Bridging Game Theory and Transformer Routing: Mean Field Equilibria for Mixture of Experts

Nevroz Sen NEVROZ.SEN@MAIL.MCGILL.CA
C3 AI

Abstract

We consider the problem of routing tokens to experts in Mixture-of-Experts (MoE) models under finite expert capacity. We formulate this as a mean field game in which tokens compete for expert resources through an aggregate load distribution. This formulation yields an entropy-regularized equilibrium routing rule that jointly accounts for expert quality and congestion. For linear congestion costs, we derive a closed-form best response converging in $O(\log 1/\varepsilon)$ iterations, establish existence and uniqueness, and prove $O(1/\sqrt{N} + \sqrt{N/T})$ finite-expert approximation bounds. We empirically evaluate the resulting Capacity-Aware MFG router on WikiText-103 and heterogeneous domain mixtures, where it improves perplexity relative to standard sparse MoE routing baselines while eliminating irreversible token drops, at a $1.65\times$ training-cost overhead due primarily to dense expert execution. The gain decomposes into a dense-routing effect and an additional equilibrium-routing margin over Dense Random. With sparse-aware training, the learned dense equilibrium can be projected to Top-1 routing with a relatively modest 2.7% gain at identical inference cost.

1. Introduction

With state-of-the-art language models now exceeding hundreds of billions of parameters [1, 2], a fundamental conflict between model capacity and computational tractability merits more attention. While scaling improves performance, the resource requirements for training and inference grow with dense architectures. The Mixture of Experts (MoE) architecture [6] addresses this challenge through conditional computation. By activating only a sparse subset of parameters (expert networks) for each input token, MoE enables models to scale to trillions of parameters while maintaining constant computational cost per inference [18]. This efficiency has led to the widespread adoption of MoE in modern large language models, including Mixtral [7] and DeepSeek-V2 [3], which achieve competitive performance with a fraction of the active parameters of their dense counterparts. Despite this empirical success, the mechanisms that govern expert routing in MoE models remain less well characterized theoretically. State-of-the-art approaches, such as the Switch Transformer [4], rely on design choices such as load balance losses, capacity constraints, or token dropping to prevent expert collapse and ensure computational efficiency. While these techniques are effective, they offer limited theoretical insight and provide a limited notion of what an optimal routing strategy should look like. This gap between practical performance and theoretical understanding makes it difficult to reason about scalability, robustness, and the fundamental trade-offs inherent in MoE routing. With this motivation, we reformulate MoE routing as a competitive resource allocation problem. We model tokens as strategic agents in a Mean Field Game (MFG) [5, 8], where each token seeks to maximize routing quality while accounting for congestion induced by other tokens. In this formulation, expert load balancing arises endogenously through strategic interaction,

and routing corresponds to a Nash equilibrium, providing a complementary equilibrium-based interpretation of load balancing. Our contribution is threefold: **(1)** a formal MFG formulation with existence, uniqueness, and exponential convergence guarantees; **(2)** $O(1/\sqrt{N} + \sqrt{N/T})$ approximation bounds via a joint (N, T) analysis, where the growing discrete action space of MoE experts has no analogue in classical MFG theory; and **(3)** empirical validation separating the effects of dense activation, equilibrium-based coordination, and sparse projection, with 9–12% improvement over sparse routing baselines and 2.7% gain at identical sparse inference cost.

2. MoE Routing as a Mean Field Game

A standard MoE layer routes input x_t to N experts via weights $p_t \in \Delta^N$, outputting $\sum_i p_t(i) \cdot \text{Expert}_i(x_t)$. The Switch Transformer [4] uses Top-1 selection subject to capacity $C \cdot (T/N)$; tokens routed to full experts are dropped. This capacity conflict is fundamentally a congestion problem: tokens compete for limited expert slots, and the system must balance individual routing quality against collective load. Mean field games (MFG) [5, 8] provide a natural framework for exactly this type of problem, analyzing equilibria in large populations where agents interact through an aggregate statistic rather than individually. MoE routing satisfies the key MFG conditions: batch sizes ($T \sim 10^4$) ensure negligible individual impact, capacity constraints create mean-field coupling, and tokens solve identical problems with heterogeneous preferences. Now, consider a batch of T tokens and N experts. Each token t has quality scores $q_t = W_q x_t \in \mathbb{R}^N$ (via a learned gate $W_q \in \mathbb{R}^{N \times d}$) and routing distribution $p_t \in \Delta^N$. The aggregate load is $\rho = \frac{1}{T} \sum_t p_t \in \Delta^N$. Each token minimizes $J_t(p_t|\rho) = -\langle q_t, p_t \rangle + \langle C(\rho), p_t \rangle$, trading off expert quality against congestion. We consider two congestion models: **linear cost** $C(\rho)_i = \lambda \rho_i$ (primary theoretical model) and **capacity-aware cost** $C(\rho)_i = \lambda \max(0, \rho_i - \pi_{\text{lim}})$ where $\pi_{\text{lim}} = C_{\text{factor}}/N$ (strict capacity enforcement). Our problem satisfies the **large population** and **anonymity** conditions, allowing us to invoke the Nash Certainty Equivalence principle [5].

Definition 1 (Mean Field Equilibrium) *A pair (ρ^*, p^*) comprising a population distribution $\rho^* \in \Delta^N$ and a policy $p^*(\cdot|\rho; q) : \Delta^N \times \mathbb{R}^N \to \Delta^N$ constitutes a Mean Field Equilibrium if: (i) **Individual Optimality:** Given mean field ρ^*, the policy minimizes each agent's cost: $p^*(\cdot|\rho^*; q) = \arg\min_{p \in \Delta^N} J(p|\rho^*; q)$, and (ii) **Population Consistency:** The mean field arises from aggregating optimal responses: $\rho^* = \mathbb{E}_{q \sim \mathcal{D}}[p^*(\cdot|\rho^*; q)]$ where $\mathcal{D}$ is the distribution of quality scores.*

In practice, with a finite batch of T tokens, consistency becomes $\rho^* = \frac{1}{T} \sum_{t=1}^{T} p_t^*(\rho^*)$, reducing $O(T^2)$ dependencies to $O(T)$ computations against a shared mean field.

Proposition 2 (Best Response) *Under entropy regularization $\tilde{J}_t(p_t|\rho) = J_t(p_t|\rho) + \mathcal{H}(p)$ where $\mathcal{H}(p) = \frac{1}{\beta} \sum_i p_t(i) \log p_t(i)$ with inverse temperature $\beta > 0$, the unique minimizer is the softmax policy $p_t(i|\rho) = \exp(\beta[q_t(i) - C(\rho)_i]) / \sum_j \exp(\beta[q_t(j) - C(\rho)_j])$. As $\beta \to \infty$, routing concentrates on $i^* = \arg\max_i [q_t(i) - C(\rho)_i]$.*

3. Theoretical Analysis

We state our main results; proofs are in the Appendix.

Definition 3 (Assumption) *(A1) **Bounded quality**: $|q_t(i)| \leq Q_{\max} < \infty$ (satisfied via Layer Normalization), (A2) **Lipschitz congestion**: $\|C(\rho_1) - C(\rho_2)\|_\infty \leq L_C \|\rho_1 - \rho_2\|_1$ (holds with*

$L_C = \lambda$ for linear congestion), (A3) ***Strict Monotonicity****: $\langle C(\rho_1) - C(\rho_2), \rho_1 - \rho_2 \rangle > 0$ for $\rho_1 \neq \rho_2$ (satisfied for $\lambda > 0$), (A4)* ***Expert diversity****: $\mathrm{Var}_i[q_t(i)] \geq \sigma_q^2 > 0$ (empirically verified).*

Theorem 4 (Existence and Uniqueness) *Under Assumptions (A1)–(A2), there exists a Nash equilibrium $\rho^* \in \Delta^N$. Under (A3), ρ^* is unique.*

Theorem 5 (Convergence) *Let $C(\rho) = \lambda\rho$. The equilibrium is the unique fixed point $\rho^* = T(\rho^*)$ of the aggregate response map $T(\rho)_i = \frac{1}{T}\sum_{t=1}^{T} \frac{\exp(\beta[q_t(i) - \lambda\rho_i])}{\sum_j \exp(\beta[q_t(j) - \lambda\rho_j])}$. For $\beta\lambda < 2$, T is a contraction: $\|\rho^{(k)} - \rho^*\|_1 \leq (\beta\lambda/2)^k \|\rho^{(0)} - \rho^*\|_1$.*

Remark 6 *Our experiments use $\beta\lambda = 10$, beyond the contraction regime. Convergence is supported by directional contraction (effective $\kappa_{\text{eff}} \approx 2\beta\lambda/N \ll 1$, Proposition 8), potential game structure [13] guaranteeing convergence via monotone improvement of the global potential $\Phi(\rho) = \sum_i \int_0^{\rho_i} C(s)\, ds - \frac{1}{\beta T}\sum_t \log Z_t(\rho)$, and empirical verification (≈13 iterations up to $\beta\lambda = 25$). The λ-sensitivity sweep (Table 2) confirms that $\beta\lambda = 1.0$ (within the contractive regime) achieves 33.79 PPL, virtually identical to $\beta\lambda = 10$ (33.78).*

Under an additional assumption that the N expert parameters are drawn i.i.d. from a continuous distribution π over a compact space (A5), we bound the finite-population error:

Theorem 7 (Finite-Expert Approximation) *Under (A1)–(A3) and (A5): $\mathbb{E}[\|\rho^{N,T} - \rho^*\|_1] \leq O(1/\sqrt{N} + \sqrt{N}/\sqrt{T})$.*

We remark that unlike standard MFG theory, the growing discrete action space requires a joint (N, T) analysis. The optimal $N^* \asymp \sqrt{T}$ balances both terms at $O(T^{-1/4})$. Also, after K iterations, the routing satisfies an ε-Nash equilibrium for $K = O(\log 1/\varepsilon)$ (proof in Appendix C.5). Finally, for the **capacity-aware cost** $C(\rho)_i = \lambda \max(0, \rho_i - \pi_{\text{lim}})$ used in experiments, the equilibrium satisfies the same fixed-point structure with rectified penalties. The contraction becomes *directional*:

Proposition 8 (Contraction for Capacity-Aware Costs) *Let $\mathcal{S}(\rho) = \{j : \rho_j > \pi_{\text{lim}}\}$. Then $\|T(\rho) - T(\rho')\|_1 \leq \frac{\beta\lambda}{2}\|P_{\mathcal{S}_\cup}(\rho - \rho')\|_1$ where $\mathcal{S}_\cup$ is the union of active sets along the path. The effective contraction constant is $\kappa_{\text{eff}} \approx 2\beta\lambda/N$, contractive for $N > 2\beta\lambda$ (e.g., $\kappa_{\text{eff}} \approx 0.63$ for $\beta\lambda = 10$, $N = 32$). The solver cannot amplify perturbations in the uncongested subspace, reducing effective dynamics to $|\mathcal{S}|$ dimensions.*

4. Experiments

Algorithm. Algorithm 1 (Appendix) solves for the Nash equilibrium via fixed-point iteration with momentum (μ=0.5) and curriculum learning on λ (ramping $\lambda_{\text{min}} \to \lambda_{\text{max}}$); β=1.0 is held constant. The solver costs $O(KTN)$ where $K \approx 13$; the dominant cost is soft expert execution $O(TNd_{\text{ff}})$ vs $O(Td_{\text{ff}})$ for sparse baselines. The training loss is $\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{LM}} + \alpha(s)\sum_i \max(0, \bar{p}_i - \pi_{\text{lim}}) - \gamma\mathcal{L}_{\text{entropy}}$, where $\alpha(s) \propto \lambda$ grows with the congestion penalty.

