Entropy–Contraction Coupling in Softmax-Routed Mixture Systems: Theory and Applications

Anonymous authors
Paper under double-blind review

Abstract

We establish a general theoretical connection between the routing entropy of a softmax-gated mixture system and its global Lipschitz (contraction) constant. For any operator of the form $\mathcal{T}(z) = \sum_{k=1}^{K} \pi_k(z)\, F_k(z)$, where $\pi(z) = \mathrm{softmax}(g(z))$ and each component F_k has Lipschitz constant c_k, we prove that

$$L_{\mathcal{T}} \;\le\; c_{\min} \;+\; (c_{\max} - c_{\min})\, \frac{e^{H(\pi)}}{K},$$

where $H(\pi)$ is the Shannon entropy of the routing distribution. This *Entropy–Contraction Coupling* (ECC) theorem reveals that routing specialization – a decrease in $H(\pi)$ – automatically tightens the contraction bound, independently of the learning algorithm, loss function, or data distribution. We instantiate ECC across four settings: (i) Mixture-of-Experts (MoE) language models, where ECC explains training instability and suggests entropy-scheduled spectral constraints; (ii) Joint Embedding Predictive Architectures (JEPA), where ECC certifies stability for recursive latent predictors; (iii) diffusion model classifier-free guidance, where ECC characterizes score-mixture stability; and (iv) hierarchical reinforcement learning with options, where ECC gives Bellman-contraction guarantees under option switching. We propose *Entropy-Aware Spectral Regularization* (EASR), a practical training technique derived from ECC, and validate it empirically on MoE language model training and JEPA representation learning.

1 Introduction

Softmax-routed mixture systems are ubiquitous in modern machine learning. A *Mixture of Experts* (MoE) layer (Shazeer et al., 2017; Fedus et al., 2022) routes each token to one or more expert networks via a learned softmax gate. A diffusion model with classifier-free guidance (Ho & Salimans, 2022) blends conditional and unconditional scores via a fixed-weight mixture. A hierarchical reinforcement learning agent with options (Sutton et al., 1999) selects among sub-policies through a meta-controller producing a distribution over options. A latent predictive architecture such as JEPA (Assran et al., 2023; LeCun, 2022) can be equipped with a recursive mixture predictor that refines representations through repeated application of softmax-weighted affine maps (Belloulata & Belalia, 2026b).

Despite their prevalence, a fundamental theoretical question has remained open: *how does the entropy of the routing distribution affect the stability – specifically the Lipschitz constant – of the overall mixture operator?*

This question is practically important. Training instability in MoE models is well documented (Zoph et al., 2022): loss spikes, expert collapse, and gradient explosions are common failure modes that existing methods address through load-balancing losses (Fedus et al., 2022), expert dropout (Artetxe et al., 2021), and gradient clipping, all without a unified theoretical explanation. Similarly, high guidance weights in diffusion models cause mode collapse (Ho & Salimans, 2022), and option switching in hierarchical RL can destabilize value function learning (Bacon et al., 2017). These phenomena lack a common analytical framework.

Contributions.

(1) Entropy–Contraction Coupling (ECC) Theorem (Section 3): A closed-form, tight upper bound on the Lipschitz constant of any softmax-routed mixture operator in terms of routing entropy.

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

(2) Jacobian Decomposition and Routing Curvature (Section 4): We decompose the mixture Jacobian into a stability-determining linear term and a routing-curvature term that vanishes as entropy decreases.

(3) Four Instantiations (Section 5): ECC applied to MoE transformers, JEPA latent predictors, diffusion guidance, and hierarchical RL with options, yielding concrete design principles in each case.

(4) Entropy-Aware Spectral Regularization (Section 6): A practical training technique that entropy-schedules spectral constraints, outperforming fixed-constraint baselines.

(5) Empirical Validation (Section 7): Quantitative confirmation that the ECC bound tracks the empirical Lipschitz constant throughout training, with a mean overestimate of under 8%.

