ON THE STATISTICAL MECHANICS OF ATTENTION

Anonymous authors
Paper under double-blind review

ABSTRACT

We revisit the statistical mechanics of Transformer attention via modern Hopfield models. We present a free-energy formalism to characterize parameter regimes and critical thresholds for robust attention inference. The resulting free-energy landscape incorporates randomness in query states and attention weights. Attention then emerges as the equation of motion of the corresponding statistical-mechanical system (i.e., modern Hopfield model) through the standard free-energy minimization principle. This allows us to identify first-order phase transitions in attention between glassy and robust-inference phases. We further extend to attentions with stochastic queries and keys. Together, our results provide a rigorous statistical-mechanics framework for characterizing phase structure and inference reliability across a family of Transformer attention mechanisms.

1 INTRODUCTION

We study the statistical mechanics of Transformer attention through modern Hopfield models and characterize parameter regimes and critical threshold for robust attention inference. Our key contribution is a free energy formalism that summarizes the randomness of queries and memories into system's dynamical tendency. Through this construction, attention mechanism hence identifies as a gradient rule on the landscape via free energy minimization principle. Importantly, attention emerges as the equation of motion in standard statistical mechanics. Our result reveals first-order phase transitions of attention between glassy and robust-inference phases. We further extend to models with stochastic properties. Our framework provides a rigorous framework for identifying phase structures across the attention family and invites physics-grounded studies.

Recent works connect the *inference* of Transformer attention (Vaswani et al., 2017) to the *retrieval* of modern Hopfield models (Ramsauer et al., 2020; Krotov & Hopfield, 2020; Hu et al., 2023). Let $Q \in \mathbb{R}^{d_{\text{in}} \times d_k}$, $K \in \mathbb{R}^{d_n \times d_k}$, and $V \in \mathbb{R}^{d_n \times d_v}$ be query, key, and value. Associative memories use Q to retrieve information of stored patterns K. Specifically, softmax attention performs this retrieval via

$$\text{Attn}(Q, K, V) = \text{Softmax}\Big(\frac{QK^\top}{\sqrt{d_k}}\Big)V = p(Q, K)V \in \mathbb{R}^{d_{\text{in}} \times d_v}.$$

Modern Hopfield update rule (Ramsauer et al., 2020) and Transformer attention (Vaswani et al., 2017) arrive at this same form, although one derives it without statistical-mechanical construction and one defines it directly. However, Hopfield models are auto-associative, i.e., $V = K$. In contrast, Transformers are hetero-associative, i.e., V and K are separate parameters. Our analysis hence focuses on the attention weight $p(Q, K)$. During inference, attention then becomes accessible for arbitrary values through $p(Q, K)V$, including auto- and hetero-associative setup.

This connection is decomposable into two aspects. First, Hu et al. (2023) bridge between modern Hopfield models and attention family including softmax, sparsemax, and beyond (Martins & Astudillo, 2016). On the other hand, Hu et al. (2023) reveal that softmax solves an entropy-regularized optimization problem via variation (Peters et al., 2019); see (Santos et al., 2025) for an overview. Together, they view Transformer attentions as physical models and lead to useful methods (Widrich et al., 2020; Hoover et al., 2023; Auer et al., 2023; Hu et al., 2024a; Wu et al., 2024; Dehmamy et al., 2025). However, the prior work only investigate the retrieval rule itself. The parameter regimes for robust inference in these models remains unclear.

In physics, phase structure characterizes regime-dependent behavior of a physical system. Under this attention-Hopfield dual, phase structure reveals retrieval reliability and stability. Amit et al. (1985; 1987); Nishimori (2001) analyze this for classical Hopfield models (Hopfield, 1982) via the mean-field replica method. However, they consider the i.i.d. and infinite-dimension assumptions.

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.

That is, they are less explicit about finite-dimensional metastable states and realistic structured patterns. The extension to modern Hopfield model phase structure becomes crucial but still unsolved.

To address those issues, we utilize the free-energy minimization principle from statistical mechanics (Landau & Lifshitz, 1980; Binder, 1987; Zinn-Justin, 2002) to study phase structures of modern Hopfield models and hence of Transformer attention. Within this framework, different entropy induces different sparsity and hence different attention mechanism (Martins & Astudillo, 2016; Peters et al., 2019; Ramsauer et al., 2020). Through numerical experiments, we explore first-order phase transitions and hence abrupt reorganization of attention weight beyond classical Hopfield models (Hopfield, 1982; Mezard et al., 1986). This reveals an parameter-dependent mechanism behind modern Hopfield models and hence the phase structure of Transformer attention.

Contributions. The main contributions of our work are as follows:

- **Transformer free energy formalism.** Our formulation connects Transformer training variables to the modern Hopfield model through its free energy. This allows us to analyze collective behaviors from macroscopic variables[1] rather than only update rules. Then the consequent attention (Propositions 3.2 and E.7) emerge naturally. This road map unifies entropy, sparsity, attention mechanism, retrieval dynamics, and distributions within a statistical mechanics framework.
- **Metastability analysis.** Given the fixed stored patterns, their correlation spectrum allows us to identify conditions that govern spectral stability (Proposition 2.1). This includes the analyses of query dynamics and metastable states through the shape of energy landscape.
- **Phase transition.** We demonstrate phase transitions through coexistence (Proposition 4.1) and sudden-jump events (Proposition 4.2) theoretically. This provides a concrete explanation of transitions toward retrieval failure along the evolution of local minima on the energy landscape.[2]
- **Experiments.** Our numerical experiments confirm the theoretical predictions of the phase transitions. We explore strict windows in the sparse regime (Figure 8 and Figure 10), and the smooth nature in the softmax regime (Figure 9 and Figure 11). Furthermore, we conclude that sudden-jump events are rare, and most of the observed transitions are due to the support change (Remark G.4).

Organization. Section 2 presents preliminaries. Section 3 presents physical formulation of attention through modern Hopfield models. Section 4 presents our phase transition results. Section 5 presents numerical validations. Section 6 presents concluding remarks and practical implications. We defer related work discussion and statistical-mechanics tutorial in Sections B and C due to space limit. Section E extends to the stochastic version, and Section F analyzes the dynamical stability. Sections G and H show more numerical details and additional experiments.

2 Preliminaries

We first discuss convex analysis and geometric properties in statistical mechanics. This induces entropy and sparsity via the free energy framework. The purpose of this section is to identify the conjugate variables that characterize the retrieval dynamics and attention mechanism.

2.1 Statistical Mechanics and Energy Formalism

We introduce partition function characterizing a system with all possible configurations, and adopt it to discuss free energy through the standard statistical mechanics (Zinn-Justin, 2002; Srednicki, 2007).

Definition 2.1 (Partition Function). Let $\mathcal{H} : \mathcal{X} \to \mathbb{R}$ be a given function called an energy function. We define the partition function as the Gibbs average over all configurations.

$$Z[J] \coloneqq \int_{\mathcal{X}} \exp\{-\beta\mathcal{H}[x] + \sum_{i=1}^{d} J_i x_i\} \, \mathrm{d}\mu(x),$$

[1] Here "microscopic" and "macroscopic" refer to two levels of description. A microscopic state is a sample-level configuration of all individual degrees of freedom, e.g., the vector of outcomes $(s_1, \ldots, s_N)$ from N dice rolls. In our setting, such configurations are sampled from the Gibbs distribution $p_J(x) \propto \exp\{-\beta H[x] + J^\top x\}$. A macroscopic state is an aggregate statistic obtained by averaging over microscopic configurations, e.g., the empirical mean $N^{-1}\sum_{i=1}^{N} s_i$ or the histogram of dice outcomes. Thus, the macroscopic variable is $\varphi = \mathbb{E}_{x\sim p_J}[x]$, and $\Gamma[\varphi]$ is an energy landscape over averaged variables. Its stationary points describe stable averaged responses rather than individual microscopic samples.

[2] We distinguish that the energy landscape here refers to free energy, rather than energy function (Hamiltonian). We adopt this terminology throughout the whole paper.

Table 1: Roadmap from the technical terms to their role in attention.

