\in[J], k\in[K], \\
\langle \mathbf{w}^{(t)}_{m,j,p}, \xi_{i,q}\rangle &= \widetilde{O}(\sigma_0),\ \forall j\in[J], i\in[n], q\in Q_i.
\end{aligned} \tag{62}$$

Besides,

$$\langle \mathbf{w}^{(t)}_{m,j^*_m,p}, \mathbf{v}^*_{k_m}\rangle \ge (1-\sigma_0^{0.2})\eta(t-T^{(m)}),\quad \forall t\ge T^{(m)}. \tag{63}$$

Now we are ready to prove the lemma. Draw a new sample $(\mathbf{x},y)\in\Omega_k$. By Lemma 15 and Lemma 2, we have computed the bound for the inner products between weights and feature signal, cluster-center signal, feature noise, as well as fresh random noise. In particular, the following equation holds with probability at least $1-1/d$

$$\big|\langle \mathbf{w}^{(T_1)}_{m,j,p},\, \boldsymbol{\xi}_q\rangle\big| = \widetilde{O}(\sigma_0),\quad \forall\, m\in[M],\, j\in[J],\, q\in Q. \tag{64}$$

Note that the feature signal of $\mathbf{x}$ can be in any patch $p_1\in[P]$ chosen uniformly at random. Therefore, combining with (60) we have $\mathbb{P}(k^*_{m,p_1}=k') = 1/K,\ \forall\, k'\in[K]$. In particular, we have

$$\mathbb{P}(k\neq k^*_{m,p_1} \text{ and } k\neq k^*_{m,p_2}) = (1-1/K)^2. \tag{65}$$

When $k^*_{m,p_1}\neq k$ and $k^*_{m,p_1}\neq k^*_{m,p_2}$, if we draw $(\mathbf{x},y)\in\Omega^-_{k,k^*_{m,p_2}}$, that is, the feature noise is $\mathbf{v}_{k^*_{m,p_2}}$ and the sign of the feature noise ϵ is not equal to the label y. Then, under the event that (64) holds, we have that

$$\begin{aligned}
y\, f_m(\mathbf{x},\mathbf{W}^{(T_1)}) &= y\sum_{j\in[J]}\sum_{p\in[P]}\sigma\Big(\big\langle \mathbf{w}^{(T_1)}_{m,j,p},\, \mathbf{x}^{(p)}\big\rangle\Big) \\
&= y\,\sigma\Big(\big\langle \mathbf{w}^{(T_1)}_{m,j^*_{m,p_2},p_2},\, -\gamma y\, \mathbf{v}_{k^*_{m,p_2}}\big\rangle\Big) + y\sum_{(j,p)\neq(j^*_{m,p_2},p_2)}\sigma\Big(\big\langle \mathbf{w}^{(T_1)}_{m,j,p},\, \mathbf{x}^{(p)}\big\rangle\Big) \\
&\le -\gamma^3\big(1-\sigma_0^{0.2}\big)^3\eta^3(T_1-T^{(m)})^3 + \widetilde{O}(JPM^3\sigma_0^3) \\
&\le -\gamma^3\big(1-3\sigma_0^{0.2}\big)\,\widetilde{O}(\sigma_0^{1.5}) + \widetilde{O}(JPM^3\sigma_0^3) \\
&\le -\Omega(\sigma_0^{1.5}),
\end{aligned} \tag{66}$$

where the first inequality follows from Lemma 15, the last inequality is due to the choice of M. Because (64) holds with probability at least $1 - 1/d$, we have shown that

$$\mathbb{P}_{(x,y)\sim\mathcal{D}}\Big(y\, f_m(\mathbf{x};\, \mathbf{W}^{(T_1)}) \le 0 \,\Big|\, (x,y) \in \Omega^-_{k,k^*_{m,p_2}} \text{ and } k \neq k^*_{m,p_1}\Big) \ge 1 - \frac{1}{d}.$$