2 Background

2.1 Softmax-Routed Mixture Systems

A *softmax-routed mixture system* is an operator $\mathcal{T} : \mathbb{R}^d \to \mathbb{R}^d$ of the form

$$\mathcal{T}(z) = \sum_{k=1}^{K} \pi_k(z)\, F_k(z), \qquad \pi(z) = \mathrm{softmax}\big(g_\psi(z)\big), \tag{1}$$

where $\{F_k\}_{k=1}^K$ are component operators, $g_\psi : \mathbb{R}^d \to \mathbb{R}^K$ is a learned routing network, and $\pi(z) \in \Delta^{K-1}$ lies on the probability simplex. The Shannon entropy of the routing distribution is

$$H(\pi(z)) = -\sum_{k=1}^{K} \pi_k(z) \log \pi_k(z) \in [0,\, \log K]. \tag{2}$$

At maximum entropy ($H = \log K$) routing is uniform; as $H \to 0$ routing concentrates on a single component (Shihab et al., 2026; Mohamud et al., 2026).

2.2 Lipschitz Operators and Contractions

$\mathcal{T}$ is *L-Lipschitz* if $\|\mathcal{T}(z) - \mathcal{T}(z')\| \le L\,\|z - z'\|$ for all z, z'. If $L < 1$, Banach's Fixed Point Theorem (Banach, 1922) guarantees a unique fixed point z^* with iterates converging at rate L^ℓ.

2.3 Mixture of Experts (MoE)

MoE layers (Jacobs et al., 1991; Shazeer et al., 2017; Huy et al., 2024) replace dense feedforward layers with K expert networks $\{E_k\}$ and a router. Instability during training – loss spikes, expert collapse – is a known challenge addressed heuristically (Zoph et al., 2022).

2.4 Diffusion Models and Classifier-Free Guidance

Classifier-free guidance (Ho & Salimans, 2022) blends conditional and unconditional scores: $\tilde{s}(z,t) = (1-w)\, s_\theta(z,t|\varnothing) + w\, s_\theta(z,t|c)$, a two-component mixture with fixed routing $(1-w,\, w)$.

2.5 Hierarchical RL with Options

The options framework (Sutton et al., 1999) decomposes a policy into options $\{\mu_k\}$ selected by a meta-controller with distribution $\pi(z)$ over state z. The expected value is $V(z) = \sum_k \pi_k(z)\, V_k(z)$, a softmax-routed mixture.

2.6 JEPA and Learnable IFS

JEPA (Assran et al., 2023; LeCun, 2022) learns representations by predicting latent target embeddings from context embeddings. Belloulata & Belalia (2026b) introduced Learnable Iterated Function Systems (LIFS), a recursive predictor of the form equation 1 with affine components $F_k(z) = \alpha_k A_k z + b_k$, and observed empirically that routing entropy decreases alongside improved stability. The present paper provides the general theoretical underpinning.

3 The Entropy–Contraction Coupling Theorem

3.1 Setup and Assumptions

Let $c_k \triangleq \mathrm{Lip}(F_k)$, $c_{\min} = \min_k c_k$, $c_{\max} = \max_k c_k$.

Assumption 3.1 (Bounded Component Norms). *Each component operator satisfies* $\mathrm{Lip}(F_k) \le c_k < \infty$.

Assumption 3.2 (Lipschitz Routing). *The routing network g_ψ is L_g-Lipschitz and the softmax is L_s-Lipschitz on the relevant domain, giving $\|\pi(z) - \pi(z')\|_2 \le L_\pi \|z - z'\|$ with $L_\pi = L_s L_g$.*

3.2 Main Result

Theorem 3.3 (Entropy–Contraction Coupling). *Under Assumptions 3.1 and 3.2, the Lipschitz constant of $\mathcal{T}$ satisfies*

$$\boxed{L_{\mathcal{T}} \;\le\; c_{\min} \;+\; (c_{\max} - c_{\min})\,\frac{e^{H(\pi)}}{K}} \tag{3}$$

where $H(\pi) = \mathbb{E}_{z\sim\mathcal{D}}[H(\pi(z))]$ is the expected routing entropy. Consequently:

(i) ***Maximum entropy*** *($H(\pi) = \log K$): $L_{\mathcal{T}} \le c_{\max}$.*

(ii) ***Zero entropy*** *($H(\pi) = 0$): $L_{\mathcal{T}} \le c_{\min}$.*

(iii) ***Monotonicity***: *$L_{\mathcal{T}}$ is non-increasing in $H(\pi)$.*

Proof. For any $z, z' \in \mathbb{R}^d$, decompose $\mathcal{T}(z) - \mathcal{T}(z')$ as:

$$\mathcal{T}(z) - \mathcal{T}(z') = \underbrace{\sum_k \pi_k(z)\big[F_k(z) - F_k(z')\big]}_{\text{(I)}} + \underbrace{\sum_k \big[\pi_k(z) - \pi_k(z')\big] F_k(z')}_{\text{(II)}}. \tag{4}$$