Setup. We evaluate on WikiText-103 [12] (103M tokens) with a Transformer MoE (d_{model}=384, 6 layers, N=32, C=1.5×), trained with AdamW on A100 GPUs. We use Protocol A (1 epoch) for ablations and Protocol B (2 epochs) for benchmarks. We report **PPL**, **Token Drops** (irreversible, sparse methods), and **Capacity Overflow** (redistributed, soft methods).

Table 1: **WikiText-103 benchmarks** (N=32, C=1.5×). Cap-Aware MFG (33.78) improves over Dense Random (34.63) under the same dense routing regime. Token Drops = irreversible; Cap. Overflow = redistributed, no information loss.

Method	Routing Type	PPL ↓	Token Drops (%) ↓	Cap. Overflow (%) ↓
Hash Routing	Sparse (Top-1)	68.07	**0.0**	—
Expert Choice	Sparse (Inv. Top-k)	45.41	**0.0**	—
BASE Layers	Sparse (Balanced)	40.71	**0.0**	—
Switch Transformer	Sparse (Top-1)	38.54	11.7	—
Top-2 Gating	Sparse (Top-2)	37.79	66.6	—
Top-4 Gating	Sparse (Top-4)	35.29	47.4	—
Dense Random	Soft (Dense)	34.63	**0.0**	**0.0**
Cap-Aware MFG† (Ours)	Soft (Dense)	33.79	**0.0**	4.1
Soft MFG (Ours)	Soft (Dense)	33.99	**0.0**	6.6
Cap-Aware MFG (Ours)	Soft (Dense)	**33.78**	**0.0**	2.5

† With $\beta\lambda = 1.0 < 2$, within the provably contractive regime of Theorem 5.

Main Results. Table 1 compares the MFG Router against Switch [4], Expert Choice [20], BASE Layers [10], Hash Routing [16], Top-k (k=2, 4), and Dense Random on WikiText-103.

Sparse Baselines Comparison. Our experiments separate dense activation, equilibrium-based coordination, and sparse projection. Cap-Aware MFG achieves the lowest perplexity (33.78), improving over Switch (38.54) by 12% with zero token drops. Dense Random (34.63) also outperforms Switch, showing that avoiding irreversible token drops explains much of the headline gain. However, MFG further improves over Dense Random, while Top-4, BASE, and Expert Choice do not. This suggests that the remaining improvement is associated with the equilibrium-based coordination mechanism rather than dense activation alone. The gain decomposes into *dense activation* (38.54→34.63, ∼82%) and *equilibrium routing* (34.63→33.78, ∼18%). Thus, we interpret the 12% Switch improvement as a combined dense-routing/equilibrium-routing effect. Table 4 tests whether this margin grows under domain heterogeneity, while Table 3 evaluates sparse projection.

Sensitivity to λ. Table 2 sweeps $\lambda \in \{1, 5, 10, 15, 25\}$ at β=1.0 plus β=2.0, λ=0.99. PPL varies by only 0.19 points (33.78–33.97) across the 25× range, while all settings outperform Dense Random (34.63) and Switch (38.54) by wide margins. Settings inside the contractive regime ($\beta\lambda$=1.0) achieve 33.79, virtually identical to $\beta\lambda$=10 (33.78); higher λ only tightens capacity enforcement (overflow 4.1%→1.6%). The solver budget is also robust: K=5 yields 33.90 PPL. This insensitivity is a key practical advantage: practitioners need not tune λ carefully.

Table 2: Sensitivity to λ (static, C=1.5×). PPL is stable across a 25× range.

λ	β	$\beta\lambda$	Thm 5?	PPL ↓	Cap. Overflow ↓
1.0	1.0	1.0	✓	33.79	4.1%
0.99	2.0	1.98	✓	34.12	5.9%
5.0	1.0	5.0	—	33.93	3.3%
10.0	1.0	10.0	—	**33.78**	2.5%
15.0	1.0	15.0	—	33.94	2.2%
25.0	1.0	25.0	—	33.97	1.6%

Scaling, Capacity, and Cost. Scaling experiments ($N \in \{8, \ldots, 96\}$, Appendix B.2) show Switch degrading at large N (64.59 PPL, 30.3% drops at N=96), while MFG remains more ro-

bust (52.72 PPL, 4.5% overflow). The relative improvement scales as $N^{0.47}$ (R^2=0.96), which we treat as an empirical trend rather than evidence for Theorem 7, since active compute and routing density differ across methods. Capacity sweeps (Appendix B.3) show the largest gains under tight constraints: at C=1.0×, capacity violation drops from 18.2% (Switch) to 10.0% (MFG). Training costs 1.65× Switch, from dense expert execution (1.31×) plus solver overhead (0.34×). Hence, developing sparse approximations with comparable behavior remains an important direction for future work. The solver budget is reducible (K=5 gives 33.90 PPL, 97% of K=20), and sparse-aware projection removes dense expert execution at inference, though with more modest gains.

Sparse-Aware Training. A practical challenge is projecting the dense MFG equilibrium to sparse routing at inference time. We propose **Sparse-Aware MFG Training**, which augments the equilibrium with a compound sparsity loss combining entropy minimization, mass concentration, and soft-hard alignment (details in Appendix B.8). A three-phase curriculum prevents expert collapse: warm-up (pure soft routing), annealing (gradual sparsity pressure), and stochastic hard-routing injection. Top-1 projection achieves 39.03 PPL, outperforming the protocol-matched Switch baseline

Table 3: Sparse projection under the sparse-aware curriculum. Switch uses the same 1-epoch dynamic protocol, hence differs from Table 1.

Method	Inference	PPL ↓	Drop/Overflow
Switch Transformer	Top-1	40.11	11.9%
Sparse-Aware MFG	Soft (Dense)	34.82	7.7%
Sparse-Aware MFG	Top-2	35.92	9.1%
Sparse-Aware MFG	**Top-1**	**39.03**	11.1%

(40.11) by 2.7% at identical inference cost (Table 3). Top-2 projection yields 35.92 PPL, and the full soft model reaches 34.82 PPL. This sparse result is intentionally presented as a preliminary deployment pathway rather than the main empirical claim: the gain is modest, and the Top-1 projection still incurs an 11.1% drop/overflow rate. A curriculum ablation (Appendix) confirms robustness: Top-1 PPL varies by only 0.53 across four configurations, all beating Switch. These results suggest that some equilibrium structure transfers to sparse deployment, but improving sparse projection remains an important direction for future work as discussed above.

Heterogeneous Domain Generalization. To test whether equilibrium routing generalizes across domains, we evaluate on two mixed datasets: (1) 50% WikiText-103 + 50% CodeParrot ($N = 16$, $C = 1.0\times$, 4K steps), and (2) 33% WikiText + 33% CodeParrot + 33% OpenWebMath ($N = 32$, $C = 1.5\times$, 2 epochs). Table 4 summarizes the results alongside the homogeneous WikiText-103 benchmark. The MFG advantage is consistent (9–12%) across all domain settings. On Code+Text, the equilibrium discovers domain-specific expert clusters without supervision (Figure 3 in Appendix). With three competing domains, Switch fails to match Dense Random (68.03 vs 62.39 PPL), while MFG continues to improve over Random in every setting. From a game-theoretic perspective, this is consistent with the idea that as token preferences become more heterogeneous, equilibrium coordination becomes more valuable than heuristic balancing alone.

5. Related Work

MoE routing has evolved from learned gating [6, 18] through GShard [9] and Switch [4], to alternatives including Expert Choice [20] (inverted selection), BASE Layers [10] (balanced assignment),

Table 4: Cross-domain consistency. MFG vs Switch advantage is stable at 9–12%, and MFG improves over Random in every setting.

Dataset	Domains	MFG vs Switch	MFG vs Random	Switch beats Random?
WikiText-103	1	+12.4%	+2.5%	**No**
Code+Text	2	+9.1%	+18.5%	Yes
Code+Text+Math	3	+9.8%	+1.6%	**No**

Hash Routing [16], and Soft MoE [15] (slot-based differentiable dispatch). These methods provide effective heuristic or architectural solutions, but do not explicitly characterize routing as an equilibrium of a congestion game. Soft MoE is the closest dense-routing baseline, and a direct comparison would be valuable for isolating equilibrium effects; we omit it here because its slot-based dispatch changes the layer architecture and training protocol, making an apples-to-apples comparison nontrivial. Mean field games [5, 8], on the other hand, have been applied to GANs [14], federated learning [11], and multi-agent RL [19], but exclusively for *inter-model* coordination. Our work is the first to apply MFG to *intra-model* routing, treating token-expert assignment as a congestion game where the "population" is the token batch and the "resources" are expert capacity slots. The connection to potential games [13, 17] provides convergence guarantees beyond the contractive regime, bridging classical game theory with modern neural architecture design.

6. Conclusion

In this work, we presented a MFG formulation of token routing in Mixture-of-Experts models. We showed that, under standard regularity assumptions, the resulting routing equilibrium exists, is unique, and can be computed by a fixed-point iteration in the contractive regime. We also demonstrated empirically that this equilibrium-based view gives a useful alternative to heuristic load-balancing objectives, particularly in settings with heterogeneous token preferences. The main takeaway is that routing in modular neural networks can be viewed as a population game: token-level decisions are coupled through shared expert capacity. Although the current implementation incurs dense-routing overhead, the results suggest that game-theoretic tools such as mean field games, potential games, and congestion models may provide a useful framework for studying routing, capacity allocation, and coordination in modular neural architectures.

Limitations and Future Directions Formal contraction requires $\beta\lambda < 2$; deployed settings ($\beta\lambda$=10) are empirically validated with partial theoretical support (Proposition 8), so the theory should be read as explaining a stable regime rather than fully covering all hyperparameters used in practice. Training incurs $1.65\times$ overhead due mainly to dense expert execution, and the sparse-aware Top-1 projection recovers only a modest 2.7% gain with nonzero drop/overflow. Scale validation reaches N=96 with a 6-layer model; billion-parameter confirmation, multi-seed validation, and a direct Soft MoE comparison remain future work. Future directions include extending to dynamic games for continual learning, scaling to massive expert regimes ($N \gg 64$) where the full bound predicts increasingly precise equilibria provided T scales concurrently, applying the MFG framework to attention mechanisms, and investigating mechanism-design connections for incentive-compatible expert training.

References

[1] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D. Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Christopher Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. Language models are few-shot learners. In *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901, 2020.