Therefore, we further have that

$$\begin{aligned}
&\mathbb{P}_{(x,y)\sim\mathcal{D}}\Big(y\, f_m(\mathbf{x};\, \mathbf{W}^{(T_1)}) \le 0 \,\Big|\, (x,y) \in \Omega_k\Big) \\
&\ge \mathbb{P}_{(x,y)\sim\mathcal{D}}\Big(y\, f_m(\mathbf{x};\, \mathbf{W}^{(T_1)}) \le 0 \,\Big|\, (x,y) \in \Omega^-_{k,k^*_{m,p_2}} \text{ and } k \neq k^*_{m,p_1}\Big) \cdot \mathbb{P}_{(x,y)\sim\mathcal{D}}\Big((x,y) \in \Omega^-_{k,k^*_{m,p_2}} \text{ and } k \neq k^*_{m,p_1} \,\Big|\, (x,y) \in \Omega_k\Big) \\
&\ge \Omega\Big(\frac{1}{K}\Big),
\end{aligned} \tag{67}$$

which completes the proof. □

Theorem 1. *Suppose the training data size is $n = \Omega(d)$. Choose experts number $M \in [\Theta(K \log K \log\log d), \Theta(\sigma_0^{-0.075})]$, filter size $J = \Theta(\log M \log\log d)$, initialization scale $\sigma_0 \in [d^{-1/3}, d^{-0.01}]$, learning rate $\eta = \widetilde{O}(\sigma_0)$, $\eta_r = \Theta(M^2)\eta$. Then with probability at least $1 - o(1)$, Algorithm 1 is able to output $(\Theta^{(T)}, \mathbf{W}^{(T)})$ within $T = \widetilde{O}(\eta^{-1})$ iterations such that the non-linear MoE defined in Section* ***New Architecture for MoE with MLP Experts*** *satisfies*

- *Training error is zero, i.e., $y_i F(x_i; \Theta^{(T)}, \mathbf{W}^{(T)}) > 0,\ \forall i \in [n]$.*
- *Test error is nearly zero, i.e., $\mathbb{P}_{(x,y)\sim\mathcal{D}}\big(yF(x; \Theta^{(T)}, \mathbf{W}^{(T)}) \le 0\big) = o(1)$.*

More importantly, the experts can be divided into a disjoint union of K non-empty sets $[M] = \bigsqcup_{k\in[K]} \mathcal{M}_k$ and

- *(Each expert is good on one cluster) Each expert $m \in \mathcal{M}_k$ performs well on the cluster Ω_k, $\mathbb{P}_{(x,y)\sim\mathcal{D}}\big(yf_m(x; \mathbf{W}^{(T)}) \le 0 \,\big|\, (x,y) \in \Omega_k\big) = o(1)$.*
- *(Router only distributes example to good expert) With probability at least $1 - o(1)$, an example $x \in \Omega_k$ will be routed to one of the experts in $\mathcal{M}_k$.*

Proof. We will give the proof for $T = T_2$, i.e., at the end of the second stage.

Test Error is small. We first prove the following result for the experts. For all expert $m \in \mathcal{M}_k$, we have that

$$\mathbb{P}_{(\mathbf{x},y)\sim\mathcal{D}}\big(yf_m(\mathbf{x}; \mathbf{W}^{(T)}) \le 0 \,\big|\, (\mathbf{x},y) \in \Omega_k\big) \;=\; o(1). \tag{68}$$

The proof is similar to the proof of Lemma 5.2. We consider the m-th expert in the MoE layer, suppose that $m \in \mathcal{M}_k$. Then if we draw a new sample $(\mathbf{x}, y) \in \Omega_k$.

By Lemma 18, we have already got the bound for the inner product between weights and feature signal, cluster–center signal, and feature noise. However, we need to recalculate the bound of the inner product between weights and random noises because we have fresh random noises i.i.d. drawn from $\mathcal{N}\big(0,\ (\sigma_p^2/d)\cdot\mathbf{I}_d\big)$. Notice that we use normalized gradient descent with step size η, so we have that

$$\big\|\mathbf{w}^{(T)}_{m,j,p} - \mathbf{w}^{(0)}_{m,j,p}\big\|_2 \;\le\; \eta T \;=\; \widetilde{O}(1). \tag{69}$$

Therefore, by triangle inequality, we have that

$$\big\|\mathbf{w}^{(T)}_{m,j,p}\big\|_2 \;\le\; \big\|\mathbf{w}^{(0)}_{m,j,p}\big\|_2 + \widetilde{O}(1) \;\le\; \widetilde{O}(\sigma_0\sqrt{d}). \tag{70}$$

Because the inner product $\langle \mathbf{w}^{(T)}_{m,j,p}, \boldsymbol{\xi}_q\rangle$ follows the distribution $\mathcal{N}\Big(0,\ (\sigma_p^2/d)\cdot\big\|\mathbf{w}^{(T)}_{m,j,p}\big\|_2^2\Big)$, with probability at least $1 - 1/(dPMJ)$ we have that

$$\big|\langle \mathbf{w}^{(T)}_{m,j,p}, \boldsymbol{\xi}_q\rangle\big| = O\Big(\sigma_p d^{-1/2}\big\|\mathbf{w}^{(t)}_{m,j,p}\big\|_2 \log(dPMJ)\Big) \le \widetilde{O}(\sigma_0). \tag{71}$$