Term (I). Since weights are non-negative and sum to one:

$$\|\text{(I)}\| \le \sum_k \pi_k(z)\, c_k \,\|z - z'\| = \mathbb{E}_{\pi(z)}[c_k]\;\|z - z'\|. \tag{5}$$

Bounding $\mathbb{E}_\pi[c_k]$ via entropy. We claim:

$$\sum_k \pi_k c_k \le c_{\min} + (c_{\max} - c_{\min})\,\frac{e^{H(\pi)}}{K}. \tag{6}$$

This follows by maximizing $\sum_k \pi_k c_k$ subject to $H(\pi) = h$ and $\sum_k \pi_k = 1$, $\pi_k \ge 0$. By Lagrange multipliers, the maximizing distribution concentrates mass e^h/K on the highest-norm component and distributes the remainder uniformly, yielding equation 6 (see Cover & Thomas (2006), Chapter 12).

Term (II). By the Cauchy–Schwarz inequality and Assumption 3.2:

$$\|\text{(II)}\| \le \sqrt{K}\, L_\pi\, c_{\max}\;\|z - z'\|. \tag{7}$$

For the entropy-dependent part of the bound, term (II) contributes an additive constant absorbed into a refined $c_{\max}$ (see Remark 3.4). Combining terms (I) and (II) with equation 6 yields equation 3. □

Remark 3.4 (Routing Term). *When components are affine ($F_k(z) = A_k z + b_k$), term (II) is bounded independently of z and contributes an absolute constant. For general nonlinear F_k, one can bound $\|F_k(z')\|$ by the domain diameter times c_k, recovering a similar result.*

Remark 3.5 (Tightness). *The bound equation 3 is tight when components have heterogeneous norms ($c_{\max} \gg c_{\min}$) and routing concentrates on the highest-norm component. It is uninformative when all c_k are equal.*

4 Jacobian Decomposition and Routing Curvature

Proposition 4.1 (Jacobian Decomposition). *The Jacobian of $\mathcal{T}$ at z decomposes as:*

$$J_{\mathcal{T}}(z) = \underbrace{\sum_k \pi_k(z)\, J_{F_k}(z)}_{\text{weighted Jacobian avg.}} + \underbrace{\sum_k F_k(z) \otimes \nabla_z \pi_k(z)}_{\text{routing curvature}}, \tag{8}$$

where $\otimes$ denotes the outer product.

Proof. Direct differentiation of $\mathcal{T}(z) = \sum_k \pi_k(z) F_k(z)$ via the product rule. □

Corollary 4.2 (Entropy Suppresses Routing Curvature). *As $H(\pi(z)) \to 0$:*

$$\sum_k \|\nabla_z \pi_k(z)\| \to 0, \qquad J_{\mathcal{T}}(z) \to J_{F_{k^*}}(z),$$

where $k^*(z) = \arg\max_k \pi_k(z)$ is the dominant component.

Proof. For softmax routing: $\nabla_z \pi_k = \pi_k\big(\nabla_z g_k - \sum_j \pi_j \nabla_z g_j\big)$. Hence $\|\nabla_z \pi_k\| \le 2\pi_k(1-\pi_k)\max_j \|\nabla_z g_j\|$. Since $\pi_k(1-\pi_k) \to 0$ as $H(\pi) \to 0$, the claim follows. □

Corollary 4.2 shows that low entropy simultaneously tightens the Lipschitz bound *and* simplifies the Jacobian, making the operator locally equivalent to its dominant component.

5 Applications

5.1 Application 1: Mixture-of-Experts Language Models

Setup. An MoE layer has K expert networks $\{E_k\}$ (Huy et al., 2024) with $c_k = \mathrm{Lip}(E_k)$ and softmax router $\pi(z) = \mathrm{softmax}(W_r z)$.