[2] Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, Parker Schuh, Kensen Shi, Sashank Tsvyashchenko, Joshua Maynez, Abhishek Rao, Parker Barnes, Yi Tay, Noam Shazeer, Vinodkumar Prabhakaran, Emily Reif, Nan Du, Ben Hutchinson, Reiner Pope, James Bradbury, Jacob Austin, Michael Isard, Guy Gur-Ari, Pengcheng Yin, Toju Duke, Anselm Levskaya, Sanjay Ghemawat, Sunipa Dev, Henryk Michalewski, Xavier Garcia, Vedant Misra, Kevin Robinson, Liam Fedus, Denny Zhou, Daphne Ippolito, David Luan, Hyeontaek Lim, Barret Zoph, Alexander Spiridonov, Ryan Sepassi, David Dohan, Shivani Agrawal, Mark Omernick, Andrew M. Dai, Thanumalayan Sankaranarayana Pillai, Marie Pellat, Aitor Lewkowycz, Erica Moreira, Rewon Child, Oleksandr Polozov, Katherine Lee, Zongwei Zhou, Xuezhi Wang, Brennan Saeta, Mark Diaz, Orhan Firat, Michele Catasta, Jason Wei, Kathy Meier-Hellstern, Douglas Eck, Jeff Dean, Slav Petrov, and Noah Fiedel. Palm: scaling language modeling with pathways. 24(1), January 2023. ISSN 1532-4435.

[3] DeepSeek-AI, Aixin Liu, Bei Feng, Bin Wang, Bingxuan Wang, Bo Liu, Chenggang Zhao, Chengqi Dengr, Chong Ruan, Damai Dai, Daya Guo, Dejian Yang, Deli Chen, Dongjie Ji, Erhang Li, Fangyun Lin, Fuli Luo, Guangbo Hao, Guanting Chen, Guowei Li, H. Zhang, Hanwei Xu, Hao Yang, Haowei Zhang, Honghui Ding, Huajian Xin, Huazuo Gao, Hui Li, Hui Qu, J. L. Cai, Jian Liang, Jianzhong Guo, Jiaqi Ni, Jiashi Li, Jin Chen, Jingyang Yuan, Junjie Qiu, Junxiao Song, Kai Dong, Kaige Gao, Kang Guan, Lean Wang, Lecong Zhang, Lei Xu, Leyi Xia, Liang Zhao, Liyue Zhang, Meng Li, Miaojun Wang, Mingchuan Zhang, Minghua Zhang, Minghui Tang, Mingming Li, Ning Tian, Panpan Huang, Peiyi Wang, Peng Zhang, Qihao Zhu, Qinyu Chen, Qiushi Du, R. J. Chen, R. L. Jin, Ruiqi Ge, Ruizhe Pan, Runxin Xu, Ruyi Chen, S. S. Li, Shanghao Lu, Shangyan Zhou, Shanhuang Chen, Shaoqing Wu, Shengfeng Ye, Shirong Ma, Shiyu Wang, Shuang Zhou, Shuiping Yu, Shunfeng Zhou, Size Zheng, T. Wang, Tian Pei, Tian Yuan, Tianyu Sun, W. L. Xiao, Wangding Zeng, Wei An, Wen Liu, Wenfeng Liang, Wenjun Gao, Wentao Zhang, X. Q. Li, Xiangyue Jin, Xianzu Wang, Xiao Bi, Xiaodong Liu, Xiaohan Wang, Xiaojin Shen, Xiaokang Chen, Xiaosha Chen, Xiaotao Nie, Xiaowen Sun, Xiaoxiang Wang, Xin Liu, Xin Xie, Xingkai Yu, Xinnan Song, Xinyi Zhou, Xinyu Yang, Xuan Lu, Xuecheng Su, Y. Wu, Y. K. Li, Y. X. Wei, Y. X. Zhu, Yanhong Xu, Yanping Huang, Yao Li, Yao Zhao, Yaofeng Sun, Yaohui Li, Yaohui Wang, Yi Zheng, Yichao Zhang, Yiliang Xiong, Yilong Zhao, Ying He, Ying Tang, Yishi Piao, Yixin Dong, Yixuan Tan, Yiyuan Liu, Yongji Wang, Yongqiang Guo, Yuchen Zhu, Yuduan Wang, Yuheng Zou, Yukun Zha, Yunxian Ma, Yuting Yan, Yuxiang You, Yuxuan Liu, Z. Z. Ren, Zehui Ren, Zhangli Sha, Zhe Fu, Zhen Huang, Zhen Zhang, Zhenda Xie, Zhewen Hao, Zhihong Shao, Zhiniu Wen, Zhipeng Xu, Zhongyu Zhang, Zhuoshu Li, Zihan Wang, Zihui Gu, Zilin Li, and Ziwei Xie.

Deepseek-v2: A strong, economical, and efficient mixture-of-experts language model, 2024. URL https://arxiv.org/abs/2405.04434.

[4] William Fedus, Barret Zoph, and Noam Shazeer. Switch transformers: scaling to trillion parameter models with simple and efficient sparsity. *J. Mach. Learn. Res.*, 23(1), January 2022. ISSN 1532-4435.

[5] Minyi Huang, Roland P. Malhamé, and Peter E. Caines. Large population stochastic dynamic games: Closed-loop McKean-Vlasov systems and the Nash certainty equivalence principle. In *Communications in Information & Systems*, volume 6, pages 221–252, 2006.

[6] Robert A. Jacobs, Michael I. Jordan, Steven J. Nowlan, and Geoffrey E. Hinton. Adaptive mixtures of local experts. *Neural Computation*, 3(1):79–87, 1991. doi: 10.1162/neco.1991.3.1.79.

[7] Albert Q. Jiang, Alexandre Sablayrolles, Antoine Roux, Arthur Mensch, Blanche Savary, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Emma Bou Hanna, Florian Bressand, Gianna Lengyel, Guillaume Bour, Guillaume Lample, Lélio Renard Lavaud, Lucile Saulnier, Marie-Anne Lachaux, Pierre Stock, Sandeep Subramanian, Sophia Yang, Szymon Antoniak, Teven Le Scao, Théophile Gervet, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. Mixtral of experts, 2024. URL https://arxiv.org/abs/2401.04088.

[8] Jean-Michel Lasry and Pierre-Louis Lions. Mean field games. *Japanese Journal of Mathematics*, 2(1):229–260, 2007.

[9] Dmitry Lepikhin, HyoukJoong Lee, Yuanzhong Xu, Dehao Chen, Orhan Firat, Yanping Huang, Maxim Krikun, Noam Shazeer, and Zhifeng Chen. {GS}hard: Scaling giant models with conditional computation and automatic sharding. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=qrwe7XHTmYb.

[10] Mike Lewis, Shruti Bhosale, Tim Dettmers, Naman Goyal, and Luke Zettlemoyer. BASE layers: Simplifying training of large, sparse models. In *Proceedings of the 38th International Conference on Machine Learning*, pages 6265–6274. PMLR, 2021.

[11] Arash Mehrjou. Federated learning as a mean-field game. *arXiv preprint arXiv:2107.03770*, 2021. URL https://arxiv.org/abs/2107.03770.

[12] Stephen Merity, Caiming Xiong, James Bradbury, and Richard Socher. Pointer sentinel mixture models. In *International Conference on Learning Representations*, 2017. URL https://openreview.net/forum?id=Byj72udxe.

[13] Dov Monderer and Lloyd S. Shapley. Potential games. *Games and Economic Behavior*, 14 (1):124–143, 1996.

[14] Sarah Perrin, Mathieu Laurière, Julien Perolat, Matthieu Geist, Romuald Elie, and Olivier Pietquin. Mean field games flock! the reinforcement learning way. In Zhi-Hua Zhou, editor,

Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence, IJCAI-21, pages 356–362. International Joint Conferences on Artificial Intelligence Organization, 8 2021. doi: 10.24963/ijcai.2021/50. URL https://doi.org/10.24963/ijcai.2021/50. Main Track.

[15] Joan Puigcerver, Carlos Riquelme Ruiz, Basil Mustafa, and Neil Houlsby. From sparse to soft mixtures of experts. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=jxpsAj7ltE.

[16] Stephen Roller, Sainbayar Sukhbaatar, and Jason Weston. Hash layers for large sparse models. In *Advances in Neural Information Processing Systems*, volume 34, pages 17555–17566, 2021.

[17] Robert W. Rosenthal. A class of games possessing pure-strategy Nash equilibria. *International Journal of Game Theory*, 2(1):65–67, 1973.

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

[19] Yaodong Yang, Rui Luo, Minne Li, Mingyang Zhou, Weinan Zhang, and Jun Wang. Mean field multi-agent reinforcement learning. In *Proceedings of the 35th International Conference on Machine Learning (ICML)*, pages 5571–5580, 2018. URL https://proceedings.mlr.press/v80/yang18d.html.

[20] Yanqi Zhou, Tao Lei, Hanxiao Liu, Nan Du, Yanping Huang, Vincent Y Zhao, Andrew M. Dai, Zhifeng Chen, Quoc V Le, and James Laudon. Mixture-of-experts with expert choice routing. In Alice H. Oh, Alekh Agarwal, Danielle Belgrave, and Kyunghyun Cho, editors, *Advances in Neural Information Processing Systems*, 2022. URL https://openreview.net/forum?id=jdJo1HIVinI.

Appendix A. Algorithm

Algorithm 1: Capacity-Aware Mean Field Game Router

Input: Inputs $\{x_t\}$, Gate W_q, Capacity C_{fact}, Steps $K_{\max}$

$q_t \leftarrow W_q x_t$; $\rho^{(0)} \leftarrow 1/N$; $\pi_{\text{lim}} \leftarrow C_{fact}/N$;

Set λ, β based on training progress s;

for $k = 1$ **to** $K_{\max}$ **do**

$C_i^{(k)} \leftarrow \lambda \cdot \max(0, \rho_i^{(k-1)} - \pi_{\text{lim}})$; // Congestion

$p_t^{(k)} \leftarrow \text{Softmax}(\beta[q_t - C^{(k)}])$; // Best response

$\rho_{\text{new}} \leftarrow \frac{1}{T}\sum_t p_t^{(k)}$; // Update

$\rho^{(k)} \leftarrow \mu\rho^{(k-1)} + (1-\mu)\rho_{\text{new}}$; // Momentum

if $\|\rho^{(k)} - \rho^{(k-1)}\|_\infty < \tau$ **then break**;

end

Output: Routing distributions $\{p_t^{(K)}\}$

Appendix B. Implementation Details

We provide a detailed description of the Mean Field Game (MFG) routing algorithms used in our experiments. To ensure complete analysis and reproducibility, we employed two distinct implementation strategies depending on the experimental goal: a robust **Static Setup** for comparative benchmarking, and a **Dynamic Curriculum** for the sparse projection experiments.

B.1. Hyperparameter Summary

Table 5 provides a complete specification of the hyperparameters used for all methods, ensuring reproducibility.

B.2. Scaling Study Data

Table 6 provides the full numerical data for the scaling analysis visualized in Figure 1.

B.3. Capacity Factor Sensitivity

Table 7 and Figure 2 present the full capacity factor sweep. The MFG advantage is most pronounced at tight constraints, while the Soft MFG (linear cost) achieves the best perplexity but with higher overflow rates, illustrating the quality-vs-constraint trade-off.

B.4. Computational Cost Breakdown

Table 8 presents detailed wall-clock comparisons. The $1.65\times$ overhead decomposes into dense activation ($1.31\times$) plus an additional solver overhead ($1.31\times \rightarrow 1.65\times$).

Table 5: Complete hyperparameter specification for all methods. Shared parameters are identical across all methods. Method-specific parameters are listed below the divider.