Concept	Meaning for Attention	Dictionary
Fenchel Conjugate	It transforms a microscopic quantity into a macroscopic one, and constructs the free-energy objective.	Definition D.6, Proposition 2.1, and Theorem 2.1
Tsallis Entropy Ψ_α	Attention Family: softmax ($\alpha = 1$), entmax ($1 < \alpha < 2$), and sparsemax ($\alpha = 2$).	Section D.3, Theorem 2.1, and Proposition 3.3
Query-Key Similarity (Overlap $z = \frac{1}{\sqrt{d}}\xi^\top x$)	It measures the alignment between queries and memories. A larger overpal indicates a stronger retrieval preference for that memory.	Definition 3.2
Free-energy Γ_α	It defines free-energy landscape and enables assigning energy value for each local minimum. It encodes equation of motion as attention rule.	Definition 2.1, Propositions 3.2 and 3.3
Coexistence (Energy Crossing)	It occurs when two distinct stable retrieval states have the same free energy Γ_α simultaneously.	Definition 4.2 and Proposition 4.1
Spinodal	It occurs when the energy barrier of a local mimimum disappers. This drives a retrieval state to the other.	Definition 4.3 and Proposition 4.2
Exponential Memory Load $\kappa = \log(M)/d$	It quantifies the growth rate of the number of memories. This affects retrieval interference and reliability.	Figure 3 and Sections G.5 and G.6

where $\beta := 1/T$ is the inverse of temperature, $J \in \mathbb{R}^d$ is the generating source, and $\mathrm{d}\mu(x)$ denotes the Lebesgue measure in $\mathcal{X}$. The partition function is positive. For simplicity, we consider $\mathcal{X} = \mathbb{R}^d$.

Remark 2.1. The generating source term represents the external magnetic field in a spin system. Furthermore, if the model is discrete, then we replace $\int_{\mathcal{X}} \mathrm{d}\mu(x)$ by $\sum_{x\in\mathcal{X}}$ with $\mathcal{X} = \{1, ..., K\}$.

Definition 2.2 (Stationary Equations and Stationary Points)**.** The stationary equations are

$$\nabla_x\{\beta\mathcal{H}[x] - \sum_{i=1}^{d} J_i x_i\}|_{x=x_J} = 0\ ,\ \nabla_x^2\{\beta\mathcal{H}[x] - \sum_{i=1}^{d} J_i x_i\}|_{x=x_J} \succ 0\ , \tag{2.1}$$

where $\succ$ means positive definite. We say that the solution x_J is a stationary or (local) stable point.

We obtain stable points in Definition 2.2 through the energy function $\mathcal{H}[x]$. However, it contains microscopic noise components. Hence, x_J is not a well-behaved stable point. To extract the aggregated stability from the original energy function $\mathcal{H}[x]$, we should introduce the free energy (effective action) $\Gamma[\varphi]$. It later helps characterize the macroscopic averaged energy of the system.

Definition 2.3 (Log-Partition Function and Free Energy)**.** We have two different free energies that describe the partition function. To distinguish them, we define them carefully.

- **Log-Partition Function**: It is defined as

$$W[J] := \log Z[J] = \log\left(\int_{\mathcal{X}} \exp\{-\beta\mathcal{H}[x] + \sum_{i=1}^{d} J_i x_i\}\, \mathrm{d}\mu(x)\right). \tag{2.2}$$

- **Free Energy (Effective Action)**: It is the Fenchel conjugate of the log-partition function.

$$\Gamma[\varphi] := \sup_{J\in\mathbb{R}^d}\{\sum_{i=1}^{d} J_i\varphi_i - W[J]\}, \qquad \text{with } \varphi \in \mathbb{R}^d. \tag{2.3}$$

Remark 2.2. From Section D.1, $W[J]$ and $\Gamma[\varphi]$ are both convex, $W[J]$ is smooth, $\nabla_J W[J]$ is single-valued, and $\Gamma[\varphi]$ is closed. However, we do not assume strict convexity. Consequently, the

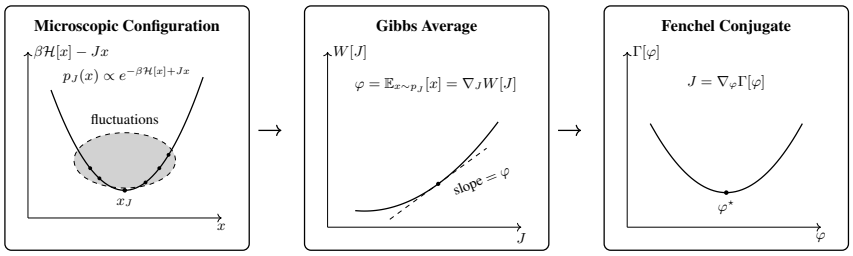


Figure 1: **Illustration of the procedure from the microscopic to macroscopic energy landscapes. Left.** "(Microscopic) Fluctuations" indicates x is random, has nonzero variance, and follows a Gibbs distribution $p_J(x)$. From the configuration built from microscopic x, the local stable point x_J exhibits randomness. **Middle.** $W[J] = \log\big(\int_{\mathcal{X}} \exp\{-\beta\mathcal{H}[x] + Jx\}\big)$ encodes macroscopic responses, i.e., $\varphi = \mathbb{E}_{x\sim p_J}[x] = \nabla_J W[J]$. Hence, the slope of $W[J]$ gives a deterministic average. **Right.** $\Gamma[\varphi] = \sup_J\{J\varphi - W[J]\}$ is the Fenchel conjugate of $W[J]$. Its response is $J = \nabla_\varphi \Gamma[\varphi]$, so local stable macroscopic points occurs at $J = 0$. $\Gamma[\varphi]$ describes the energy landscape built from macroscopic φ after integrating out microscopic fluctuations.

gradient maps $\nabla_J W[J]$ and $\nabla_\varphi \Gamma[\varphi]$ are not guaranteed to be bijective according to Fact 1. In this way, distinct microscopic configurations are able to represent the same macroscopic state.

The optimization problem becomes minimizing the free energy $\Gamma[\varphi]$ with respect to φ. Next, we describe the physical interpretation/derivation of the Fenchel-Young duality.

Proposition 2.1 (Fenchel-Young Duality Relation (Srednicki, 2007; Blondel et al., 2020)). Let p_J be the Gibbs measure with external source J. From the standard statistical field theory, the induced macroscopic field $\varphi = \mathbb{E}_{p_J}[x]$ reveals the following source-field relation:

$$\varphi = \mathbb{E}_{x\sim p_J}[x] \in \partial W[J] \;\Leftrightarrow\; J \in \partial\Gamma[\varphi]\,,$$

where ∂ is the subdifferential from Definition D.5. If W and Γ are differentiable, then ∂ reduce to gradients ∇. We call them equations of motion (EOMs) in the physics literature. It also refers to the Fenchel-Young duality relation. For φ and J, we call them the classical solutions or saddle points.

Proof. Please see Section D.2 for a detailed proof. □

In Proposition 2.1, we derive that the Fenchel conjugate variable φ is macroscopic $\mathbb{E}_{x\sim p_J}[x]$ exactly. φ eliminates fluctuations and induces deterministic landscape $\Gamma[\varphi]$. Please see Figure 1 for illustration.

2.2 Connection between Free Energy and Tsallis Entropy

In this section, we connect our free energy to Tsallis entropy. The value of α determines the sparsity structure of the retrieval dynamics. As we shall see in Section 3, this choice covers the general attention mechanisms, ranging from dense softmax ($\alpha = 1$) to the sparse case ($\infty > \alpha > 1$).

For preliminaries on Tsallis entropy, α-deformed exponential and logarithm family, please see Section D.3 for completeness. They generalize free energy to the α-deformed version.

Theorem 2.1 (Tsallis Entropy and Deformed Free Energy). If we replace the exponential functions and logarithm functions by α-deformed functions, then we have the deformed log-partition function

$$-W_\alpha[J] = \beta\mathbb{E}^{(\alpha)}_{x\sim p_J}[\mathcal{H}[x]] - \frac{\Psi_\alpha[p_J]}{Z_\alpha^{\alpha-1}[J]\int_{\mathcal{X}} p_J^\alpha(x)\,\mathrm{d}\mu(x)} - \sum_{i=1}^{d} J_i \mathbb{E}^{(\alpha)}_{x\sim p_J}[x_i],$$

where $\mathbb{E}^{(\alpha)}_{x\sim p_J}[\cdot] \coloneqq \int_{\mathcal{X}} p_J^\alpha(x)[\cdot]\mathrm{d}\mu(x)/\int_{\mathcal{X}} p_J^\alpha(x)\mathrm{d}\mu(x)$ is the escort expectation value, and

$$\Gamma_\alpha = \beta\mathbb{E}^{(\alpha)}_{x\sim p_{J^\star}}[\mathcal{H}[x]] - \frac{\Psi_\alpha[p_{J^\star}]}{Z_\alpha^{\alpha-1}[J^\star]\int_{\mathcal{X}} p_{J^\star}^\alpha(x)\,\mathrm{d}\mu(x)}, \qquad \text{with } \varphi = \mathbb{E}^{(\alpha)}_{x\sim p_{J^\star}}[x], \tag{2.4}$$

where $J^\star = \arg\max_J\{\sum_{i=1}^d J_i\varphi_i - W_\alpha[J]\}$. We assume $\mathcal{X}$ is finite-volume to avoid divergence.