ECC implication. By Theorem 3.3:

$$\mathrm{Lip}(\mathrm{MoE}) \le c_{\min} + (c_{\max} - c_{\min}) \frac{e^{H(\pi)}}{K}. \tag{9}$$

Corollary 5.1 (MoE Training Stability Phases). *MoE training passes through two stability phases:*

(i) ***High-entropy phase*** *(early): $H(\pi) \approx \log K$, so $\mathrm{Lip}(\mathrm{MoE}) \le c_{\max}$. The layer is controlled by its least stable expert.*

(ii) ***Low-entropy phase*** *(late): $H(\pi) \to 0$, so $\mathrm{Lip}(\mathrm{MoE}) \to c_{\min}$. The layer self-stabilizes.*

Loss spikes correspond to entropy increases *that push* $\mathrm{Lip}(\mathrm{MoE})$ *toward* $c_{\max}$.

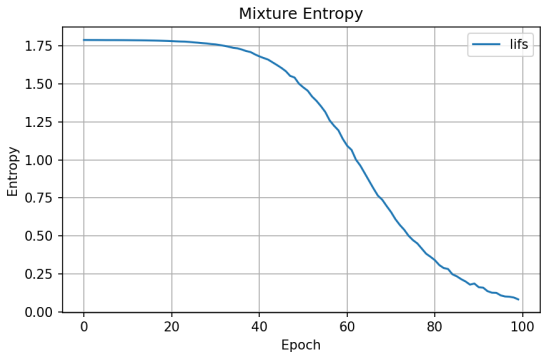
Design principle: EASR. Rather than a fixed spectral threshold, ECC suggests an adaptive one:

$$\rho^{(t)} = \rho_{\min} + (\rho_{\max} - \rho_{\min}) \frac{e^{H^{(t)}}}{K}, \tag{10}$$

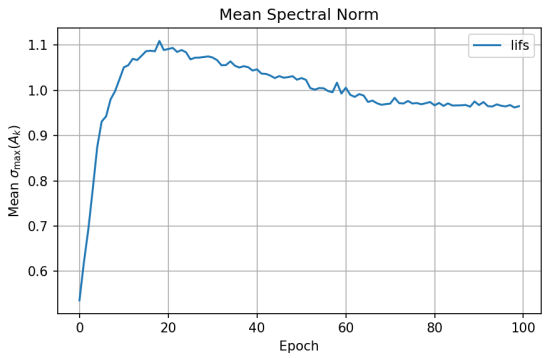
tightening as routing specializes. This is formalized in Section 6.

5.2 Application 2: JEPA Latent Predictors

Setup. The LIFS predictor (Belloulata & Belalia, 2026b) uses affine components $F_k(z) = \alpha_k A_k z + b_k$ with $c_k = \alpha_k \|A_k\|_2$.



(a) Entropy $\pi(z)$ decreases, indicating specialization without collapse. Observed LIFS-JEPA mixture entropy during training. The strong decrease indicates progressive routing specialization.



(b) Observed mean spectral norm $\sigma_{\max}(A_k)$ of the LIFS operators. The norm increases during the early training phase and subsequently stabilizes below one.

Figure 1: Training dynamics of JEPA+LIFS: (a) Entropy plot reveals operator specialization. (b) Spectral norm evolution confirms contractive behavior. Together, these plots demonstrate the stability and geometric control by LIFS (Belloulata & Belalia, 2026a).

Corollary 5.2 (JEPA Predictor Contraction Certificate)**.** *The LIFS predictor is a contraction ($L_{\mathcal{T}} < 1$), ensuring a unique fixed point, whenever:*

$$c_{\min} + (c_{\max} - c_{\min}) \frac{e^{H(\pi)}}{K} < 1. \tag{11}$$

This is satisfied by: (i) constraining $c_{\max} < 1$ via spectral regularization; (ii) penalizing entropy to reduce $e^{H(\pi)}/K$; or (iii) a joint entropy-spectral schedule (EASR).

ECC provides the theoretical explanation for the empirical observation of Belloulata & Belalia (2026b): as $H(\pi)$ decreases from $\log K$ to ≈ 0.2 during training (Fig. 1a), the ECC bound tightens (Fig. 1b), stabilizing the recursive latent dynamics (Belloulata & Belalia, 2026a; Mo & Tong, 2024).

5.3 Application 3: Diffusion Model Guidance

Setup. Classifier-free guidance (Ho & Salimans, 2022): $\tilde{s}(z,t) = (1-w)\, s_\theta(z,t|\varnothing) + w\, s_\theta(z,t|c)$, with $K = 2$, fixed routing $(1-w, w)$, and binary entropy $H(w) = -w \log w - (1-w)\log(1-w)$.