Applying union bound for $m \in [M], j \in [J], q \in Q$ gives that, with probability at least $1 - 1/d$,

$$\big|\langle \mathbf{w}^{(T)}_{m,j,p}, \boldsymbol{\xi}_q\rangle\big| \;=\; \widetilde{O}(\sigma_0), \quad \forall\, m \in [M],\, j \in [J],\, q \in Q. \tag{72}$$

Now, under the event that (72) holds, we have that

$$
\begin{aligned}
y f_m(\mathbf{x}, \mathbf{W}^{(T)}) &= y \sum_{j\in[J]} \sum_{p\in[P]} \sigma\Big(\big\langle \mathbf{w}_{m,j,p}^{(T)}, \mathbf{x}^{(p)} \big\rangle\Big) \\
&= y\,\sigma\Big(\big\langle \mathbf{w}_{m,j_m^*,p_1}^{(T)}, \alpha y \mathbf{v}_k \big\rangle\Big) \;+\; y \sum_{(j,p)\neq(j_m^*,p_1)} \sigma\Big(\big\langle \mathbf{w}_{m,j,p}^{(T)}, \mathbf{x}^{(p)} \big\rangle\Big) \\
&\geq \alpha^3\big(1-\sigma_0^{0.2}\big)^3\eta^3(T-T^{(m)})^3 - \widetilde{O}(JPM^3\sigma_0^3) \\
&\geq \alpha^3\big(1-3\sigma_0^{0.2}\big) - \widetilde{O}(JPM^3\sigma_0^3) \\
&\geq \Omega(1),
\end{aligned} \tag{73}
$$

where the first inequality is by Lemma 18, the last inequality is due to the choice of M. Because (72) holds with probability at least $1-1/d$, we have prove that

$$
\mathbb{P}_{(\mathbf{x},y)\sim\mathcal{D}}\big(y f_m(\mathbf{x}; \mathbf{W}^{(T)}) \leq 0 \,\big|\, (\mathbf{x},y)\in\Omega_k\big) \leq \frac{1}{d}. \tag{74}
$$

Next we show that, with probability at least $1-o(1)$, an example $\mathbf{x}\in\Omega_k$ will be routed to one of the experts in $\mathcal{M}_k$. For any $\mathbf{x}\in\Omega_k$ and any $m\notin\mathcal{M}_k$, we need to check that $h_m(\mathbf{x};\Theta^{(T)}) < \max_{m'} h_{m'}(\mathbf{x};\Theta^{(T)})$. Indeed, we have

$$
\begin{aligned}
h_m(\mathbf{x};\Theta^{(T)}) - \max_{m'} h_{m'}(\mathbf{x};\Theta^{(T)}) &= h_m(\overline{\mathbf{x}};\Theta^{(T)}) - \max_{m'} h_{m'}(\overline{\mathbf{x}};\Theta^{(T)}) + \widetilde{O}(M^2P^2K^2\sigma_0^{2.5}) \\
&= \beta\langle\Theta_{m,p_3}^{(T)}, \mathbf{c}_k\rangle - \beta\max_{m'}\langle\Theta_{m',p_3}^{(T)}, \mathbf{c}_k\rangle + \widetilde{O}(M^2P^2K^2\sigma_0^{2.5}) \\
&\leq -\beta\,\Omega(P^{-1}K^{-1}M^{-9}) + \widetilde{O}(M^2P^2K^2\sigma_0^{1.5}) < 0,
\end{aligned} \tag{75}
$$

where the first equality is due to Lemma 20, the first inequality is due to Lemma 24. Because the lemmas hold with probability at least $1-1/d$, this completes our proof.

Training Error is zero . The proof for training error is much easier, because we no longer need to deal with the fresh noises, and we no longer need to use high-probability bounds for those inner products with fresh noises. That's the reason we can get exactly zero training error. We first prove the following result for the experts. For all expert $m\in\mathcal{M}_k$, we have that

$$
y_i f_m(\mathbf{x}_i; \mathbf{W}^{(T)}) \leq 0, \quad \forall\, i\in\Omega_k. \tag{76}
$$