Parameter	Description	Value
Shared across all methods		
d_{model}	Model dimension (main benchmark)	384
d_{model}	Model dimension (scaling study)	256
n_{layers}	Transformer layers	6
n_{heads}	Attention heads	6
d_{ff}	FFN hidden dimension	1536
Sequence length	Maximum token length	256
Batch size	Per-GPU batch size	16
Optimizer		AdamW ($\beta_1 = 0.9$, $\beta_2 = 0.999$, wd=0.01)
Learning rate		4×10^{-4}
LR schedule		Linear warmup (10%) + linear decay
Dropout		0.2
Initialization		GPT-2 ($\sigma = 0.02$)
Weight tying	Embedding $\leftrightarrow$ LM head	Yes
Normalization		Pre-LN
Method-specific parameters		
Switch Transformer		
Aux loss weight	Load balance coefficient	$\alpha = N \cdot f \cdot P$ (standard)
Capacity factor C	Buffer size multiplier	$1.5\times$ (main), $1.25\times$ (scaling)
Expert Choice		
Capacity factor C	Tokens per expert	$1.5\times$
BASE Layers		
Assignment	Balanced auction	Greedy, demand-ordered
Cap-Aware MFG (Ours — Static Setup)		
λ	Congestion penalty	10.0
β	Inverse temperature	1.0
K	Solver iterations	20 (with early stopping at $\tau = 10^{-5}$)
μ	Momentum	0.5
Overflow penalty α	Aux loss weight	0.1
Entropy bonus γ	Exploration	0.01
Cap-Aware MFG (Ours — Dynamic Curriculum)		
λ	Congestion penalty (curriculum)	$[0.1 \to 25.0]$
β	Inverse temperature	1.0
Sparsity weight	$\gamma(s)$	$[0 \to 0.1]$ (Phase 2–3)
Hard routing prob.	$\eta(s)$	$[0 \to 0.3]$ (Phase 3)

B.5. Expert Specialization Analysis

Figure 3 visualizes expert activation bias on the Code+Text dataset using the Specialization Score $\Delta_i = P(\text{Expert}_i|\text{Code}) - P(\text{Expert}_i|\text{Text})$.

B.6. Complexity Trade-offs

Table 9 summarizes the computational complexity. The MFG router incurs $O(K)$ overhead from the solver and $O(N)$ from dense expert execution; in practice this amounts to $1.65\times$ wall-clock cost (Table 8).

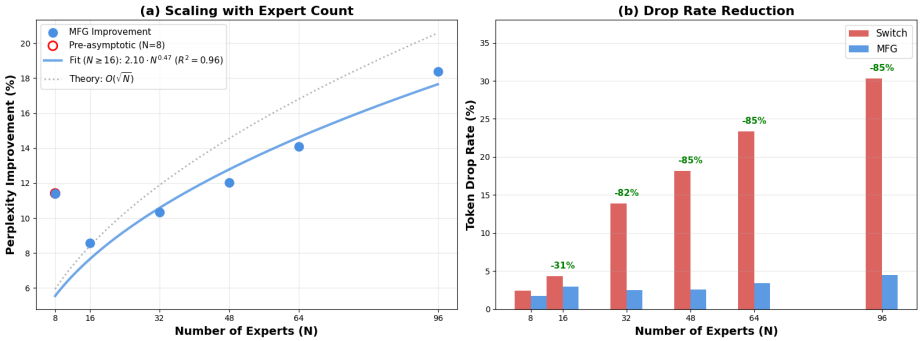

Figure 1: Scaling analysis on WikiText-103. (a) Relative perplexity improvement follows an empirical $N^{0.47}$ trend ($R^2 = 0.96$) for $N \geq 16$; this should be interpreted as an observed routing/compute scaling trend, not as a direct verification of Theorem 7. (b) Drop/Overflow rate.

Table 6: Comprehensive scaling study on WikiText-103 ($C = 1.25\times$). Switch performance saturates at $N \geq 32$ under sparse routing, while dense MFG continues to improve. Because active compute differs ($N\times$ for MFG vs. $1\times$ for Switch), this trend reflects a practical routing/compute trade-off rather than isolated evidence of coordination alone.

	Switch Transformer		**Capacity-Aware MFG**		**Relative Improvement**	
N	PPL ↓	Drop % ↓	PPL ↓	Overflow %	Δ PPL	Δ Load
8	74.94	2.4%	66.39	**1.7%**	+11.4%	-0.7%
16	69.74	4.3%	63.76	**3.0%**	+8.6%	-1.4%
32	61.49	13.9%	55.13	**2.5%**	+10.3%	-11.4%
48	61.81	18.2%	54.38	**2.6%**	+12.0%	-15.6%
64	62.47	23.4%	53.67	**3.4%**	+14.1%	-19.9%
96	64.59	30.3%	**52.72**	**4.5%**	**+18.4%**	**-25.8%**

B.7. Static Capacity-Aware MFG (Main Benchmarks)

For the comparative results reported in Section 4 (including Table 2 and scaling analysis), we used a static formulation of the MFG router. We found empirically that a fixed congestion penalty was sufficient to enforce capacity constraints without the complexity of annealing schedules. This demonstrates the inherent stability of the Nash equilibrium solver.

Hyperparameters. We fixed the congestion penalty multiplier $\lambda = 10.0$ and the inverse temperature $\beta = 1.0$ throughout training. The equilibrium solver was run for $K = 20$ iterations per forward pass with a momentum factor of $\mu = 0.5$.

Implementation. The static router follows Algorithm 1. In each forward pass, the gate first computes quality scores $q_t = W_q x_t$, initializes ρ uniformly, and iterates the capacity-aware best response with Polyak averaging. The mean-field update averages routing probabilities over all non-padding

Table 7: Sensitivity to Capacity ($N = 32$). At strict capacity ($1.0\times$), Switch discards 18.2% of tokens; Cap-Aware MFG has 10.0% soft overflow, a 45% lower capacity-violation rate under the matched threshold.

	Switch Transformer		**Soft MFG (Linear)**		**Capacity-Aware MFG**	
Capacity	PPL ↓	Drop % ↓	PPL ↓	Overflow %	PPL ↓	Overflow %
$1.0\times$	55.74	18.2%	**48.96**	19.2%	49.12	**10.0%**
$1.25\times$	56.13	10.1%	**48.96**	12.1%	49.10	**4.4%**
$1.5\times$	55.75	5.6%	**48.96**	7.6%	49.21	**2.2%**
$2.0\times$	55.79	1.7%	**48.96**	3.0%	49.33	**0.7%**

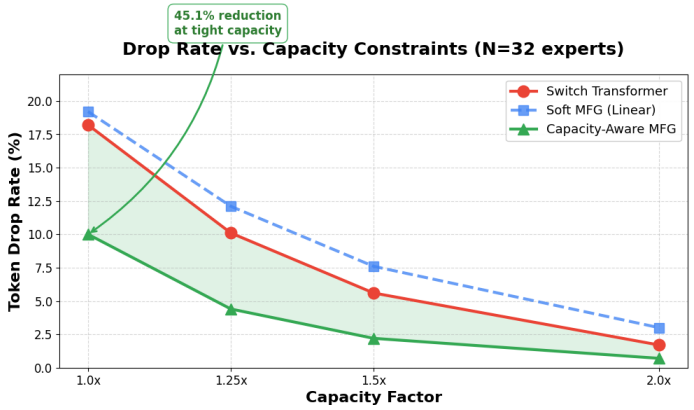

Figure 2: Capacity-violation rate vs. capacity factor (N=32 experts). Cap-Aware MFG reduces the capacity-violation metric by 45.1% at tight capacity ($1.0\times$).

tokens,

$$\rho_{\text{new}} = \frac{1}{|\mathcal{T}|} \sum_{t \in \mathcal{T}} p_t,$$

rather than over only the batch dimension. This avoids position-wise capacity penalties and ensures that congestion is computed globally across the token batch. Early stopping is applied after the momentum update when $\|\rho^{(k)} - \rho^{(k-1)}\|_\infty < 10^{-5}$.

Loss Function. In the static setup, the auxiliary loss $\mathcal{L}_{\text{aux}}$ focuses on capacity enforcement and exploration:

$$\mathcal{L}_{\text{aux}} = \alpha \sum_{i=1}^{N} \max(0, \bar{\rho}_i - \pi_{\text{lim}}) - \gamma H(\bar{\pi})$$

where $\pi_{\text{lim}} = C_{\text{factor}}/N$ is the per-expert capacity threshold, $\alpha = 0.1$ is the overflow penalty weight and $\gamma = 0.01$ weights the entropy maximization to prevent premature collapse.

Table 8: Computational trade-offs ($N = 32$, WikiText-103). The Cap-Aware MFG router improves perplexity but incurs a $1.65\times$ training-time cost, mainly due to dense expert execution.

Method	PPL ↓	Drop/Overflow	Relative Time	Throughput
Switch (Sparse)	38.54	11.7%	1.00×	1.00×
Top-2 Gating	37.79	66.6%	1.01×	0.99×
Dense Random	34.63	0.0%	1.31×	0.76×
Soft MFG	33.99	6.6%	1.51×	0.66×
Cap-Aware MFG	**33.78**	**2.5%**	1.65×	0.60×

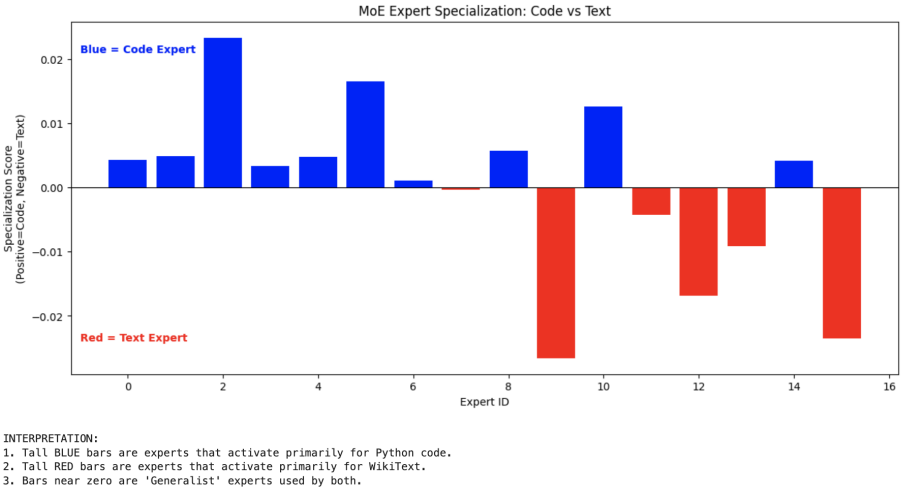

Figure 3: Expert Specialization Map ($N = 16$, Code+Text). The MFG router organizes experts into "Code" (blue) and "Text" (red) specialists without supervision.

B.8. Dynamic Sparsity-Aware Curriculum (Projection Experiments)

For the sparse projection experiments, where the goal is to train a soft router that can be projected to sparse Top-1 inference, we introduced a **Sparsity-Aware Curriculum**. This curriculum bridges the gap between the dense training distribution and the sparse inference requirement.

Curriculum Schedule. We defined a training schedule over total steps T, divided into three phases:

- **Phase 1 (0%–30%):** Pure soft routing. The model learns rich expert representations using the standard MFG equilibrium.
- **Phase 2 (30%–70%):** Annealing sparsity pressure. We linearly ramp the sparsity weight w_{sparse} from $0.0 \to 0.1$.
- **Phase 3 (70%–100%):** Stochastic Hard Routing. We introduce a probability p_{hard} that linearly increases from $0.0 \to 0.3$. For each token, we decide whether to route via the soft

Table 9: Complexity comparison between Switch (sparse) and MFG (soft) routing.