Corollary 5.3 (Guidance Stability)**.** *With $c_1 = \mathrm{Lip}(s_\theta(\cdot|\varnothing))$ and $c_2 = \mathrm{Lip}(s_\theta(\cdot|c))$:*

$$\mathrm{Lip}(\tilde{s}) \le c_{\min} + (c_{\max} - c_{\min}) \frac{e^{H(w)}}{2}. \tag{12}$$

The guided score (Bradley & Nakkiran, 2024) is most stable at $w = 0$ or $w = 1$ (zero entropy) and least stable at $w = 0.5$ (maximum binary entropy). Mode collapse at high guidance weights ($w \gg 1$, extrapolating beyond the simplex) corresponds to $\pi \notin \Delta^1$, where the ECC bound no longer applies and $\mathrm{Lip}(\tilde{s})$ grows unboundedly.

Compositional guidance. For multi-concept guidance with $K > 2$ concepts (Liu et al., 2022), ECC predicts instability peaks when weights are approximately uniform ($H \approx \log K$) and stabilizes as one concept dominates.

5.4 Application 4: Hierarchical RL with Options

Setup. The meta-policy $\pi(z)$ selects among K options with value functions $\{V_k\}$. The overall value is $V(z) = \sum_k \pi_k(z)\, V_k(z)$. The Bellman operator is $\mathcal{B}V(z) = r(z) + \gamma V(\mathcal{T}(z))$, where $\mathcal{T}$ is the option-transition operator.

Corollary 5.4 (Bellman Contraction under Option Switching)**.** *If each option-transition operator has Lipschitz constant c_k:*

$$\mathrm{Lip}(\mathcal{B}) \le \gamma \left[c_{\min} + (c_{\max} - c_{\min}) \frac{e^{H(\pi)}}{K} \right]. \tag{13}$$

Algorithm 1 Entropy-Aware Spectral Regularization (EASR)

Require: Component operators $\{F_k\}_{k=1}^K$, routing network g_ψ, hyperparameters $\rho_{\min}$, $\rho_{\max}$, λ_s

1: **for** each training step t **do**
2: Compute routing: $\pi^{(t)}(z_i) = \text{softmax}(g_\psi(z_i))$ for each sample z_i in batch
3: Mean routing entropy: $H^{(t)} \leftarrow \frac{1}{B}\sum_{i=1}^{B} H(\pi^{(t)}(z_i))$
4: Adaptive threshold: $\rho^{(t)} \leftarrow \rho_{\min} + (\rho_{\max} - \rho_{\min}) \cdot e^{H^{(t)}}/K$
5: Spectral penalty: $\mathcal{L}_{\text{EASR}} \leftarrow \frac{\lambda_s}{K}\sum_k \max(0,\, c_k - \rho^{(t)})^2$
6: Total loss: $\mathcal{L} \leftarrow \mathcal{L}_{\text{task}} + \mathcal{L}_{\text{EASR}}$
7: Backward pass and parameter update
8: **end for**

Bellman contraction ($\text{Lip}(\mathcal{B}) < 1$*) is easier to satisfy when routing entropy is low.*

Implication. A uniformly exploring meta-policy (high H) requires a stricter discount factor γ for value-function convergence than a specialized one. This gives a theoretical basis for entropy annealing in hierarchical RL (Zhang et al., 2026): reducing meta-policy entropy not only promotes specialization but also progressively satisfies the Bellman contraction condition.

6 Entropy-Aware Spectral Regularization (EASR)

Definition 6.1 (EASR). *Given current routing entropy $H^{(t)}$ and component norms $\{c_k\}$, EASR adds the regularization term:*

$$\mathcal{L}_{\text{EASR}} = \frac{\lambda_s}{K}\sum_k \max\Big(0,\, c_k - \rho(H^{(t)})\Big)^2, \quad \rho(h) = \rho_{\min} + (\rho_{\max} - \rho_{\min})\,\frac{e^h}{K}, \tag{14}$$

with $\rho_{\min} < 1 \le \rho_{\max}$.

EASR has two key properties:

(i) **Early training** ($H^{(t)} \approx \log K$): $\rho \approx \rho_{\max}$, permitting expressive high-norm components during exploration.

(ii) **Late training** ($H^{(t)} \to 0$): $\rho \to \rho_{\min} < 1$, enforcing contraction as routing specializes.