By Lemma 18, we have that for all $i\in\Omega_k$

$$
\begin{aligned}
y f_m(\mathbf{x}_i, \mathbf{W}^{(T)}) &= y_i \sum_{j\in[J]} \sum_{p\in[P]} \sigma\Big(\big\langle \mathbf{w}_{m,j,p}^{(T)}, \mathbf{x}_i^{(p)} \big\rangle\Big) \\
&= y_i\,\sigma\Big(\big\langle \mathbf{w}_{m,j_m^*,p_1}^{(T)}, \alpha_i y_i \mathbf{v}_k \big\rangle\Big) \;+\; y_i \sum_{(j,p)\neq(j_m^*,p_1)} \sigma\Big(\big\langle \mathbf{w}_{m,j,p}^{(T)}, \mathbf{x}_i^{(p)} \big\rangle\Big) \\
&\geq \alpha_i^3\big(1-\sigma_0^{0.2}\big)^3\eta^3(T-T^{(m)})^3 - \widetilde{O}(JPM^3\sigma_0^3) \\
&\geq \alpha_i^3\big(1-3\sigma_0^{0.2}\big) - \widetilde{O}(JPM^3\sigma_0^3) \\
&> 0,
\end{aligned} \tag{77}
$$

where the first inequality is due to Lemma 18, the last inequality is due to the choice of M. We then show that an example $(\mathbf{x}_i, y_i)\in\Omega$ is routed to one of the experts in $\mathcal{M}_k$. Suppose the m-th expert is not in $\mathcal{M}_k$. We only need to check the value of $h_m(\mathbf{x}_i;\Theta^{(T)}) < \max_{m'} h_{m'}(\mathbf{x}_i;\Theta^{(T)})$, which is straightforward by Lemma 24 and Lemma 20.

□

Auxiliary Lemmas

In this section, we present several auxiliary results that we will use throughout our analysis.

The following lemma shows that when two gating-network outputs are close, the router distributes examples to the corresponding experts with nearly the same probability.

Lemma 25 (Chen et al. (2022), Lemma C.3)**.** *Let* $\mathbf{h}\in\mathbb{R}^M$ *be the output of the gating network and let* $\{r_m\}_{m=1}^M$ *be noise drawn independently from* $\mathrm{Unif}[0,1]$. *Define the routing probabilities* $p_m \coloneqq \mathbb{P}\big(\arg\max_{m'}\{h_{m'}+r_{m'}\}=m\big)$. *Then for any* $m,m'\in[M]$,

$$
|p_m - p_{m'}| \;\leq\; M^2\,|h_m - h_{m'}|\,.
$$

Proof. Construct $\widehat{\mathbf{h}}$ as a copy of $\mathbf{h}$ and permute its m-th and m'-th entries. Let $\widehat{\mathbf{p}}$ be the corresponding probability vector induced by $\widehat{\mathbf{h}}$. Then $|p_m - p_{m'}| = \|\mathbf{p} - \widehat{\mathbf{p}}\|_\infty$ and $|h_m - h_{m'}| = \|\widehat{\mathbf{h}} - \mathbf{h}\|_\infty$. Applying Lemma 5.1 completes the proof. □

The next lemma shows that the router will not route examples to experts whose gating outputs are sufficiently small, which both saves computation and improves performance.

Lemma 26 (Chen et al. (2022), Lemma C.4)**.** *Suppose the noise $\{r_m\}_{m=1}^M$ is drawn independently from* $\mathrm{Unif}[0,1]$*. If an expert m satisfies*

$$h_m(\mathbf{x};\mathbf{\Theta}) \;\le\; \max_{m'} h_{m'}(\mathbf{x};\mathbf{\Theta}) - 1,$$

then the example $\mathbf{x}$ will not be routed to expert m.

Proof. From $h_m(\mathbf{x};\mathbf{\Theta}) \le \max_{m'} h_{m'}(\mathbf{x};\mathbf{\Theta}) - 1$, for any realization of the uniform noises $\{r_{m'}\}_{m'\in[M]}$ we have

$$h_m(\mathbf{x};\mathbf{\Theta}) + r_m \;\le\; \max_{m'} h_{m'}(\mathbf{x};\mathbf{\Theta}) \;\le\; \max_{m'} \big\{h_{m'}(\mathbf{x};\mathbf{\Theta}) + r_{m'}\big\},$$

where the first inequality uses $r_m \le 1$ and the second uses $r_{m'} \ge 0$ for all $m' \in [M]$. Thus m cannot be the maximizer, so $\mathbf{x}$ is not routed to expert m. □