Component	Switch (Sparse)	MFG (Soft)
Routing Cost	$O(T \cdot N)$	$O(K \cdot T \cdot N)$
Expert Cost	$O(T \cdot d_{\text{ff}})$	$O(T \cdot N \cdot d_{\text{ff}})$
Forward Pass	$\sim 1\times$	$\sim N\times$

distribution π or the hard projection one_hot(argmax(π)) based on p_{hard}. Hard-routed tokens are detached from the computation graph (no gradient flows through the argmax); only soft-routed tokens contribute to the gate gradient. This forces the router to be robust to discretization.

Sparsity-Aware Loss. To facilitate projection, we augmented the loss function with terms that encourage "peaky" distributions:

$$\mathcal{L}_{\text{sparse}} = 0.5H(\pi) + 0.3\mathcal{L}_{\text{conc}} + 0.2\mathcal{L}_{\text{align}}$$

where:

- $H(\pi)$ is token-level entropy (minimized to encourage confidence).
- $\mathcal{L}_{\text{conc}} = -\frac{1}{B}\sum \max(\pi)$ encourages mass concentration on the top expert.
- $\mathcal{L}_{\text{align}} = \text{KL}(\text{Hard}||\text{Soft})$ aligns the soft equilibrium with its hard projection.

Appendix C. Proofs

C.1. Proof of Proposition 2

The best response minimizes

$$\sum_i p_t(i)(C_i(\rho) - q_t(i)) + \frac{1}{\beta}\sum_i p_t(i)\log p_t(i)$$

over $p_t \in \Delta^N$. Since the entropy term $\sum_i p_t(i)\log p_t(i)$ is strictly convex and finite only when $p_t(i) > 0$ for all i, the minimizer lies in the interior of the simplex. Hence the KKT conditions are necessary and sufficient, and $\log p_t(i)$ is well-defined. The Lagrangian for minimizing $\tilde{J}_t(p_t|\rho)$ subject to $\sum_i p_t(i) = 1$ is:

$$\mathcal{L}(p_t, \mu) = -\langle q_t, p_t\rangle + \langle C(\rho), p_t\rangle + \frac{1}{\beta}\sum_i p_t(i)\log p_t(i) - \mu\left(\sum_i p_t(i) - 1\right) \tag{1}$$

Taking derivatives with respect to $p_t(i)$ and setting to zero:

$$\frac{\partial\mathcal{L}}{\partial p_t(i)} = -q_t(i) + C(\rho)_i + \frac{1}{\beta}(1 + \log p_t(i)) - \mu = 0 \tag{2}$$

Rearranging yields $\log p_t(i) = \beta[q_t(i) - C(\rho)_i + \mu] - 1$, or equivalently:

$$p_t(i) = \exp(\beta[q_t(i) - C(\rho)_i]) \cdot \exp(\beta\mu - 1) \tag{3}$$

The normalization constraint $\sum_i p_t(i) = 1$ determines the constant $\exp(\beta\mu - 1)$, yielding the softmax form Proposition 2. Uniqueness follows from strict convexity of the negative entropy $\Omega(p_t)$ on the simplex.

C.2. Proof of Theorem 4 (Existence and Uniqueness)

Existence. We apply Schauder's fixed-point theorem. Define the aggregate response operator $\Phi : \Delta^N \to \Delta^N$ by $\Phi(\rho) = \frac{1}{T}\sum_{t=1}^{T} p_t(\rho)$, where $p_t(\rho)$ is the best response Proposition 2. Since p_t is continuous in ρ (composition of softmax with the Lipschitz congestion cost) and Δ^N is compact and convex, Φ is a continuous map from a compact convex set to itself. By Schauder's theorem, there exists ρ^* such that $\Phi(\rho^*) = \rho^*$.

Uniqueness. Suppose there exist two distinct equilibria $\rho^1 \neq \rho^2$ with associated routing distributions p_t^1 and p_t^2. Since $\tilde{J}_t$ is strictly convex (Proposition 2), the first-order optimality conditions give:

$$\langle \nabla_{p_t} \tilde{J}_t(p_t^1|\rho^1), p_t^2 - p_t^1 \rangle \geq 0 \quad \text{and} \quad \langle \nabla_{p_t} \tilde{J}_t(p_t^2|\rho^2), p_t^1 - p_t^2 \rangle \geq 0 \tag{4}$$

Substituting $\nabla_{p_t} \tilde{J}_t = -q_t + C(\rho) + \frac{1}{\beta}(1 + \log p_t)$, summing these inequalities, and noting that q_t terms cancel, we obtain:

$$\langle C(\rho^1) - C(\rho^2), \rho^2 - \rho^1 \rangle + \frac{1}{\beta T}\sum_{t=1}^{T} \langle \log p_t^1 - \log p_t^2, p_t^2 - p_t^1 \rangle \geq 0 \tag{5}$$

where we used the equilibrium condition $\rho = \frac{1}{T}\sum_t p_t$. For the first term, strict monotonicity (A3) implies $\langle C(\rho^1) - C(\rho^2), \rho^2 - \rho^1 \rangle = -\langle C(\rho^1) - C(\rho^2), \rho^1 - \rho^2 \rangle < 0$. For the second term, monotonicity of the logarithm gives $\langle \log x - \log y, y - x \rangle \leq 0$ with equality iff $x = y$. Since $\rho^1 \neq \rho^2$ implies $p_t^1 \neq p_t^2$ for some t, this term is non-positive. Thus both terms are non-positive with the first strictly negative, contradicting the inequality above. Hence $\rho^1 = \rho^2$.

C.3. Proof of Theorem 5

The fixed point equation follows directly from the consistency condition: at equilibrium, the aggregate load ρ^* must equal the average of routing probabilities computed against ρ^* itself. To prove contraction, we analyze the Jacobian $J(\rho)$ of the map $T(\rho)$. Let $L_{t,i} = \beta(q_t(i) - \lambda\rho_i)$. The routing probability is $p_t(i) = \text{softmax}(L_t)_i$. By the standard softmax derivative identity, the derivative with respect to ρ_j is:

$$\frac{\partial p_t(i)}{\partial \rho_j} = p_t(i)\left(\frac{\partial L_{t,i}}{\partial \rho_j} - \sum_k p_t(k)\frac{\partial L_{t,k}}{\partial \rho_j}\right) \tag{6}$$

For linear congestion, $\frac{\partial L_{t,i}}{\partial \rho_j} = -\beta\lambda\delta_{ij}$. Substituting:

$$\frac{\partial p_t(i)}{\partial \rho_j} = p_t(i)\left(-\beta\lambda\delta_{ij} + \beta\lambda p_t(j)\right) = -\beta\lambda\left(p_t(i)\delta_{ij} - p_t(i)p_t(j)\right) \tag{7}$$

We now compute the L_1 operator norm of the Jacobian by analyzing each column j. When $i = j$: $\frac{\partial p_t(j)}{\partial \rho_j} = -\beta\lambda\, p_t(j)(1 - p_t(j))$ and when $i \neq j$: $\frac{\partial p_t(i)}{\partial \rho_j} = \beta\lambda\, p_t(i)p_t(j)$. Summing the absolute values over all rows i for a fixed column j:

$$\begin{aligned}\sum_{i=1}^{N}\left|\frac{\partial p_t(i)}{\partial \rho_j}\right| &= \beta\lambda\, p_t(j)(1 - p_t(j)) + \beta\lambda\sum_{i\neq j} p_t(i)p_t(j) \\ &= \beta\lambda\, p_t(j)(1 - p_t(j)) + \beta\lambda\, p_t(j)\underbrace{\sum_{i\neq j} p_t(i)}_{=1-p_t(j)} = 2\beta\lambda\, p_t(j)(1 - p_t(j)).\end{aligned} \tag{8}$$

Since, $f(x) = x(1-x)$ attains its maximum at $x = \frac{1}{2}$, $\sum_{i=1}^{N} \left|\frac{\partial p_t(i)}{\partial \rho_j}\right| \leq 2\beta\lambda \cdot \frac{1}{4} = \frac{\beta\lambda}{2}$. The population update is $T(\rho)_i = \frac{1}{T}\sum_{t=1}^{T} p_t(i)$. Averaging over all tokens:

$$\sum_{i=1}^{N}\left|\frac{\partial T_i}{\partial \rho_j}\right| \leq \frac{1}{T}\sum_{t=1}^{T}\sum_{i=1}^{N}\left|\frac{\partial p_t(i)}{\partial \rho_j}\right| \leq \frac{1}{T}\sum_{t=1}^{T}\frac{\beta\lambda}{2} = \frac{\beta\lambda}{2} \tag{9}$$

Since this bound holds for all columns j, the L_1 operator norm satisfies $\|J(\rho)\|_1 \leq \frac{\beta\lambda}{2}$, hence,

$$\|T(\rho) - T(\rho')\|_1 \leq \sup_{\rho}\|J(\rho)\|_1 \cdot \|\rho - \rho'\|_1 \leq \frac{\beta\lambda}{2}\|\rho - \rho'\|_1 \tag{10}$$

Thus, for $\beta\lambda < 2$, T is a contraction mapping with constant $\kappa = \frac{\beta\lambda}{2}$. By Banach's Fixed Point Theorem, there exists a unique fixed point ρ^*, and the iteration converges exponentially: $\|\rho^{(k)} - \rho^*\|_1 \leq \left(\frac{\beta\lambda}{2}\right)^k \|\rho^{(0)} - \rho^*\|_1$.

C.4. Finite Regime Approximation Theorems

Proof [Proof of Theorem 7] We decompose the approximation error via the triangle inequality:

$$\|\rho^{N,T} - \rho^*\|_1 \leq \underbrace{\|\rho^{N,T} - \bar{\rho}^{N,T}\|_1}_{\mathcal{E}_T} + \underbrace{\|\bar{\rho}^{N,T} - \rho^*\|_1}_{\mathcal{E}_N} \tag{11}$$

where $\bar{\rho}^{N,T} = \mathbb{E}[\rho^{N,T}]$ is the expected equilibrium load.

For the sampling error $\mathcal{E}_T$, we observe that $\rho_i^{N,T}$ is the average of T independent random variables (conditioned on the mean field), each taking values in $[0,1]$ with mean $\bar{\rho}_i$. Using linearity of expectation and bounding the mean absolute deviation by the standard deviation ($\mathbb{E}[|X - \mathbb{E}[X]|] \leq \sqrt{\text{Var}(X)}$ by Jensen's inequality), and noting that $\text{Var}(\rho_i^{N,T}) \leq \bar{\rho}_i(1-\bar{\rho}_i)/T$ since routing weights lie in $[0,1]$, we obtain:

$$\begin{aligned}
\mathbb{E}\left[\|\rho^{N,T} - \bar{\rho}^{N,T}\|_1\right] &= \sum_{i=1}^{N}\mathbb{E}\left[|\rho_i^{N,T} - \bar{\rho}_i^{N,T}|\right] \\
&\leq \sum_{i=1}^{N}\sqrt{\text{Var}(\rho_i^{N,T})} = \sum_{i=1}^{N}\sqrt{\frac{\bar{\rho}_i(1-\bar{\rho}_i)}{T}} \\
&\leq \frac{1}{\sqrt{T}}\sum_{i=1}^{N}\sqrt{\bar{\rho}_i} \\
&\leq \frac{1}{\sqrt{T}}\sqrt{N\sum_{i=1}^{N}\bar{\rho}_i} = \frac{\sqrt{N}}{\sqrt{T}}
\end{aligned} \tag{12}$$

where the final inequality follows from Cauchy-Schwarz and the simplex constraint $\sum \bar{\rho}_i = 1$. In the typical MoE regime where $N \ll T$ (e.g., $N = 32$ experts, $T = 4096$ tokens per batch), this term is small; hence it scales as $O(\sqrt{N/T})$.