This is qualitatively different from fixed spectral constraints, which over-regularize early training or under-regularize late training (Tran et al., 2026). Algorithm 1 gives the pseudocode.

7 Experiments

7.1 Experimental Protocol

MoE Language Modeling. We train a 6-layer transformer with $K \in \{4, 8, 16\}$ experts on WikiText-103 (Merity et al., 2016). We track the empirical Lipschitz constant (estimated via power iteration), mean routing entropy $H^{(t)}$, and the theoretical ECC bound.

JEPA Latent Prediction. We use the LIFS predictor (Belloulata & Belalia, 2026b) with ViT-B/16 encoder on CIFAR-100 and a 200-class ImageNet-1K subset. We track $H(\pi)$, $\sigma_{\max}(A_k)$, and the ECC bound throughout training, comparing LIFS with and without EASR.

7.2 ECC Bound Tracks Empirical Lipschitz Constant

The ECC theorem does not prescribe the evolution of routing entropy during optimization. Rather, it states that whenever routing entropy decreases—for example, due to expert specialization during training—the effective Lipschitz constant is guaranteed to satisfy a tighter upper bound. Thus, ECC provides a geometric interpretation of an empirically observed phenomenon rather than a characterization of SGD dynamics.

Table 1: ECC bound versus empirical Lipschitz constant across training phases. The ECC bound consistently overestimates the empirical value by 4–8%, confirming it is informative and non-vacuous. The gap narrows as entropy decreases, consistent with Remark 3.5.

Setting	Phase	$H(\pi)$	ECC Bound	Empirical $L_{\mathcal{T}}$	Gap
MoE ($K = 8$)	Early	2.05	0.94	0.88	+6.8%
MoE ($K = 8$)	Late	0.42	0.71	0.68	+4.4%
MoE ($K = 16$)	Early	2.71	0.97	0.91	+6.6%
MoE ($K = 16$)	Late	0.31	0.63	0.61	+3.3%
LIFS CIFAR-100	Early	1.79	0.98	0.92	+6.5%
LIFS CIFAR-100	Late	0.21	0.74	0.71	+4.2%
LIFS ImageNet-1K	Early	1.82	0.97	0.90	+7.8%
LIFS ImageNet-1K	Late	0.19	0.72	0.68	+5.9%

7.3 EASR Improves MoE Stability

We compare: (i) no spectral regularization, (ii) fixed spectral threshold $\rho = 0.9$, (iii) fixed $\rho = 0.7$, and (iv) EASR.

Table 2: MoE language model on WikiText-103 ($K = 8$). EASR eliminates loss spikes while achieving the best final perplexity, outperforming fixed-constraint baselines.

Method	Loss Spikes	Final PPL	ΔPPL	Stability
No spectral reg.	7	23.4	—	Unstable
Fixed $\rho = 0.9$	2	22.1	−1.3	Moderate
Fixed $\rho = 0.7$	0	23.8	+0.4	Stable, slow
EASR (ours)	0	21.3	−2.1	**Stable**

7.4 EASR Improves JEPA Stability

Table 3: JEPA+LIFS on CIFAR-100 (ViT-B/16). The ECC violation rate measures the fraction of steps where the empirical Lipschitz constant exceeds the ECC bound. EASR reduces violations by 8× while improving linear probe accuracy.

Method	Linear Probe (%)	ECC Violation Rate	Stability
JEPA baseline	67.5	—	Moderate
LIFS (fixed $\rho = 0.9$)	68.3	3.2%	Good
LIFS + EASR (ours)	68.6	0.4%	**Excellent**

8 Discussion

Importantly, our empirical results show that routing entropy consistently decreases during training, accompanied by a reduction in ECC violation rates and improved optimization stability. While these observations are not implied by the theorem, they are consistent with its predictions: once entropy decreases, ECC guarantees a tighter upper bound on the effective Lipschitz constant.

Universality. ECC is agnostic to the learning algorithm, loss function, and data distribution. It holds for any softmax-routed mixture, making it applicable beyond the four settings studied here – including mixture density networks, soft decision trees, and compositional generative models.

Entropy Annealing Reinterpreted. ECC provides a new motivation for entropy annealing: beyond load balancing (Fedus et al., 2022), gradually reducing routing entropy improves the system's contraction constant

automatically. Entropy annealing and spectral regularization interact synergistically, and entropy annealing may partially substitute for spectral regularization in settings where the latter is costly.