Lemma 27 (Allen-Zhu and Li (2020), Lemma C.19)**.** *Let $\{x_t, y_t\}_{t=1,\dots}$ be two positive sequences that satisfy*

$$x_{t+1} \ge x_t + \eta \cdot C_t x_t^2, \qquad y_{t+1} \le y_t + S\eta \cdot C_t y_t^2,$$

and $|x_{t+1} - x_t|^2 + |y_{t+1} - y_t|^2 \le \eta^2$. Suppose $x_0, y_0 = o(1)$, $x_0 \ge y_0 S(1+G)$, and

$$\eta \le \min\left\{\frac{G^2 x_0}{\log(A/x_0)},\ \frac{G^2 y_0}{\log(1/G)}\right\}.$$

Then we have for all $A > x_0$, let T_x be the first iteration such that $x_t \ge A$, then we have

$$y_{T_x} \le O(y_0 G^{-1}).$$

Proof. We only need to replace $O(\eta A^{q-1})$ in the proof of Lemma C.19 by $O(\eta)$, because we use normalized gradient descent, i.e., $C_t x_t^2 \le 1$. For completeness, we present the whole proof here.

For all $g = 0, 1, 2, \dots$, let T_g be the first iteration such that $x_t \ge (1+\delta)^g x_0$, let b be the smallest integer such that $(1+\delta)^b x_0 \ge A$. For simplicity of notation, we replace x_t with A whenever $x_t \ge A$. Then by the definition of T_g, we have that

$$\sum_{t\in[T_g, T_{g+1})} \eta C_t [(1+\delta)^g x_0]^2 \le x_{T_{g+1}} - x_{T_g} \le \delta(1+\delta)^g x_0 + O(\eta),$$

where the last inequality holds because we are using normalized gradient descent, i.e., $\max_t |x_{t+1} - x_t| \le \eta$. This implies that

$$\sum_{t\in[T_g, T_{g+1})} \eta C_t \le \frac{\delta}{(1+\delta)^g x_0} + \frac{O(\eta)}{x_0^2}.$$

Recall that b is the smallest integer such that $(1+\delta)^b x_0 \ge A$, so we can calculate

$$\sum_{t\ge 0, x_t \le A} \eta C_t \le \left[\sum_{g=0}^{b-1} \frac{\delta}{(1+\delta)^g x_0}\right] + \frac{O(\eta) b}{x_0^2} = \frac{1+\delta}{x_0} + \frac{O(\eta) b}{x_0^2} \le \frac{1+\delta}{x_0} + \frac{O(\eta)\log(A/x_0)}{x_0^2 \log(1+\delta)}.$$

Let T_x be the first iteration t in which $x_t \ge A$. Then we have that

$$\sum_{t=0}^{T_x} \eta C_t \le \frac{1+\delta}{x_0} + \frac{O(\eta)\log(A/x_0)}{\delta x_0^2}. \tag{78}$$

On the other hand, let $A' = G^{-1} y_0$ and b' be the smallest integer such that $(1+\delta)^{b'} y_0 \ge A'$. For simplicity of notation, we replace y_t with A' when $y_t \ge A'$. Then let T'_g be the first iteration such that $y_t \ge (1+\delta)^g y_0$, then we have that

$$\sum_{t\in[T'_g, T'_{g+1})} \eta S C_t [(1+\delta)^g y_0]^{q-1} \ge y_{T'_{g+1}} - y_{T'_g} \ge \delta(1+\delta)^g y_0 - O(\eta).$$

Therefore, we have that

$$\sum_{t\in[T'_g, T'_{g+1})} S\eta C_t \ge \frac{\delta}{(1+\delta)^g (1+\delta)^2 y_0} - \frac{O(\eta)}{y_0^2}.$$

Recall that b' is the smallest integer such that $(1+\delta)^{b'} y_0 \geq A'$. We then have that

$$\sum_{t \geq 0, x_t \leq A} \eta S C_t \geq \sum_{g=0}^{b'-2} \frac{\delta}{(1+\delta)^g (1+\delta)^2 y_0} - \frac{O(\eta) b'}{y_0^2}.$$

Let T_y be the first iteration t in which $y_t \geq A'$, so we can calculate

$$\sum_{t=0}^{T_y} \eta S C_t \geq \frac{1 - O(\delta + G)}{y_0} - \frac{O(\eta) \log(A'/y_0)}{y_0^2 \delta}. \tag{79}$$

Compare (78) and (79). Choosing $\delta = G$ and $\eta \leq \min\{\frac{G^2 x_0}{\log(A/x_0)}, \frac{G^2 y_0}{\log(1/G)}\}$, together with $x_0 \geq y_0 S(1+G)$, we obtain the result.

□