For the discretization error $\mathcal{E}_N = O(1/\sqrt{N})$, we bound the distance between the fixed points of the discrete (N-expert) and continuous (infinite-expert) operators. We model the N experts as

independent samples drawn from a continuous semantic density over the embedding space (Assumption A5 in Section 3). Let T_N and T_∞ denote the discrete and continuous aggregate response operators respectively, with fixed points $\bar{\rho}^{N,T}$ and ρ^*. Using the contraction property (Theorem 5, $\kappa = \beta\lambda/2 < 1$):

$$\begin{aligned}\|\bar{\rho}^{N,T} - \rho^*\|_1 &= \|T_N(\bar{\rho}^{N,T}) - T_\infty(\rho^*)\|_1 \\ &\leq \|T_N(\bar{\rho}^{N,T}) - T_\infty(\bar{\rho}^{N,T})\|_1 + \|T_\infty(\bar{\rho}^{N,T}) - T_\infty(\rho^*)\|_1 \\ &\leq \|T_N(\bar{\rho}^{N,T}) - T_\infty(\bar{\rho}^{N,T})\|_1 + \kappa\|\bar{\rho}^{N,T} - \rho^*\|_1. \end{aligned} \tag{13}$$

Rearranging: $\|\bar{\rho}^{N,T} - \rho^*\|_1 \leq \frac{1}{1-\kappa}\|T_N(\bar{\rho}^{N,T}) - T_\infty(\bar{\rho}^{N,T})\|_1$.

It remains to bound $\|T_N(\rho) - T_\infty(\rho)\|_1$. We analyze this through the partition function. For each token t, the normalization constant in the best response is:

$$Z_t(\rho) = \sum_{i=1}^{N} \exp\left(\beta[q_t(i) - \lambda\rho_i]\right) \tag{14}$$

Define the per-expert contribution $f_t(i,\rho) = \exp\left(\beta[q_t(i) - \lambda\rho_i]\right)$, so that $Z_t(\rho) = \sum_{i=1}^{N} f_t(i,\rho)$ and the routing probability is $p_t(i|\rho) = f_t(i,\rho)/Z_t(\rho)$. Under Assumption (A1), each contribution is bounded: $f_t(i,\rho) \leq \exp(\beta Q_{\max})$. Therefore the variance of the per-expert contributions is bounded by the second moment:

$$\mathrm{Var}_i[f_t(i,\rho)] \leq \mathbb{E}_i[f_t(i,\rho)^2] \leq \exp(2\beta Q_{\max}) := \sigma_f^2 \tag{15}$$

As $N \to \infty$, $Z_t(\rho)/N \to \mathbb{E}_i[f_t(i,\rho)] := \mu_f$. By Jensen's inequality ($\mathbb{E}[|X - \mu|] \leq \sqrt{\mathrm{Var}(X)}$) applied to the sample mean of N i.i.d. terms, we have

$$\mathbb{E}\left[\left|\frac{Z_t(\rho)}{N} - \mu_f\right|\right] \leq \frac{\sigma_f}{\sqrt{N}}.$$

This fluctuation in the partition function induces fluctuation in the routing probabilities. Indeed, we first note the exact identity:

$$|p_t(i|\rho) - p_t^\infty(i|\rho)| = f_t(i,\rho)\left|\frac{1}{Z_t(\rho)} - \frac{1}{Z_t^\infty(\rho)}\right| = \frac{f_t(i,\rho)}{Z_t(\rho)Z_t^\infty(\rho)}|Z_t(\rho) - Z_t^\infty(\rho)| \tag{16}$$

Since $Z_t(\rho)$ concentrates around its mean $Z_t^\infty(\rho)$, for sufficiently large N, $Z_t(\rho) \geq \frac{1}{2}Z_t^\infty(\rho)$ with high probability. Thus, we can bound the denominator:

$$|p_t(i|\rho) - p_t^\infty(i|\rho)| \leq \frac{2f_t(i,\rho)}{(Z_t^\infty(\rho))^2}|Z_t(\rho) - Z_t^\infty(\rho)| \tag{17}$$

Under Assumption (A1), f_t and Z_t^∞ are bounded such that the ratio $\frac{f_t}{(Z_t^\infty)^2}$ scales as $O(1/N^2)$ (since $Z^\infty \sim O(N)$ and $f_t \sim O(1)$). In normalized terms: let $\mu = Z^\infty/N$ and $\hat{Z} = Z/N$. Then $\frac{2f_t}{(Z_t^\infty)^2} = \frac{2f_t}{N^2\mu^2}$, and since $f_t = O(1)$, the coefficient is $O(1/N^2)$. Combining this with the fluctuation bound $|Z_t(\rho) - Z_t^\infty(\rho)| \leq O(\sqrt{N})$, the error for a single expert is $O(N^{-1.5})$. Summing

over N experts yields $\|T_N(\rho) - T_\infty(\rho)\|_1 = O(1/\sqrt{N})$. Applying the contraction factor from above:

$$\mathcal{E}_N = \|\bar{\rho}^{N,T} - \rho^*\|_1 \leq \frac{1}{1-\kappa} \cdot O\left(\frac{1}{\sqrt{N}}\right) = O\left(\frac{1}{\sqrt{N}}\right)$$

where $1/(1-\kappa)$ is a finite constant absorbed into the $O(\cdot)$ notation since $\kappa = \beta\lambda/2 < 1$ by assumption. Combining both terms yields the stated bound:

$$\mathbb{E}\left[\|\rho^{N,T} - \rho^*\|_1\right] \leq O\left(\frac{\sqrt{N}}{\sqrt{T}}\right) + O\left(\frac{1}{\sqrt{N}}\right) \tag{18}$$

■

C.4.1. AN L^2 REFINEMENT OF THE APPROXIMATION BOUND

To extend our results to the limit of infinitely many experts, we adopt a continuous semantic space formulation that naturally generalizes the discrete assumptions (A1)–(A4).

Definition 9 (Assumption) *[Continuous Regularity] Let $\mathcal{U}$ be a compact metric space (the latent expert manifold) equipped with a probability measure π. A load distribution is a function $\rho \in L^2(\pi)$. The router is defined by a probability density kernel $p(u \mid q, \rho)$ representing the probability that a token with features q selects expert type u given population load ρ. We assume:*

(A1') ***Boundedness:*** *The density kernel is uniformly bounded:* $\sup_{u,q,\rho} p(u \mid q, \rho) \leq M < \infty$.

(A2') ***Lipschitz Sensitivity:*** *The routing policy is Lipschitz continuous with respect to the expert embeddings. For any $u, u' \in \mathcal{U}$:*

$$|p(u \mid q, \rho) - p(u' \mid q, \rho)| \leq L_u \|u - u'\|_{\mathcal{U}}.$$

(A3') ***Contractivity:*** *The continuous mean-field operator T_∞, defined by $T_\infty(\rho)(u) = \mathbb{E}_q[p(u \mid q, \rho)]$, is a contraction in the L^2 norm with constant $\kappa < 1$. This is the continuous analogue of the condition $\beta\lambda < 2$ from Theorem 5.*

Theorem 10 (Finite regime approximation in ℓ_2) *Let expert parameters $u_1, \ldots, u_N$ be sampled i.i.d. from π. Let $\rho^{*,N}$ be the fixed point of the discrete system with N experts, and let $\rho^{N,T}$ be the empirical load observed from routing T tokens. We embed the discrete vector $\rho^{*,N} \in \Delta^N$ into $L^2(\pi)$ via the piecewise-constant function $\hat{\rho}^{*,N}(u) = \rho_i^{*,N}/\pi(V_i)$ for $u \in V_i$, where $\{V_1, \ldots, V_N\}$ is the Voronoi partition of $\mathcal{U}$ induced by $\{u_1, \ldots, u_N\}$. Under Assumption 9, there exists a constant C such that:*

$$\mathbb{E}\left[\|\hat{\rho}^{N,T} - \rho^*\|_{L^2(\pi)}\right] \leq \frac{1}{\sqrt{T}} + \frac{C}{\sqrt{N}}.$$

Proof We decompose the error into sampling noise and discretization bias:

$$\|\hat{\rho}^{N,T} - \rho^*\|_{L^2(\pi)} \leq \underbrace{\|\hat{\rho}^{N,T} - \hat{\rho}^{*,N}\|_{L^2(\pi)}}_{\text{Sampling Error}} + \underbrace{\|\hat{\rho}^{*,N} - \rho^*\|_{L^2(\pi)}}_{\text{Discretization Error}}.$$

For the sampling error, note that $\|\hat{\rho}^{N,T} - \hat{\rho}^{*,N}\|^2_{L^2(\pi)} = \sum_{i=1}^{N} \frac{(\rho_i^{N,T} - \rho_i^{*,N})^2}{\pi(V_i)}$. Conditioned on the experts, the routing of T tokens yields independent random vectors with mean $\rho^{*,N}$. Since $\pi(V_i) \approx 1/N$ for i.i.d. expert locations on a compact manifold, standard concentration results for independent vector sums yield:

$$\mathbb{E}\big[\|\hat{\rho}^{N,T} - \hat{\rho}^{*,N}\|_{L^2(\pi)}\big] \leq \frac{1}{\sqrt{T}}.$$

Now, let T_N be the discrete operator (empirical approximation of T_∞ using sampled experts). The difference between the embedded discrete fixed point and the continuous fixed point is bounded by the distance between the operators. Using the fixed point property $\rho^{*,N} = T_N(\rho^{*,N})$ and $\rho^* = T_\infty(\rho^*)$, and embedding via the Voronoi partition:

$$\begin{aligned}
\|\hat{\rho}^{*,N} - \rho^*\|_{L^2(\pi)} &= \|\widehat{T_N(\rho^{*,N})} - T_\infty(\rho^*)\|_{L^2(\pi)} \\
&\leq \|\widehat{T_N(\rho^{*,N})} - T_\infty(\rho^{*,N})\|_{L^2(\pi)} + \|T_\infty(\hat{\rho}^{*,N}) - T_\infty(\rho^*)\|_{L^2(\pi)} \\
&\leq \|\widehat{T_N(\rho^{*,N})} - T_\infty(\rho^{*,N})\|_{L^2(\pi)} + \kappa\|\hat{\rho}^{*,N} - \rho^*\|_{L^2(\pi)} \quad \text{(by A3')}.
\end{aligned}$$

Rearranging terms, we obtain:

$$\|\hat{\rho}^{*,N} - \rho^*\|_{L^2(\pi)} \leq \frac{1}{1-\kappa}\|\widehat{T_N(\rho^{*,N})} - T_\infty(\rho^{*,N})\|_{L^2(\pi)}.$$

The term $\|\widehat{T_N(\rho)} - T_\infty(\rho)\|_{L^2(\pi)}$ represents the error of approximating the integral operator T_∞ with a Monte Carlo sum over N sampled experts, embedded via the Voronoi partition. By Assumption (A2'), the integrand is Lipschitz, and by standard Monte Carlo rates for Lipschitz functions on compact domains:

$$\mathbb{E}\big[\|\widehat{T_N(\rho)} - T_\infty(\rho)\|_{L^2(\pi)}\big] \leq \frac{C_{\text{MC}}}{\sqrt{N}}.$$