Proof. Please see Section D.4 for a detailed proof. □

Remark 2.3. If $\alpha = 1$, an escort expectation value becomes an ordinary expectation value. Hence, if we further consider the zero-temperature limit $\beta \to \infty$, then minimizing the free energy Γ_α is equivalent to minimizing the energy function $\mathcal{H}$. In the high-temperature limit, minimizing Γ_α is equivalent to maximizing the Tsallis entropy Ψ_α. The latter case indicates a disorder regime.

Remark 2.4. Theorem 2.1 relates to the evidence lower bound (ELBO) in variational autoencoders (VAE), (Kingma & Welling, 2013). Specifically, with a posterior $q_\phi(z|x)$, we have the inequality

$$\log p_\theta(x) \geq \mathbb{E}_{z\sim q_\phi(z|x)}[\log p_\theta(x,z)] - \mathbb{E}_{z\sim q_\phi(z|x)}[\log q_\phi(z|x)].$$

We recognize that the second term on the right hand side is Shannon entropy, (D.3). Then,

$$\text{ELBO} \coloneqq \mathbb{E}_{z\sim q_\phi(z|x)}[\log p_\theta(x,z)] + S[q_\phi(\cdot|x)].$$

It is an energy-entropy decomposition, where the expected log joint probability corresponds to a negative energy. Hence, maximizing ELBO is equivalent to minimizing a free energy (2.4).

3 Physics of Attention: Modern Hopfield Model (MHM)

In this section, we define an energy function of MHMs in (Hu et al., 2023; Wu et al., 2023). Then we express the free energy in terms of Tsallis entropy and derive the emergent attention mechanism.

3.1 Modern Hopfield Model (MHM)

Inspired by the optimization objectives in (Ramsauer et al., 2020; Krotov & Hopfield, 2020; Hu et al., 2023; Wu et al., 2023), we introduce the explicit joint energy function of MHM as follows.

Definition 3.1 (Modern Hopfield Model (MHM))**.** The joint energy function of the MHM is

$$\widetilde{\mathcal{H}}_{\text{MHM}}[x,p,\xi] \coloneqq \frac{1}{2}\sum_{i=1}^{d} x_i^2 - \sum_{\mu=1}^{M} z_\mu p_\mu, \tag{3.1}$$

where $p \in \Delta^M$ lives on a probability simplex, i.e., p is non-negative and satisfies $\sum_{\mu=1}^{M} p_\mu = 1$, and $z_\mu \coloneqq \sum_{i=1}^{d} \xi_i^\mu x_i/\sqrt{d}$. Moreover, we denote $\mathcal{H}_{\text{MHM}}[x,\xi] \coloneqq \inf_{p\in\Delta^M} \widetilde{\mathcal{H}}_{\text{MHM}}[x,p,\xi]$.

In our setting, query states, the probability distribution, and the stored patterns are all fixed, i.e., the distributions of x and p collapse to their escort average. Specifically, a Taylor expansion of $f(x)$ gives $\mathbb{E}^{(\alpha)}_{(x,p)\sim\rho_J}[f(x)] = f(\varphi) + O(\text{Var}^{(\alpha)}(x))$, where $\rho_J(x,p)$ is the joint probability distribution, $\varphi = \mathbb{E}^{(\alpha)}_{(x,p)\sim\rho_J}[x]$ is the expectation value, and $\text{Var}^{(\alpha)}$ is the variance. Then the fixed-parameter setting corresponds to the zero variance. Hence, we are able to evaluate all quantities $f(x)$ at φ under this fixed-parameter setting exactly. Thus, we identify φ with x, and similarly for p and z. [3]

Definition 3.2 (Statistics-MHM Mapping)**.** The mapping is:

- **Query State Vector.** $x \in \mathbb{R}^d$: It represents continuous query states. Due to our non-fluctuating setting, it is also the macroscopic dynamical variable parameterizing $\Gamma_\alpha[x] = \Gamma_\alpha[\varphi]$.
- **External Source.** $J \in \mathbb{R}^d$: It represents an external bias of the query states. $J = 0$ identifies saddle points in the energy landscape $\Gamma_\alpha[x]$ due to the Fenchel-Young relation (Proposition 2.1).
- **Probability Function.** $p \in \Delta^M$: It represents the retrieval distribution over memories.
- **Tsallis Entropy.** Ψ_α: It represents a regularizer controlling how sparse the memory weights p are.
- **Conjugate Field.** $z \in \mathbb{R}^M$: It represents query-key similarity between the current query state and each memory pattern via the relation: $z_\mu \coloneqq \sum_{i=1}^{d} \xi_i^\mu x_i/\sqrt{d}$. Higher z_μ suggests higher similarity.

The quantities in Definition 3.2 are related via previous statistical mechanics approach. Specifically,

Proposition 3.1 (Query-Probability Relation)**.** A query state relates to a probability function via

$$x_i = \frac{1}{\sqrt{d}}\sum_{\mu=1}^{M} \xi_i^\mu p_\mu + \frac{1}{\beta} J_i. \tag{3.2}$$

Proof. Please see Section D.5 for a detailed proof. □

In Section 2.2, we relate free energy to Tsallis entropy via (2.4). Inspired by that relation, we are capable of expressing the free energy in terms of the variables of MHMs as follows.

[3] Here we restrict attention to the deterministic regime. This allows us to isolate the physics most directly relevant to the attention mechanism. In Section E, we consider the more general setting in which both x and p are random variables with nonzero variance. This leads to a richer and more complete statistical-mechanics structure, and allows us to analyze fluctuations and response properties.

Theorem 3.1 (Free Energy of MHM). The joint free energy of the MHM is

$$\widetilde{\Gamma}_\alpha[x,p,\xi] = \frac{\beta}{2}\sum_{i=1}^{d} x_i^2 - \beta \sum_{\mu=1}^{M} z_\mu p_\mu. \tag{3.3}$$

Hence, after taking the infimum with respect to p, we obtain the free energy

$$\Gamma_\alpha[x,\xi] = \frac{\beta}{2}\sum_{i=1}^{d} x_i^2 - \beta \max_{\mu\in[1,M]} z_\mu. \tag{3.4}$$

Proof. Please see Section D.6 for a detailed proof. □

This free energy selects the maximal component of the overlap $\max_\mu z_\mu$. The hard selection mechanism arises from zero entropy due to the collapse of p. This captures a supremum over simplex vertices rather than the full simplex Δ^M. The relaxation from a hard selection to a distribution over multiple memories becomes a natural extension. Motivated by the sparsity of the entropy regularizer in (Hu et al., 2023), we consider the relaxation through Fenchel conjugation as follows.

We introduce the entropy regularizer in (3.4). This yields a generalized (relaxed) model:

$$\Gamma_\alpha^{(\mathrm{mod})}[x,\xi] = \frac{\beta}{2}\sum_{i=1}^{d} x_i^2 - \Omega_\alpha^\star[\beta z], \qquad \Omega_\alpha := -\Psi_\alpha, \tag{3.5}$$

where $\Omega_\alpha^\star[\beta z] = \sup_{p\in\Delta^M}\{\sum_{\mu=1}^M \beta z_\mu p_\mu - \Omega_\alpha[p]\}$ is the Fenchel conjugate of the negative Tsallis entropy (D.5). It recovers the energy form in (Hu et al., 2023; Wu et al., 2023) exactly.

Remark 3.1. This generalization allows us to relax p to have a distribution over multiple memories on the full simplex Δ^M. This leads to several theoretical and practical advantages:

1. **Equilibrium Selection vs. Averaging.** The MHM performs a hard maximization, selecting a dingle dominant component of $\Gamma_\alpha^{(\mathrm{mod})}[x]$. In contrast, expectation-based rules compute weighted averages over configurations. The generalization (3.5) interpolates between these regimes, i.e., it reduces to the original MHM in $\beta \to \infty$ limit. Please see Section E for details.
2. **Robustness to Noise.** As we will see in Proposition 3.3, the generalized model emphasizes the dominant components of z in for $\alpha > 1$. Hence, it suppresses contributions from noisy patterns. In contrast, expectation-based rules average all components, making them more sensitive to noise.
3. **Sparsity.** The choice of entropy function Ψ_α controls the support of sparsity. In particular, for $\alpha > 1$, the maximizer $p^\star$ is sparse, so some of the components are zero. This is also a consequence of robustness to noise, which does not appear in expectation-based formulations.
4. **Computational Accessibility.** Fenchel conjugation leads to an explicit variational object and is more accessible with gradient-based methods. In contrast, expectation-based formulations require sampling or approximation of Gibbs averages, which is more demanding.