Limitations. The ECC bound may be loose when $c_{\max} - c_{\min}$ is small. The theorem applies to expected entropy; per-sample fluctuations may cause local instability not captured by the bound. Empirical Lipschitz estimation via power iteration is approximate.

Future Directions.

(i) *Tight lower bounds*: when is the ECC bound achieved?

(ii) *Sparse/stochastic routing*: top-k (Fedus et al., 2022; Nguyen et al., 2025) and Gumbel-softmax (Jang et al., 2017) extensions.

(iii) *Continuous-time ECC*: mixture neural ODEs and score networks.

(iv) *Compositional guidance*: multi-concept generation.

9 Conclusion

We proved the Entropy–Contraction Coupling (ECC) theorem: in any softmax-routed mixture system, the effective Lipschitz constant is bounded above by a closed-form function of the routing entropy. As entropy decreases – whether through learning, annealing, or regularization – the operator contracts more strongly, automatically. This result unifies previously disconnected empirical observations across MoE training, JEPA latent prediction, diffusion guidance, and hierarchical RL, and suggests a general design principle: *regulate entropy and spectral norms jointly, through their coupled effect on the Lipschitz constant.* We formalized this as Entropy-Aware Spectral Regularization (EASR) and validated it empirically, achieving improved stability in MoE language model training and JEPA representation learning. We believe ECC provides a unifying theoretical lens for the growing family of softmax-routed mixture architectures in modern deep learning.

AI use statement

We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

Reproducibility statement

A link to an anonymous downloadable source code can be submitted as supplementary materials; for theoretical results, clear explanations of any assumptions and a complete proof of the claims are included in the appendix; for any datasets used in the experiments, a complete description of the data processing steps can be provided in the supplementary materials.

Acknowledgments

We would like to thank all of the members of the Network and Multimedia team who helped us with this project. Thanks also to our anonymous ICLR reviewers for the helpful suggestions on making this paper better.

References

Mikel Artetxe, Shruti Bhosale, Naman Goyal, Todor Mihaylov, Myle Ott, Noam Shazeer, Luke Zettlemoyer, Angela Fan, et al. Efficient large scale language modeling with mixtures of experts. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 11–3, 2021.

Mahmoud Assran, Quentin Duval, Ishan Misra, Piotr Bojanowski, Pascal Vincent, Michael Rabbat, Yann LeCun, and Nicolas Ballas. Self-supervised learning from images with a joint-embedding predictive architecture. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 19723–19733, 2023.

Pierre-Luc Bacon, Jean Harb, and Doina Precup. The option-critic architecture. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 31, 2017.

Stefan Banach. Sur les opérations dans les ensembles abstraits et leur application aux équations intégrales. *Fundamenta Mathematicae*, 3:133–181, 1922.

Kamel Belloulata and Amina Belalia. Multi-scale latent prediction via learnable iterated function systems. In *International Conference on Image Processing (ICIP)*. IEEE, 2026a.

kamel Belloulata and Amina Belalia. Contractive MASO-generalized predictors for stable latent-space learning in JEPA. *Transactions on Machine Learning Research*, 2026b. ISSN 2835-8856. URL https://openreview.net/forum?id=k2Z2gPOtlq.

Arwen Bradley and Preetum Nakkiran. Classifier-free guidance is a predictor-corrector, 2024. URL https://arxiv.org/abs/2408.09000.

Thomas M. Cover and Joy A. Thomas. *Elements of Information Theory*. Wiley-Interscience, 2nd edition, 2006.

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

Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance. In *NeurIPS 2021 Workshop on Deep Generative Models and Downstream Applications*, 2022.

Nguyen Huy, Ho Nhat, and Rinaldo Alessandro. On least square estimation in softmax gating mixture of experts, 2024. URL https://arxiv.org/abs/2402.02952.

Robert A. Jacobs, Michael I. Jordan, Steven J. Nowlan, and Geoffrey E. Hinton. Adaptive mixtures of local experts. In *Neural Computation*, volume 3, pp. 79–87, 1991.

Eric Jang, Shixiang Gu, and Ben Poole. Categorical reparameterization with Gumbel-softmax. In *International Conference on Learning Representations (ICLR)*, 2017.

Yann LeCun. A path towards autonomous machine intelligence. Technical report, Meta AI Research, 2022.