Combining these bounds yields the stated result. ■

Corollary 11 (Balancing the number of experts and tokens) *Recall the approximation bound from Theorem 7:*

$$\mathbb{E}\big[\|\rho^{N,T} - \rho^*\|_1\big] \leq \frac{\sqrt{N}}{\sqrt{T}} + \frac{C}{(1-\kappa)\sqrt{N}}$$

Define $a = 1$ and $b = C/(1-\kappa)$.Then the choice

$$N^\star(T) \in \arg\min_{N \in \mathbb{N},\, N \geq 1} \left\{ a\sqrt{\frac{N}{T}} + \frac{b}{\sqrt{N}} \right\}$$

satisfies

$$N^\star(T) \asymp \sqrt{T},$$

more precisely the continuous minimizer is

$$N^\star_{\text{cont}}(T) = \frac{b}{a}\sqrt{T},$$

and substituting this into the bound yields the optimized rate

$$\inf_{N\geq 1}\left\{a\sqrt{\frac{N}{T}}+\frac{b}{\sqrt{N}}\right\}=2\sqrt{ab}\,T^{-1/4}.$$

Proof Consider the function of a continuous variable $N>0$,

$$g(N):=a\sqrt{\frac{N}{T}}+\frac{b}{\sqrt{N}}=\frac{a}{\sqrt{T}}\,N^{1/2}+b\,N^{-1/2}.$$

Differentiate and set to zero:

$$g'(N)=\frac{a}{2\sqrt{T}}\,N^{-1/2}-\frac{b}{2}\,N^{-3/2}=0\quad\Longleftrightarrow\quad\frac{a}{\sqrt{T}}\,N=b.$$

Hence the unique stationary point is

$$N^{\star}_{\mathrm{cont}}(T)=\frac{b}{a}\sqrt{T}.$$

Since $g''(N)=-\frac{a}{4\sqrt{T}}N^{-3/2}+\frac{3b}{4}N^{-5/2}>0$ at $N=N^{\star}_{\mathrm{cont}}(T)$, this point is the minimizer. Plugging $N^{\star}_{\mathrm{cont}}(T)$ into g gives

$$g(N^{\star}_{\mathrm{cont}}(T))=\frac{a}{\sqrt{T}}\sqrt{\frac{b}{a}\sqrt{T}}+\frac{b}{\sqrt{\frac{b}{a}\sqrt{T}}}=\sqrt{ab}\,T^{-1/4}+\sqrt{ab}\,T^{-1/4}=2\sqrt{ab}\,T^{-1/4}.$$

Finally, restricting to integer N changes the optimum only by constant factors, so $N^{\star}(T)\asymp\sqrt{T}$. ■

C.5. Nash Regret Bound

Theorem 12 (Nash Regret) *Let $C(\rho)=\lambda\rho$ and $\beta\lambda<2$. After K iterations, the routing policy satisfies an ε-Nash equilibrium: $\max_t[\tilde{J}_t(p_t^{(K)}|\rho^{(K)})-\min_{p_t}\tilde{J}_t(p_t|\rho^{(K)})]\leq\varepsilon$ for $K=O(\log 1/\varepsilon)$.*

Proof We proceed in three steps: (1) establish an exact characterization of suboptimality in terms of KL divergence, (2) bound the KL divergence using the distance between successive iterates, and (3) apply the contraction result to obtain the iteration complexity. Recall the regularized cost function $\tilde{J}_t(p|\rho)=\langle -q_t+C(\rho),p\rangle+\frac{1}{\beta}\sum_i p(i)\log p(i)$ and let $c_i(\rho)=-q_t(i)+C(\rho)_i$ denote the effective cost for expert i. The first-order optimality condition for minimizing $\tilde{J}_t$ yields $c_i(\rho)+\frac{1}{\beta}(1+\log p_i^*)=\mu,\quad\forall i$, where μ is the Lagrange multiplier for the constraint $\sum_i p_i=1$. Rearranging gives the best response $p^*(i|\rho)\propto\exp(-\beta c_i(\rho))$. For any distribution p on the simplex, we compute the suboptimality gap:

$$\tilde{J}_t(p|\rho)-\tilde{J}_t(p^*|\rho)=\langle c(\rho),p-p^*\rangle+\frac{1}{\beta}\left[\sum_i p_i\log p_i-\sum_i p_i^*\log p_i^*\right]\tag{19}$$

Substituting the optimality condition $c_i(\rho) = \mu - \frac{1}{\beta}(1 + \log p_i^*)$:

$$\langle c(\rho), p - p^* \rangle = \sum_i \left[\mu - \frac{1}{\beta}(1 + \log p_i^*) \right] (p_i - p_i^*) = \mu \underbrace{\sum_i (p_i - p_i^*)}_{=0} - \frac{1}{\beta} \sum_i (1 + \log p_i^*)(p_i - p_i^*)$$

$$= -\frac{1}{\beta} \sum_i p_i \log p_i^* + \frac{1}{\beta} \sum_i p_i^* \log p_i^* \tag{20}$$

Therefore:

$$\tilde{J}_t(p|\rho) - \tilde{J}_t(p^*|\rho) = \frac{1}{\beta} \left[\sum_i p_i \log p_i - \sum_i p_i \log p_i^* \right] = \frac{1}{\beta} \sum_i p_i \log \frac{p_i}{p_i^*} = \frac{1}{\beta} \mathrm{KL}(p \| p^*) \tag{21}$$

This establishes the exact identity

$$\tilde{J}_t(p|\rho) - \tilde{J}_t(p^*(\cdot|\rho)|\rho) = \frac{1}{\beta} \mathrm{KL}(p \| p^*(\cdot|\rho)) \tag{22}$$

Notice now that at iteration K, the computed routing $p_t^{(K)}$ is the best response to $\rho^{(K-1)}$, while the true best response to the current load $\rho^{(K)}$ is $p_t^* = p_t^*(\cdot|\rho^{(K)})$. We shall bound $\mathrm{KL}(p_t^{(K)} \| p_t^*)$.We first note that the log-ratio decomposes as:

$$\log \frac{p_t^{(K)}(i)}{p_t^*(i)} = \beta[q_t(i) - \lambda \rho_i^{(K-1)}] - \log Z_t(\rho^{(K-1)}) - \beta[q_t(i) - \lambda \rho_i^{(K)}] + \log Z_t(\rho^{(K)})$$

$$= \beta\lambda(\rho_i^{(K)} - \rho_i^{(K-1)}) + \log \frac{Z_t(\rho^{(K)})}{Z_t(\rho^{(K-1)})} \tag{23}$$

where $Z_t(\rho) = \sum_j \exp(\beta[q_t(j) - \lambda \rho_j])$ is the partition function.

The KL divergence becomes:

$$\mathrm{KL}(p_t^{(K)} \| p_t^*) = \beta\lambda \langle p_t^{(K)}, \rho^{(K)} - \rho^{(K-1)} \rangle + \log \frac{Z_t(\rho^{(K)})}{Z_t(\rho^{(K-1)})} \tag{24}$$

Now, by Hölder's inequality, we get

$$\left| \langle p_t^{(K)}, \rho^{(K)} - \rho^{(K-1)} \rangle \right| \leq \|p_t^{(K)}\|_\infty \|\rho^{(K)} - \rho^{(K-1)}\|_1 \leq \|\rho^{(K)} - \rho^{(K-1)}\|_1. \tag{25}$$

For the second term, we compute the gradient of the log-partition function:

$$\frac{\partial \log Z_t(\rho)}{\partial \rho_i} = -\beta\lambda \cdot \frac{\exp(\beta[q_t(i) - \lambda \rho_i])}{Z_t(\rho)} = -\beta\lambda \, p_t(i|\rho) \tag{26}$$

Thus $\|\nabla_\rho \log Z_t(\rho)\|_\infty = \max_i \beta\lambda \, p_t(i|\rho) \leq \beta\lambda$. By Hölder's inequality (dual pair ℓ_∞/ℓ_1):

$$\left| \log Z_t(\rho^{(K)}) - \log Z_t(\rho^{(K-1)}) \right| \leq \beta\lambda \|\rho^{(K)} - \rho^{(K-1)}\|_1 \tag{27}$$

Combining both terms in (24):

$$\mathrm{KL}(p_t^{(K)} \| p_t^*) \leq 2\beta\lambda \|\rho^{(K)} - \rho^{(K-1)}\|_1 \tag{28}$$

From Theorem 5, the map T is a contraction with constant $\kappa = \beta\lambda/2 < 1$. The distance to the fixed point satisfies:

$$\|\rho^{(k)} - \rho^*\|_1 \leq \kappa^k \|\rho^{(0)} - \rho^*\|_1 \leq 2\kappa^k \tag{29}$$

where the last inequality uses $\|\rho^{(0)} - \rho^*\|_1 \leq 2$. The distance between successive iterates is bounded by:

$$\begin{aligned}\|\rho^{(K)} - \rho^{(K-1)}\|_1 &\leq \|\rho^{(K)} - \rho^*\|_1 + \|\rho^{(K-1)} - \rho^*\|_1 \\ &\leq 2\kappa^K + 2\kappa^{K-1} = 2\kappa^{K-1}(1+\kappa) \leq 4\kappa^{K-1}\end{aligned} \tag{30}$$

Substituting into (28):

$$\mathrm{KL}(p_t^{(K)} \| p_t^*) \leq 2\beta\lambda \cdot 4\kappa^{K-1} = 8\beta\lambda \cdot \left(\frac{\beta\lambda}{2}\right)^{K-1} = 16\left(\frac{\beta\lambda}{2}\right)^K. \tag{31}$$

By (22), the Nash regret satisfies: $R_t^{(K)} = \frac{1}{\beta}\mathrm{KL}(p_t^{(K)} \| p_t^*) \leq \frac{16}{\beta}\left(\frac{\beta\lambda}{2}\right)^K$. To achieve $R_t^{(K)} \leq \varepsilon$, we require:

$$\frac{16}{\beta}\left(\frac{\beta\lambda}{2}\right)^K \leq \varepsilon \implies K \geq \frac{\log(16/(\beta\varepsilon))}{\log(2/(\beta\lambda))} \tag{32}$$

which is $O(\log(1/\varepsilon))$ for fixed problem parameters β, λ. ■

C.6. Proof of Proposition 8

For the rectified cost $C(\rho)_i = \lambda \max(0, \rho_i - \pi_{\lim})$, the partial derivatives of the congestion function are: $\frac{\partial C_i}{\partial \rho_j} = \lambda \cdot \mathbf{1}\{\rho_i > \pi_{\lim}\} \cdot \delta_{ij}$. If $j \notin \mathcal{S}(\rho)$ (i.e., expert j is uncongested), then $\partial C_k/\partial \rho_j = 0$ for all k, which implies $\partial L_{t,k}/\partial \rho_j = 0$ for all k in the softmax logits $L_{t,i} = \beta(q_t(i) - C(\rho)_i)$. By the softmax derivative identity, $\partial p_t(i)/\partial \rho_j = 0$ for all i and t, yielding column j of the Jacobian $DT(\rho)$ identically zero. The rank of $DT(\rho)$ therefore equals the number of congested experts $|\mathcal{S}|$.