3.2 Retrieval Dynamics and Attention Mechanism

In Section 2, we introduce EOMs according to Proposition 2.1. These equations describe the trajectory that the flow should follow. In this section, we investigate the underlying physical meaning of the trajectories in associative memory systems. In particular, they describe emergent thermal attention.

Proposition 3.2 (EOMs of Free Energies in MHMs). Let Z_α be the partition function with α-deformed Gibbs average. Then the EOM in the J-space corresponds to the Hopfield retrieval

$$\nabla_J W_\alpha[J,\xi] = x, \qquad \text{with } W_\alpha[J,\xi] := \log_\alpha Z_\alpha[J,\xi]. \tag{3.6}$$

The EOM of the joint log-partition function in the z-space corresponds to

$$\nabla_z \widetilde{W}[z,p,J] = \beta p, \qquad \text{with } \widetilde{W}[z,p,J] := -\beta\widetilde{\mathcal{H}}_{\mathrm{MHM}}[x,p,\xi] + \sum_{i=1}^{d} J_i x_i. \tag{3.7}$$

The EOMs of the joint free energy and the generalized one are

$$\nabla_z \widetilde{\Gamma}_\alpha[x,p,\xi] = \nabla_z \widetilde{\Gamma}_\alpha^{(\mathrm{mod})}[x,p,\xi] = -\beta p\,. \tag{3.8}$$

Proof. Under fixed x and p, the log-partition function is $W_\alpha[J,\xi] = -\beta/2\sum_{i=1}^d x_i^2 + \beta\sum_{\mu=1}^M z_\mu p_\mu + \sum_{i=1}^d J_i x_i$ following from the proof of Theorem 3.1. After taking the deriva-

tive of $W_\alpha[J, \xi]$ with respect to J, we obtain the first equation. For the second and the last ones, we derive them by taking the derivative of (3.1), (3.4), and (3.5) with respect to z_μ directly. □

We observe the emergent attention mechanism from the EOMs in Proposition 3.2. Specifically,

Proposition 3.3 (Emergent Attention Mechanism as EOMs in MHMs). Attention emerges as:

- If $\alpha \to \infty$, it corresponds to a hardmax function:

$$p_\mu = \text{hardmax}(z_\mu) = \begin{cases} 1\,, & \text{for maximal } z_\mu \\ 0\,, & \text{otherwise} \end{cases}.$$

- If $\alpha = 2$, it corresponds to a sparsemax function (Martins & Astudillo, 2016):

$$p_\mu = \text{Sparsemax}_\beta(z_\mu) = \frac{\beta}{2}[z_\mu - \theta]_+, \qquad \text{with } \frac{\beta}{2}\sum_{\mu=1}^{M}[z_\mu - \theta]_+ = 1.$$

- If $1 < \alpha < 2$, it corresponds to an entmax family (Peters et al., 2019).
- If $\alpha \to 1$, it corresponds to a softmax function (Blondel et al., 2020):

$$p_\mu = \text{Softmax}_\beta(z_\mu) = \frac{\exp\{\beta z_\mu\}}{\sum_{\nu=1}^{M}\exp\{\beta z_\nu\}}.$$

Proof. Please see Section D.7 for a detailed proof. □

Proposition 3.3 provides a re-derivation via EOMs of a free energy landscape, revealing the underlying physical interpretation of the attention mechanism. In Transformer language, (Vaswani et al., 2017), we denote $\text{Attn}(Q, K, V) := \sum_{\mu=1}^{M} p_\mu(Q, K)V_\mu$ as the representation of x, where Q represents the microscopic states, and K and V represent the stored memory patterns. Q, K, V are all fixed inputs. This corresponds to our zero variance or deterministic setting.

4 Phase Transitions and Critical Points of Retrieval Success

In this section, we discuss the static stability conditions and thresholds for retrieval success, where deformations of energy landscapes become crucial. For the dynamical stability, please see Section F.

Since the free energy depends on parameters such as β, M, d, and κ, they control the shape of the landscape. It is possible to have multiple local stable points for certain values of those parameters, inducing flows into a wrong retrieval state. This refers to the study of the phase structure. In our setting, the stored memories $\{\xi^\mu\}$ remain fixed, while the query state x evolves dynamically. Therefore, the phase structure and metastability arise from retrieval dynamics and the interaction between the input state and the energy landscape, rather than from training the memories themselves.

Definition 4.1 (Branch). Choose an initial parameter θ_0 that determines $\Gamma_\alpha^{(\text{mod})}[x(\theta_0), \xi]$. As $\theta \in \mathcal{I}_{\theta_0}$ varies from θ_0, we define a branch as a continuous family of the evolution from a specific local stable point $\{x_a^\star(\theta; \theta_0)\}_{\theta \in \mathcal{I}_{\theta_0}}$ satisfying the stationary condition, Definition 2.2,

$$\nabla_x \Gamma_\alpha^{(\text{mod})}[x_a^\star(\theta; \theta_0), \xi] = 0\,, \quad \text{and} \quad \nabla_x^2 \Gamma_\alpha^{(\text{mod})}[x_a^\star(\theta; \theta_0), \xi] \succ 0.$$

Branches help us analyze the evolution of stable points and the threshold for loss of stability. Notice that we formulate the dynamics in x. This differs from ∇_z due to $z_\mu = \sum_{i=1}^{d} \xi_i^\mu x_i$. The associated projection degeneracy leads to degenerate local minima. For a detailed analysis, please see Section F.

Next, we study the phase transitions of the generalized model (3.5) following phase stability analysis in statistical physics (Binder, 1987; Zinn-Justin, 2002). This starts with two definitions.

Definition 4.2 (Coexistence Point). Given two distinct branches, a, b, with their stable points, $\{x_a^\star(\theta; \theta_0)\}$, $\{x_b^\star(\theta; \theta_0)\}$. A coexistence point is a parameter value θ_{coexist} such that

$$\Gamma_\alpha^{(\text{mod})}[x_a^\star(\theta_{\text{coexist}}; \theta_0), \xi] = \Gamma_\alpha^{(\text{mod})}[x_b^\star(\theta_{\text{coexist}}; \theta_0), \xi]. \tag{4.1}$$

However, the coexistence condition is not a necessary condition for a real jump of local minima. Even if the equality holds, there is still an energy barrier that forbids the transition.

Definition 4.3 (Spinodal Point)**.** Given a smooth branch a, a spinodal point occurs when the branch loses its local stability. That is, the point where the stationary condition fails:

$$\lambda_{\min}(\nabla_x^2 \Gamma_\alpha^{(\text{mod})}[x_a^\star(\theta_{\text{jump}}; \theta_0), \xi]) = 0. \tag{4.2}$$

This is also the point where the continuation along that branch breaks down.

Remark 4.1. If a branch is non-smooth, e.g., there is a sudden change in the effective support size, then jumps also occur at that discontinuous point. In this case, (4.2) does not hold necessarily.

We control the system by the parameters, $\theta = (\beta, M, d, \xi)$. For $\alpha = 1$, the retrieval map is the softmax map (Proposition 3.3), so $p_\mu^\star$ is positive for all memories and the wights vary smoothly with $z^\star$. Hence, varying θ does not create any discontinuity through support change. Discontinuous changes of retrieval in the dense case ($\alpha = 1$) only occur if the Hessian vanishes, i.e., (4.2) holds. In this way, we expect more new transitions in the sparse case ($\alpha = 2$), and rare in the dense case. To investigate the transitions, we focus on the critical value of β in the following. We also analyze the exponential memory-load ($\kappa := \log(M)/d$) effect in Section 5 numerically.