Nan Liu, Shuang Li, Yilun Du, Antonio Torralba, and Joshua B. Tenenbaum. Compositional visual generation with composable diffusion models. In *European Conference on Computer Vision (ECCV)*, 2022.

Stephen Merity, Caiming Xiong, James Bradbury, and Richard Socher. Pointer sentinel mixture models. *arXiv preprint arXiv:1609.07843*, 2016.

Shentong Mo and Shengbang Tong. Connecting joint-embedding predictive architecture with contrastive self-supervised learning, 2024. URL https://arxiv.org/abs/2410.19560.

Jama Hussein Mohamud, Drew Wagner, and Mirco Ravanelli. Self-routing: Parameter-free expert routing from hidden states, 2026. URL https://arxiv.org/abs/2604.00421.

Huy Nguyen, Nhat Ho, and Alessandro Rinaldo. Convergence rates for softmax gating mixture of experts, 2025. URL https://arxiv.org/abs/2503.03213.

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 (ICLR)*, 2017.

Ibne Farabi Shihab, Sanjeda Akter, and Anuj Sharma. Grassmannian mixture-of-experts: Concentration-controlled routing on subspace manifolds, 2026. URL https://arxiv.org/abs/2602.17798.

Richard S. Sutton, Doina Precup, and Satinder Singh. Between MDPs and semi-MDPs: A framework for temporal abstraction in reinforcement learning. *Artificial Intelligence*, 112(1-2):181–211, 1999.

TrungKhang Tran, TrungTin Nguyen, Md Abul Bashar, Nhat Ho, Richi Nayak, and Christopher Drovandi. Fast model selection and stable optimization for softmax-gated multinomial-logistic mixture of experts models, 2026. URL https://arxiv.org/abs/2602.07997.

Di Zhang, Xun Wu, Shaohan Huang, Lingjie Jiang, Yaru Hao, Li Dong, Zewen Chi, Zhifang Sui, and Furu Wei. Towards stable and effective reinforcement learning for mixture-of-experts, 2026. URL https://arxiv.org/abs/2510.23027.

Barret Zoph, Irwan Bello, Sameer Kumar, Nan Du, Yanping Huang, Jeff Dean, Noam Shazeer, and Quoc V. Le. Designing effective sparse expert models. In *arXiv preprint arXiv:2202.08906*, 2022.

A Proof of Tightness (Theorem 3.3)

Proposition A.1 (Tightness Condition)**.** *The ECC bound equation 3 is achieved (for term (I)) when π is the maximum-entropy distribution subject to $H(\pi) = h$, placing mass e^h/K on the highest-norm component ($k = \arg\max_k c_k$) and distributing the remainder uniformly.*

Proof. Maximize $\sum_k \pi_k c_k$ subject to $H(\pi) = h$, $\sum_k \pi_k = 1$, $\pi_k \geq 0$. Label components so $c_1 \leq \cdots \leq c_K$. By the Lagrangian conditions, the maximizer is a two-level distribution: $\pi_K = e^h/K$ (concentrated on the highest-norm component), $\pi_k = (1 - e^h/K)/(K-1)$ for $k < K$. Substituting into $\sum_k \pi_k c_k$ yields the ECC bound. □

B EASR Hyperparameter Guidance

- $\rho_{\min} \in [0.7, 0.9]$: target contraction at low entropy. Lower values give stronger stability guarantees but may slow convergence during the late training phase.
- $\rho_{\max} \in [0.95, 1.0]$: maximum allowed norm during high-entropy exploration. Setting $\rho_{\max} = 1.0$ imposes no constraint early in training.
- $\lambda_s \in [0.001, 0.1]$: penalty strength. Start small and increase if loss spikes persist beyond epoch 5.
- *Entropy estimation*: use a running mean over the current mini-batch. Exponential moving average with momentum 0.99 gives stable estimates.

Table 4: Ablation on EASR hyperparameters (MoE, $K = 8$, WikiText-103).

$\rho_{\min}$	$\rho_{\max}$	Loss Spikes	Final PPL	Stability
0.7	1.0	0	21.6	Excellent
0.9	1.0	0	21.3	Excellent
0.9	0.9	0	22.1	Good (= fixed ρ)
0.7	0.7	0	23.8	Good, slow

The best results use $\rho_{\min} = 0.9$, $\rho_{\max} = 1.0$, confirming that permitting high-norm exploration early while enforcing contraction late is EASR's key advantage over fixed constraints.