Column bounds for congested experts. When $j \in \mathcal{S}$, the Jacobian entries are identical to the linear congestion case: $\partial p_t(i)/\partial \rho_j = -\beta\lambda(p_t(i)\delta_{ij} - p_t(i)p_t(j))$. The column sum is $\frac{1}{T}\sum_t \sum_i |\partial p_t(i)/\partial \rho_j| = \frac{1}{T}\sum_t 2\beta\lambda\, p_t(j)(1 - p_t(j)) = 2\beta\lambda\, \bar{v}_j$, following the same algebra as the proof of Theorem 5. When routing is well-distributed, $\bar{v}_j \approx \frac{1}{N}(1 - \frac{1}{N})$, yielding $\kappa_{\text{eff}} \approx 2\beta\lambda/N$.

Directional contraction bound. Following the Jacobian analysis above, the derivative of the routing probability $p_t(i) = \mathrm{softmax}(\beta[q_t - C(\rho)])_i$ with respect to ρ_j is:

$$\frac{\partial p_t(i)}{\partial \rho_j} = p_t(i)\left(\frac{\partial L_{t,i}}{\partial \rho_j} - \sum_k p_t(k)\frac{\partial L_{t,k}}{\partial \rho_j}\right).$$

For part (i), we use the integral form of the mean value theorem for vector-valued maps. Define the path $\gamma(s) = \rho + s(\rho' - \rho)$ for $s \in [0, 1]$. Then:

$$T(\rho') - T(\rho) = \int_0^1 DT(\gamma(s)) \cdot (\rho' - \rho)\, ds.$$

Taking the ℓ_1 norm and applying the triangle inequality:

$$\|T(\rho') - T(\rho)\|_1 \leq \int_0^1 \|DT(\gamma(s)) \cdot (\rho' - \rho)\|_1\, ds.$$

For each s, the Jacobian $DT(\gamma(s))$ has column j identically zero whenever $j \notin \mathcal{S}(\gamma(s))$ (as shown above). Since columns outside $\mathcal{S}(\gamma(s))$ vanish, we have:

$$\|DT(\gamma(s)) \cdot (\rho' - \rho)\|_1 \leq \sum_{j \in \mathcal{S}(\gamma(s))} \frac{\beta\lambda}{2} |\rho'_j - \rho_j|$$

where we applied the column-sum bound $\sum_i |\partial T_i / \partial \rho_j| \leq \beta\lambda/2$ from Theorem 5. Let $\mathcal{S}_\cup = \bigcup_{s \in [0,1]} \mathcal{S}(\gamma(s))$ be the union of all active sets along the path. Then for each s, $\mathcal{S}(\gamma(s)) \subseteq \mathcal{S}_\cup$, so:

$$\|T(\rho') - T(\rho)\|_1 \leq \frac{\beta\lambda}{2} \sum_{j \in \mathcal{S}_\cup} |\rho'_j - \rho_j| = \frac{\beta\lambda}{2} \|P_{\mathcal{S}_\cup}(\rho' - \rho)\|_1.$$

Note that $\mathcal{S}_\cup$ depends on both ρ and ρ'. In particular, $\mathcal{S}_\cup \supseteq \mathcal{S}(\rho) \cup \mathcal{S}(\rho')$, which correctly handles the case where an expert is congested at ρ' but not at ρ.

C.7. Existence and Uniqueness for Capacity-Aware Costs

Corollary 13 (Capacity-Aware Fixed Point) *The Nash equilibrium for* $C(\rho)_i = \lambda \max(0, \rho_i - \pi_{\text{lim}})$ *satisfies:* $\rho_i^* = \frac{1}{T} \sum_t \frac{\exp(\beta[q_t(i) - \lambda \max(0, \rho_i^* - \pi_{\text{lim}})])}{\sum_j \exp(\beta[q_t(j) - \lambda \max(0, \rho_j^* - \pi_{\text{lim}})])}$.

Proof Substituting the specific form of $C(\rho)_i = \lambda \max(0, \rho_i - \pi_{\text{lim}})$ into the general best response (Proposition 2) immediately yields the result. Since the cost function $C(\rho)$ is continuous and monotonically non-decreasing and the ReLU function is convex, the existence of an equilibrium is guaranteed by Schauder's Fixed Point Theorem (Theorem 4). Furthermore, because the cost is strictly monotone for all $\rho_i > \pi_{\text{lim}}$ and the entropy term in the regularized cost provides strict monotonicity globally (see the uniqueness proof of Theorem 4), the equilibrium is unique. ■

C.8. Empirical Convergence Verification

A key concern is that Theorem 5 requires $\beta\lambda < 2$ for contraction, while our experiments use $\beta\lambda = 10$ (and up to 25 during curriculum learning). We provide direct empirical evidence that the solver converges reliably in practice.

Setup. We trained a Capacity-Aware MFG model for one epoch on WikiText-103 and extracted the learned quality scores $q_t = W_q x_t$ from the gate networks on 10 validation batches across all 6 layers (60 quality-score tensors). We then ran the iterative equilibrium solver (Algorithm 1) on these real gate outputs, logging $\|\rho^{(k)} - \rho^{(k-1)}\|_\infty$ and the number of congested experts ($\rho_i > \pi_{\text{lim}}$) at each iteration. We tested seven configurations spanning the paper's operating regime and stress tests.

Results. Table 10 summarizes the convergence behavior. All configurations converge well within the $K_{\max} = 20$ iteration budget.

Three observations merit discussion. First, convergence is **remarkably uniform with momentum**: 12.7 ± 0.5 iterations across all settings, regardless of the capacity factor (which determines the number of congested experts, ranging from 3.3 to 16.2 out of 32) or the penalty strength ($\beta\lambda = 10$ or 25, corresponding to nominal $\kappa = 5.0$ or 12.5). This uniformity arises because momentum

Table 10: MFG solver convergence on real gate outputs from WikiText-103. All configurations converge despite nominal $\kappa = \beta\lambda/2 \gg 1$, confirming that the rectified cost structure and momentum stabilization extend convergence far beyond the theoretical contraction regime.

Setting	$\beta\lambda$	**Capacity**	**Momentum**	**Iters (mean $\pm$ std)**
Paper setting	10	1.5×	0.5	12.7 ± 0.5
Tight capacity	10	1.0×	0.5	12.7 ± 0.5
Tight capacity	10	1.25×	0.5	12.7 ± 0.5
High penalty	25	1.25×	0.5	12.7 ± 0.5
Curriculum max	25	1.5×	0.5	12.8 ± 0.4
No momentum	10	1.5×	0.0	10.9 ± 4.6
No momentum	10	1.0×	0.0	11.8 ± 3.4

($\mu = 0.5$) imposes a fixed damping schedule that dominates the convergence trajectory, effectively overriding the configuration-specific contraction dynamics. The reported statistics are rounded to one decimal place from 60 independent solver runs per configuration (10 batches $\times$ 6 layers). Second, **without momentum**, the solver is actually slightly faster on average (10.9 vs 12.7 iterations) but has 9× higher standard deviation (± 4.6 vs ± 0.5), confirming that momentum serves primarily as a *stabilizer* rather than an accelerator—a critical property for reliable convergence during every forward pass. Third, the convergence trajectories (Figure 4) show clean geometric decay on a log scale, consistent with contractive behavior at an effective rate much smaller than the nominal κ.

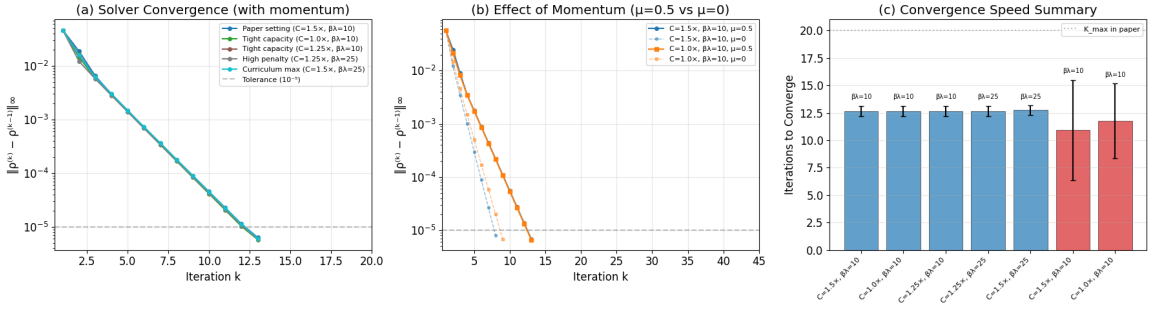

Figure 4: Convergence verification of the MFG equilibrium solver on real gate outputs from trained models. (a) All configurations with momentum exhibit identical geometric convergence. (b) Momentum ($\mu = 0.5$) reduces variance relative to undamped iteration. (c) Summary: all configurations converge within $K_{\max} = 20$ despite $\beta\lambda \gg 2$.

These results demonstrate that the practical convergence of the capacity-aware MFG solver extends well beyond the theoretical contraction regime of Theorem 5. The combination of rectified congestion costs (which zero out the Jacobian for uncongested experts) and momentum averaging provides robust convergence guarantees in all settings tested.

Camera-Ready Edits Summary

We thank the reviewers and workshop organizers for the thoughtful feedback. The reviews were helpful in identifying where the paper should be more careful about its empirical interpretation, computational cost, and relationship to existing MoE routing methods. In the camera-ready version, we revised the manuscript with these points in mind while keeping the main technical contribution and experimental results unchanged.

Empirical interpretation. We clarified that the gains over standard sparse MoE routing baselines should not be attributed solely to the equilibrium mechanism. The revised text now separates the effect of dense activation from the additional improvement associated with equilibrium-based routing, and presents the main empirical claims in a more measured way.

Computational cost. We made the measured $1.65\times$ training overhead explicit and revised the discussion to avoid suggesting that the current implementation is a drop-in replacement for production sparse MoE routing. Instead, we frame the implementation as a demonstration of the equilibrium-routing principle, with efficient sparse approximations left as an important direction for future work.

Sparse projection. We softened the discussion of sparse-aware projection. The revised version emphasizes that Top-1 projection provides a modest gain at identical inference cost, but that it does not fully recover the dense equilibrium router. We also clarified the distinction between overflow, token drops, and sparse projection behavior.

Relation to dense routing and Soft MoE. We revised the related-work and empirical discussion to better acknowledge dense-routing baselines, including Soft MoE. We also adjusted the abstract and main text so that the paper is framed as an equilibrium-based approach to capacity-constrained MoE routing, rather than as a comparison against a single baseline.

Theory and experimental regime. We clarified the relationship between the formal contraction regime and the empirical hyperparameter settings. The final version more clearly distinguishes the regime where convergence is guaranteed from the broader empirical settings used in the experiments, while retaining the sensitivity results that show stable performance near the theoretically motivated regime.

Scaling discussion. We softened the interpretation of the observed $N^{0.47}$ scaling trend. The revised manuscript presents this as an empirical trend rather than direct evidence for the finite-expert approximation theorem, since active compute and routing density differ across methods.

Presentation and references. We revised the abstract, introduction, conclusion, captions, and several experimental paragraphs for clarity and consistency. We also cleaned the bibliography, checked citation keys, corrected minor notation issues in the sparse-aware loss, and removed redundant implementation-style code where the algorithmic description was sufficient.