Proposition 4.1 (Temperature Coexistence Point for Sparse Model)**.** Let a, b be two distinct branches and $\alpha = 2$. Denote k_a and k_b as the exact support sizes (see Lemma G.1) of branches a and b, respectively. Based on the statistical mechanics framework, the coexistence point satisfies

$$\sum_{\mu,\nu \in a} [z_\mu^\star - \theta]_+ [z_\nu^\star - \theta]_+ K_{\mu\nu}(\beta_{\text{coexist}}; \beta_0) = \sum_{\mu,\nu \in b} [z_\mu^\star - \theta]_+ [z_\nu^\star - \theta]_+ K_{\mu\nu}(\beta_{\text{coexist}}; \beta_0),$$

where $K_{\mu\nu}(\beta) := \mathbb{1}_{\{\mu,\nu\}} - \frac{\beta}{2} G_{\mu\nu}$, and $G_{\mu\nu} := \sum_{i=1}^d \xi_i^\mu \xi_i^\nu / d$ is the Gram matrix. β_{coexist} does not admit a closed-form expression. However, if the distribution is uniform on the simplex and the system satisfies the well-separation condition, i.e.,

$$G_{\mu\nu} = \begin{cases} 1 & , \ \mu = \nu \\ \epsilon \ll 1 & , \ \mu \neq \nu \end{cases}, \tag{4.3}$$

then for $k_a = 1$ and $k_b = 2$, we have the analytic solution $\beta_{\text{coexist}} = 2/(1 - \epsilon)$.

Proof. Please see Section D.8 for a detailed proof. □

Proposition 4.2 (Temperature Spinodal Point for Sparse Model)**.** Consider the sparse case $\alpha = 2$ and a branch a with exact support size k. Then the spinodal critical value is

$$\beta_{\text{jump}} = \frac{2}{\lambda_{\max}(PGP)},$$

where $P_{\mu\nu} := \mathbb{1}_{\{\mu,\nu\}} - k^{-1}$, and G is the Gram matrix. If the distribution is uniform and the system satisfies the well-separation condition, (4.3), then we have the analytic solution $\beta_{\text{jump}} = 2/(1 - \epsilon)$.

Proof. Please see Section D.9 for a detailed proof. □

Remark 4.2. Under the uniform distribution condition and the well-separation condition, we conclude that $\beta_{\text{coexist}} = \beta_{\text{jump}} = 2/(1 - \epsilon)$. However, we do not have these assumptions in general. Then the real jump and the coexistence point do not occur simultaneously, i.e., $\beta_{\text{jump}} \neq \beta_{\text{coexist}}$.

5 Numerical Results

We provide numerical evidence supporting the ideas in Section 4. Specifically, we examine the temperature and memory-load dependence of the phase structure in MHMs, under the sparse ($\alpha = 2$) and dense ($\alpha = 1$) regimes. For experimental setup, please see Section G.

Temperature Scan. Figure 2 illustrates the behavior of the maximal overlap $z_{\max}^\star$ under up- and down-sweeps of β. The two regimes exhibit distinct behaviors and characteristics.

- **Sparse regime** ($\alpha = 2$)**.** We observe hysteresis under up- and down-sweeps of β, which indicates coexistence of multiple stable retrieval states. The variance from seeds collapses at large β. This suggests concentration onto a single memory and seed-independent retrieval. This is consistent with single-pattern saturation $z_{\max}^\star \approx 1/d = 1/64 \approx 0.0156$. Notably, spinodal points occur in the sparse regime rarely. The observed jumps arise from discontinuous support change. Varying the initialization reveals $\beta_{\text{coexist}} \neq \beta_{\text{jump}}$. This is consistent with Remark 4.2.

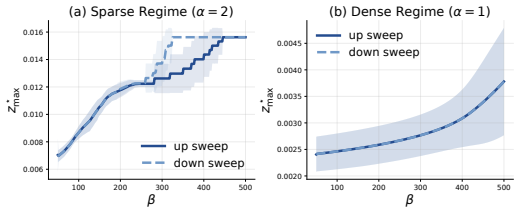


Figure 2: **Temperature dependence of the maximal overlap $z^\star_{\max}$ in the modern Hopfield model.** Patterns number $M = 100$. Pattern dimension $d = 64$. The shaded regions indicate standard deviation across different seeds. Up sweep corresponds to increasing β from low to high values using the previous solution as initialization, and vice versa for down sweep. **(a) Sparse regime** ($\alpha = 2$)**.** $z^\star_{\max}$ exhibits discontinuous jumps, and the up and down sweeps follow distinct paths. This indicates the coexistence of multiple local stable points and the structure of a discontinuous phase transition. **(b) Dense regime** ($\alpha = 1$)**.** $z^\star_{\max}$ varies with β smoothly, and the up and down sweeps coincide. This indicates a single retrieval without hysteresis. This behavior matches the absence of a spinodal instability.

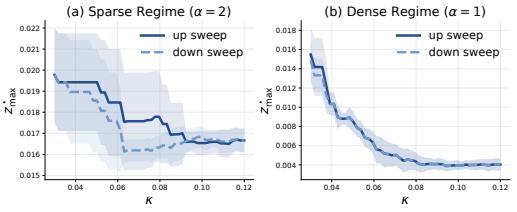


Figure 3: **Memory-load dependence of the maximal overlap $z^\star_{\max}$ in the modern Hopfield model.** Pattern dimension $d = 48$. Inverse temperature $\beta = 180$. Results are averaged over random pattern realizations, and the shaded regions indicate one standard deviation across different seeds. Up sweep corresponds to increasing $\kappa = \log(M)/d$ from low to high values using the previous solution as initialization, and vice versa for down sweep. **(a) Sparse regime** ($\alpha = 2$)**.** The up and down sweeps follow distinct branches, which indicates hysteresis and multiple coexisting local stable points. **(b) Dense regime** ($\alpha = 1$)**.** The up and down sweeps follow the same retrieval branch. Both cases suggest delocalization at high κ.

- **Dense regime** ($\alpha = 1$)**.** There is no hysteresis or spinodal event in β-scans. Hence, the system has a single retrieval branch, and its basin sharpens as β increases. However, the variance from seeds grows with β. This indicates sensitivity to seeds. Consequently, the system remains in a distributed mixture over patterns and does not collapse onto a single dominant pattern.

Our results show that the sparsity alters the phase structure of retrieval. Hence, support dynamics governs the phase behavior, rather than the curvature of the energy landscape alone.

Memory-Load Scan. Figure 3 illustrates the behavior of $z^\star_{\max}$ under the sweeps of κ.

- **Sparse regime** ($\alpha = 2$)**.** The up- and down-sweeps show a discrepancy, which suggests the hysteresis and multiple branches. Furthermore, in the intermediate-κ regime, there are discontinuous transitions due to support change. Finally, the maximal overlap $z^\star_{\max}$ decreases as κ grows. This reveals a trade-off between the high memory capability and retrieval reliability.
- **Dense regime** ($\alpha = 1$)**.** The maximal overlap $z^\star_{\max}$ decreases as κ grows as well. However, the up- and down-sweeps are indistinguishable throughout the scan. This suggests that increasing memory load introduces growing interference, rather than causing path-dependent retrieval branches. Again, this reveals a trade-off between high memory capability and retrieval reliability.

Our memory load experiment shows the capacity-reliability trade-off both in the sparse and dense cases. For the relation between support sizes and energies, please see Section G for more details.

6 Discussion and Conclusion

We present a free energy framework for Transformer attention via modern Hopfield models. This viewpoint identifies attention as the equation of motion on an energy landscape. Our work shows the empirical reliable retrieval regime with respect to entropy, temperature, and exponential memory load (i.e., $\alpha/\beta/\kappa$). We study the temperature and exponential memory load effects on phase structure. In the sparse regime, discontinuous transitions occur due to support change in memories (keys). The accompanying hysteresis indicates the coexistence of multiple stable states. In contrast, the softmax regime exhibits smooth evolution without hysteresis. Consequently, the softmax attention is a mixture of all patterns, while the query-key alignment grows with inverse-temperature smoothly. These results reveal the entropy-dependent phase behavior of modern Hopfield models and hence in Transformers.

Limitations. Our analysis relies on simplified assumptions, such as normalized patterns, prescribed correlation and noise structures, and the fixed memory setting. In contrast, realistic Transformer architectures allow interactions across heads and layers. This thereby leads to complicated noise. Furthermore, the keys and queries co-evolve through learned projection matrices and upstream layers. Hence, the theoretical analysis on phase transitions and retrieval reliability of full Transformers remain less explored. Extending our framework along these directions is crucial for future work.

We defer **practical implications** and **more experiments** to Sections A, G and H due to page limit.

REFERENCES

Shun-ichi Amari. *α-Divergence and α-Projection in Statistical Manifold*, pp. 66–103. Springer New York, New York, NY, 1985. ISBN 978-1-4612-5056-2. doi: 10.1007/978-1-4612-5056-2_3.

Daniel J. Amit, Hanoch Gutfreund, and H. Sompolinsky. Storing infinite numbers of patterns in a spin-glass model of neural networks. *Phys. Rev. Lett.*, 55:1530–1533, Sep 1985. doi: 10.1103/PhysRevLett.55.1530.

Daniel J Amit, Hanoch Gutfreund, and H Sompolinsky. Statistical mechanics of neural networks near saturation. *Annals of Physics*, 173(1):30–67, 1987. ISSN 0003-4916. doi: https://doi.org/10.1016/0003-4916(87)90092-3.

Andreas Auer, Martin Gauch, Daniel Klotz, and Sepp Hochreiter. Conformal prediction for time series with modern hopfield networks. *Advances in neural information processing systems*, 36: 56027–56074, 2023.

K Binder. Theory of first-order phase transitions. *Reports on Progress in Physics*, 50(7):783, jul 1987. doi: 10.1088/0034-4885/50/7/001.

Mathieu Blondel, André FT Martins, and Vlad Niculae. Learning with fenchel-young losses. *Journal of Machine Learning Research*, 21(35):1–69, 2020.

Sidney Coleman and Erick Weinberg. Radiative corrections as the origin of spontaneous symmetry breaking. *Phys. Rev. D*, 7:1888–1910, Mar 1973. doi: 10.1103/PhysRevD.7.1888.

Nima Dehmamy, Benjamin Hoover, Bishwajit Saha, Leo Kozachkov, Jean-Jacques Slotine, and Dmitry Krotov. Nrgpt: An energy-based alternative for gpt. *arXiv preprint arXiv:2512.16762*, 2025.

Amir. Dembo and Ofer. Zeitouni. *Large deviations techniques and applications*. Stochastic modelling and applied probability, 38. Springer, Berlin, 2nd ed. edition, 2010. ISBN 9783642033117.

Mete Demircigil, Judith Heusel, Matthias Löwe, Sven Upgang, and Franck Vermet. On a model of associative memory with huge storage capacity. *Journal of Statistical Physics*, 168(2):288–299, May 2017. ISSN 1572-9613. doi: 10.1007/s10955-017-1806-y.

Sacha Friedli and Yvan Velenik. *Statistical Mechanics of Lattice Systems: A Concrete Mathematical Introduction*. Cambridge University Press, 2017.

S. Furuichi, K. Yanagi, and K. Kuriyama. Fundamental properties of tsallis relative entropy. *Journal of Mathematical Physics*, 45(12):4868–4877, 12 2004. ISSN 0022-2488. doi: 10.1063/1.1805729.

Corrado Gini. *Variabilità e mutabilità*. Tipografia di Paolo Cuppini, Bologna, 1912.

M. O. Hill. Diversity and evenness: A unifying notation and its consequences. *Ecology*, 54(2): 427–432, 1973. doi: https://doi.org/10.2307/1934352.

Benjamin Hoover, Yuchen Liang, Bao Pham, Rameswar Panda, Hendrik Strobelt, Duen Horng Chau, Mohammed Zaki, and Dmitry Krotov. Energy transformer. *Advances in neural information processing systems*, 36:27532–27559, 2023.

J J Hopfield. Neural networks and physical systems with emergent collective computational abilities. *Proceedings of the National Academy of Sciences*, 79(8):2554–2558, 1982. doi: 10.1073/pnas.79.8.2554.

Jerry Yao-Chieh Hu, Donglin Yang, Dennis Wu, Chenwei Xu, Bo-Yu Chen, and Han Liu. On sparse modern hopfield model. *Advances in neural information processing systems*, 36:27594–27608, 2023.

Jerry Yao-Chieh Hu, Pei-Hsuan Chang, Haozheng Luo, Hong-Yu Chen, Weijian Li, Wei-Po Wang, and Han Liu. Outlier-efficient hopfield layers for large transformer-based models. In *International Conference on Machine Learning*. PMLR, 2024a.

Jerry Yao-Chieh Hu, Dennis Wu, and Han Liu. Provably optimal memory capacity for modern hopfield models: Transformer-compatible dense associative memories as spherical codes. *Advances in Neural Information Processing Systems*, 37:70693–70729, 2024b.

Diederik P Kingma and Max Welling. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013.

Scott Kirkpatrick. Frustration and ground-state degeneracy in spin glasses. *Phys. Rev. B*, 16: 4630–4641, Nov 1977. doi: 10.1103/PhysRevB.16.4630.

Felix Koulischer, Cédric Goemaere, Tom Van Der Meersch, Johannes Deleu, and Thomas Demeester. Exploring the temperature-dependent phase transition in modern hopfield networks. *arXiv preprint arXiv:2311.18434*, 2023.

Dmitry Krotov. A new frontier for hopfield networks. *Nature Reviews Physics*, 5, 05 2023. doi: 10.1038/s42254-023-00595-y.

Dmitry Krotov and John Hopfield. Dense associative memory is robust to adversarial inputs. *Neural Computation*, 30(12):3151–3167, 12 2018. ISSN 0899-7667. doi: 10.1162/neco_a_01143.

Dmitry Krotov and John Hopfield. Large associative memory problem in neurobiology and machine learning. *arXiv preprint arXiv:2008.06996*, 2020.

Solomon Kullback and Richard A Leibler. On information and sufficiency. *The annals of mathematical statistics*, 22(1):79–86, 1951.

L. D. Landau and E. M. Lifshitz. *Statistical Physics, Part 1*, volume 5 of *Course of Theoretical Physics*. Butterworth-Heinemann, Oxford, 1980. ISBN 978-0-7506-3372-7. doi: 10.1016/c2009-0-24487-4.

Jason D Lee, Max Simchowitz, Michael I Jordan, and Benjamin Recht. Gradient descent converges to minimizers. *arXiv preprint arXiv:1602.04915*, 2016.

Andre Martins and Ramon Astudillo. From softmax to sparsemax: A sparse model of attention and multi-label classification. In *International conference on machine learning*, pp. 1614–1623. PMLR, 2016.

M Mezard, G Parisi, and M Virasoro. *Spin Glass Theory and Beyond*. WORLD SCIENTIFIC, 1986. doi: 10.1142/0271.

Beren Millidge, Tommaso Salvatori, Yuhang Song, Thomas Lukasiewicz, and Rafal Bogacz. Universal hopfield networks: A general framework for single-shot associative memory models. In Kamalika Chaudhuri, Stefanie Jegelka, Le Song, Csaba Szepesvari, Gang Niu, and Sivan Sabato (eds.), *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pp. 15561–15583. PMLR, 17–23 Jul 2022.

Moshe Moshe and Jean Zinn-Justin. Quantum field theory in the large n limit: a review. *Physics Reports*, 385(3–6):69–228, October 2003. ISSN 0370-1573. doi: 10.1016/s0370-1573(03)00263-1.

Jan Naudts. Generalized thermostatistics and mean-field theory. *Physica A: Statistical Mechanics and its Applications*, 332:279–300, February 2004. ISSN 0378-4371. doi: 10.1016/j.physa.2003.10.013.

Hidetoshi Nishimori. In *Statistical Physics of Spin Glasses and Information Processing: An Introduction*. Oxford University Press, 07 2001. ISBN 9780198509417. doi: 10.1093/acprof:oso/9780198509417.001.0001.

Lars Onsager. Crystal statistics. i. a two-dimensional model with an order-disorder transition. *Phys. Rev.*, 65:117–149, Feb 1944. doi: 10.1103/PhysRev.65.117.

Ben Peters, Vlad Niculae, and André FT Martins. Sparse sequence-to-sequence models. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pp. 1504–1519, 2019.

Hubert Ramsauer, Bernhard Schäfl, Johannes Lehner, Philipp Seidl, Michael Widrich, Thomas Adler, Lukas Gruber, Markus Holzleitner, Milena Pavlović, Geir Kjetil Sandve, et al. Hopfield networks is all you need. *arXiv preprint arXiv:2008.02217*, 2020.

R Tyrrell Rockafellar. *Convex analysis*, volume 28. Princeton university press, 1997.

Saul Santos, Vlad Niculae, Daniel McNamee, and André FT Martins. Hopfield-fenchel-young networks: A unified framework for associative memory retrieval. *Journal of Machine Learning Research*, 26(265):1–51, 2025.

Claude E. Shannon. A mathematical theory of communication. *Bell System Technical Journal*, 27(3): 379–423, 1948.

David Sherrington and Scott Kirkpatrick. Solvable model of a spin-glass. *Phys. Rev. Lett.*, 35: 1792–1796, Dec 1975. doi: 10.1103/PhysRevLett.35.1792.

M. Srednicki. *Quantum field theory*. Cambridge University Press, 1 2007. ISBN 978-0-521-86449-7, 978-0-511-26720-8. doi: 10.1017/CBO9780511813917.

D. J. Thouless, P. W. Anderson, and R. G. Palmer. Solution of 'solvable model of a spin glass'. *The Philosophical Magazine: A Journal of Theoretical Experimental and Applied Physics*, 35(3): 593–601, 1977. doi: 10.1080/14786437708235992.

Constantino Tsallis. Possible Generalization of Boltzmann-Gibbs Statistics. *J. Statist. Phys.*, 52: 479–487, 1988. doi: 10.1007/BF01016429.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

Michael Widrich, Bernhard Schäfl, Milena Pavlović, Hubert Ramsauer, Lukas Gruber, Markus Holzleitner, Johannes Brandstetter, Geir Kjetil Sandve, Victor Greiff, Sepp Hochreiter, et al. Modern hopfield networks and attention for immune repertoire classification. *Advances in neural information processing systems*, 33:18832–18845, 2020.

Dennis Wu, Jerry Yao-Chieh Hu, Weijian Li, Bo-Yu Chen, and Han Liu. Stanhop: Sparse tandem hopfield model for memory-enhanced time series prediction. *arXiv preprint arXiv:2312.17346*, 2023.

Dennis Wu, Jerry Yao-Chieh Hu, Teng-Yun Hsiao, and Han Liu. Uniform memory retrieval with larger capacity for modern hopfield models. In *International Conference on Machine Learning*, pp. 53471–53514. PMLR, 2024.

Yongyi Yang, zengfeng Huang, and David P Wipf. Transformers from an optimization perspective. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh (eds.), *Advances in Neural Information Processing Systems*, volume 35, pp. 36958–36971. Curran Associates, Inc., 2022.

Jean Zinn-Justin. Quantum field theory and critical phenomena. *Int. Ser. Monogr. Phys.*, 113:1–1054, 2002.

Appendix

A Practical Implications. **14**

B Related Work **14**

C A Minimal Introduction of Statistical Mechanics **15**

C.1 From Interacting Spins to Associative Memory 15

C.2 A Concrete Hopfield Free Energy 16

C.3 Landscapes, Metastability, and Retrieval Jumps 17

D Definitions and Proofs in the Main Text **18**

D.1 Basic Tools for Convex Analysis 18

D.2 Proof of Proposition 2.1 19

D.3 Tsallis Entropy 21

D.4 Proof of Theorem 2.1 24

D.5 Proof of Proposition 3.1 25

D.6 Proof of Theorem 3.1 26

D.7 Proof of Proposition 3.3 28

D.8 Proof of Proposition 4.1 29

D.9 Proof of Proposition 4.2 30

E Fluctuated Version and Thermal Attention **32**

E.1 Physics of Temperature 32

E.2 Fluctuated Variables in Modern Hopfield Model 34

E.3 Connection between Free Energy and Tsallis Entropy in Fluctuated Case 35

E.4 Retrieval Dynamics in Fluctuated Case and Thermal Attention 37

F Dynamical Retrieval Stability Analysis **43**

G Numerical Studies **47**

G.1 Effective Support Size and Sparsity 47

G.2 Settings. 49

G.3 Temperature Analysis ($\alpha = 2$) 50

G.4 Temperature Analysis ($\alpha = 1$) 51

G.5 Exponential Memory Load Analysis ($\alpha = 2$) 53

G.6 Exponential Memory Load Analysis ($\alpha = 1$) 54

G.7 Dynamical Stability Analysis for Temperature 55

H Additional Experiments **58**

H.1 Support and Capacity–Reliability Trade-off 58

H.2 Support, Temperature, and Reliability 61

H.3 Numerical Validation of Coexistence and Spinodal Formula 63

Impact Statement

By the theoretical nature of this work, we do not anticipate any negative social impact.

LLM Usage Disclosure

We used large language models (LLMs) to aid, polish writing and debug codes, such as improving clarity, grammar, and conciseness. We also used LLMs for retrieval and discovery, for example, exhausting literature to identify potential missing related work. All technical content, proofs, experiments, and results are original contributions by the authors.

A Practical Implications.

We sketch some design principles from our results:

- First, the attention should match the task: sparse one is support-efficient, while dense one is more accurate when many relevant values contribute (Proposition 3.3 and Sections H.1 and H.2).
- Second, the retrieval depends on path under temperature changes. Proposition 4.1 and Figures 2 and 8 to Figure 11, and Figure 14 to Figure 17) Specifically, increasing versus decreasing β lead to different different memory selection in the sparse case. The outcome also depends on the initial query (i.e., input context or prompt) supplied to the retrieval update during inference.
- Third, increasing the inverse-temperature strengthens the maximal query-key similarity $z_{\max} = \max_\mu \sum_{i=1}^{d} x_i \xi_i^\mu / \sqrt{d}$ (Figures 8 and 9). In particular, sparse attention reaches single-memory retrieval saturation, while dense one sharpens smoothly with full support.
- Finally, increasing memory load reduces retrieval accuracy (Figures 10 and 11). This further reveals a capacity-reliability trade-off in the realistic Transformer architectures (Section H.1).

Together, these provide concrete guidance for selecting attention, temperature ane memory-load.

B Related Work

Classical and Modern Hopfield Models. Hopfield-type models originate from the physical interpretation of associative memory in spin systems, including the Ising model (Onsager, 1944), the Sherrington-Kirkpatrick model (Sherrington & Kirkpatrick, 1975), and the classical Hopfield model (Hopfield, 1982). Modern Hopfield models modify the energy function so that stored associative memories become local minima on the energy landscape, leading to a higher memory capacity (Krotov & Hopfield, 2020).

Attention mechanisms (Vaswani et al., 2017) have been highly successful in large-scale architectures. Ramsauer et al. (2020) establishes the connection between softmax modern Hopfield retrieval and Transformer attention. Hu et al. (2023) further extends the softmax to sparse retrieval and sparse attention by applying convex conjugate. Peters et al. (2019) introduce the entmax family of attention maps, which interpolates between dense and sparse attention. Complementarily, Millidge et al. (2022) develop universal Hopfield networks. It formulates a retrieval dynamics through three operations: similarity, separation and projection. Santos et al. (2025) systematize this line of work into a Hopfield-Fenchel-Young framework, covering Fenchel-Young losses and the variational formulation of retrieval.

Free Energy. Our formulation adopts the variational principles used in physics. Specifically, this method appears in quantum field theory (Zinn-Justin, 2002), large-N mean-field theory (Moshe & Zinn-Justin, 2003), and low-temperature theory (Landau & Lifshitz, 1980). This viewpoint describes the retrieval process as a variational problem over macroscopic parameters, rather than microscopic states. Related optimization-based interpretations of Transformer architectures are also explored in (Yang et al., 2022), where Transformer layers are viewed as unfolding steps of an energy minimization process. Unlike traditional analyses of Hopfield networks that focus on energy minimization alone, the free energy framework incorporates entropy, leading to a richer landscape with multiple metastable states.

Phase Transition. Physicists often characterize phase transitions in energy-based systems through spontaneous symmetry breaking (Coleman & Weinberg, 1973) and the emergence of metastable states. In classical Hopfield networks, Amit et al. (1985; 1987) and Mezard et al. (1986) analyze this behavior using replica methods. This reveals transitions between the retrieval and spin-glass phases, which contain numerous spurious states that lead to retrieval failure. Free energy provides a description of phase transitions through the deformation of energy landscapes and the emergence of local minima (Binder, 1987). Recent work from Koulischer et al. (2023) discusses the phase transition in the softmax case of modern Hopfield models. They identify a critical inverse temperature at which the mean minima are close to individual stored patterns, rather than one common averaged minimum. However, their analysis does not include sparse retrieval map and the entmax family.

Dense Associative Memory. Dense associative memory models extend classical Hopfield models by introducing higher-order non-linear energy functions, which improved memory load storage in energy-based models (Demircigil et al., 2017; Krotov, 2023). These interactions further reshape the energy landscape and enhance robustness to noise (Krotov & Hopfield, 2018).

Fuzzy Memory. Modern Hopfield models admit distributed retrieval over multiple memories. Recent work studies how such retrieval behavior relates to memory capacity. Specifically, Wu et al. (2024) introduce a learnable feature map that separates stored patterns in a kernel space. This makes the distribution of memory minima more uniform. On the other hand, Hu et al. (2024b) consider spherical codes, i.e., stored memories in the kernel feature space as points on a hypersphere. This induces a tight capacity bound by analyzing how many feature vectors maintain angular separation on a hypersphere. These works specify regimes where multiple patterns are activated simultaneously. Such distributed states interpolate between sparse (single-pattern) and dense (multi-pattern) retrieval regimes.

Despite these prior works, a unified statistical framework of modern Hopfield models remains limited. Existing analyses focus on either retrieval capacity or dynamical behavior, without a systematic treatment of entropy, sparsity, and metastability within a single framework. Furthermore, replica analyses characterize transitions between retrieval, spin-glass, and disordered phases, but not capture the detailed finite-dimensional geometry for modern Hopfield retrieval. In this work, we introduce a free energy formulation that unifies these aspects and study the induced finite-dimensional retrieval landscape. In particular, we focus on phase behavior inside the retrieval regime, rather than only transitions between retrieval and disordered phases. This allows us to compare dense and sparse regime, analyze metastable states, and verify retrieval success. In addition, we analyze how exponential memory load κ changes empirical retrieval behaviors. This probes memory-dependent phase regimes which are not addressed by existing Fenchel-Young or sparse Hopfield analyses.

C A Minimal Introduction of Statistical Mechanics

This appendix gives a physical route from interacting spins to associative memory and then to attention. We set Boltzmann's constant to one, so the inverse temperature is $\beta = 1/T$.

C.1 From Interacting Spins to Associative Memory

A Hamiltonian assigns an energy to each microscopic configuration; a choice of dynamics specifies how the system moves through that landscape.

Definition C.1 (Classical Hopfield Model (Hopfield, 1982))**.** Consider binary spins $x_i \in \{-1, +1\}$ with couplings $J_{ij} = d^{-1} \sum_{\mu=1}^{M} \xi_i^\mu \xi_j^\mu$ and $J_{ii} = 0$. This gives the classical Hopfield model

$$\mathcal{H}_{\text{Hop}}(x) = -\frac{1}{2d} \sum_{i \neq j}^{d} \sum_{\mu=1}^{M} \xi_i^\mu \xi_j^\mu x_i x_j, \tag{C.1}$$

where memories are denoted as $\xi^1, \ldots, \xi^M \in \{-1, +1\}^d$.

Thus, zero-temperature Hopfield dynamics is a dissipative descent toward a local energy minimum. At nonzero temperature the system instead fluctuates among configurations, as described in Section C.2.

Definition C.2 (Overlap)**.** The overlap between the current state and memory μ is

$$z_\mu(x) \coloneqq \frac{1}{\sqrt{d}} \sum_{i=1}^{d} \xi_i^\mu x_i. \tag{C.2}$$

Therefore, low energy means large overlap with one or more stored patterns.

Definition C.3 (Frustration (Kirkpatrick, 1977))**.** A system is frustrated if it does not satisfy $J_{ij} x_i x_j > 0$ for every pair with $J_{ij} \neq 0$.

A one-pattern Hopfield model is equivalent to an ordinary ferromagnet and retrieves without frustration. However, with several patterns, the pairwise alignments become incompatible. This creates frustration and additional metastable minima. Hence, stored patterns are attractors, but not every local minimum must be a stored memory. This is the elementary mechanism of associative retrieval shown in Figure 4.

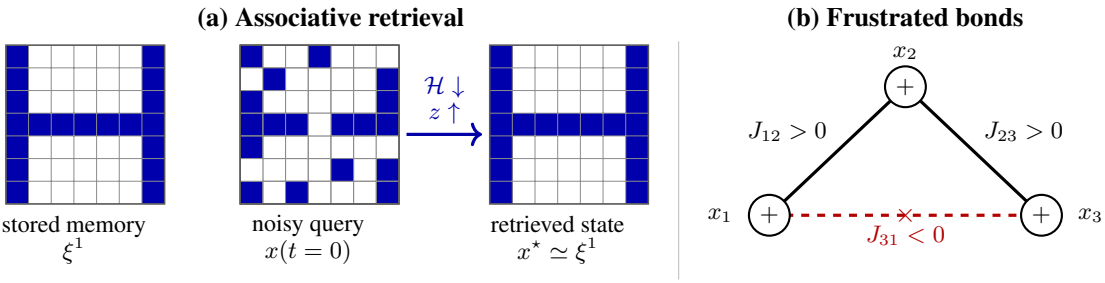


Figure 4: **Retrieval and frustration in the classical Hopfield picture. (a)** From a noisy query set, we retrieve a stored pattern by minimizing $\mathcal{H}$. Equivalently, the overlap increases during this associative retrieval. **(b)** Two ferromagnetic bonds and one antiferromagnetic bond on a triangle is impossible to satisfy pairwise energy minimization simultaneously. This is the simplest illustration of frustration.

C.2 A Concrete Hopfield Free Energy

In this section, we take one-pattern Hopfield model for a self-contained example of the statistical mechanics construction. That is, we choose $\mathcal{H}_{\text{1pat}}(x) = -\frac{1}{2}z^2(x)$. At temperature $T = 1/\beta$, the Gibbs distribution and partition function (Definition 2.1) are

$$p_J(x) = \frac{e^{-\beta\mathcal{H}_{\text{1pat}}(x)+J^\top x}}{Z(J)}, \qquad Z(J) = \sum_{x\in\{-1,+1\}^d} e^{-\beta\mathcal{H}_{\text{1pat}}(x)+J^\top x}. \tag{C.3}$$

By grouping states via Gibbs distribution and Fenchel conjugate (Definitions 2.3 and D.6),

$$\Gamma(z) = \sup_J \left\{J^\top\varphi - \log Z(J)\right\},$$

we obtain the exact form of one-pattern Hopfield free energy.

Proposition C.1 (Free Energy of One-Pattern Hopfield Model (Friedli & Velenik, 2017))**.** The free energy of a one-pattern Hopfield model is

$$\Gamma(z) = -\frac{1}{2}z^2 - \frac{1}{\beta}\log\binom{d}{\frac{d+\sqrt{d}z}{2}}.$$

In the large-d limit, it approaches to an explicit expression

$$\Gamma(z) = \beta\big(U - TS\big) \xrightarrow{d\to\infty} \Gamma_\infty(q) = -\frac{\beta}{2}q^2 - s(q), \tag{C.4}$$

where $U = -\frac{1}{2}\mathbb{E}_{p_J}[z^2]$ is the internal energy, S is the entropy (Definition D.8) in the statistical mechanics language, $q := \lim_{d\to\infty}\mathbb{E}_{p_J}[z]$ and

$$s(q) := -\frac{d+q\sqrt{d}}{2}\log\frac{1+q/\sqrt{d}}{2} - \frac{d-q\sqrt{d}}{2}\log\frac{1-q/\sqrt{d}}{2}. \tag{C.5}$$

Proof. Please see (Friedli & Velenik, 2017) for the detailed proof. □

Proposition C.2 (Stationary Equation of One-Pattern Hopfield Model (Friedli & Velenik, 2017))**.** The stationary condition of the large-d one-pattern Hopfield free energy is

$$q^\star = \sqrt{d}\tanh\left(\frac{\beta q^\star}{\sqrt{d}}\right). \tag{C.6}$$

This is an implicit equation. The solutions depend on inverse-temperature β.

Proof. Taking the derivative , we have

$$0 = \Gamma'_\infty(q^\star) = -\beta q^\star + \frac{\sqrt{d}}{2}\log\frac{1+q^\star/\sqrt{d}}{1-q^\star/\sqrt{d}}.$$

This hence gives the stationary points $q^\star$ and proves the proposition. □

At the origin, $\Gamma''_\infty(q^\star = 0) = -\beta + 1$. Therefore, as illustrated in Figure 5:

- for $\beta < 1$ (high temperature), $q^\star = 0$ is the unique minimum. However, $q^\star = \lim_{d\to\infty}\mathbb{E}_{p_J}[z] = 0$ indicates disorderness, i.e., zero overlap.
- at $\beta_c = 1$, the Hessian at the origin vanishes.
- for $\beta > 1$ (low temperature), $q^\star = 0$ becomes unstable and two minima $q^\star_\pm \neq 0$ appear. Since the local minima are symmetric, the retrieval is still unique and stable.

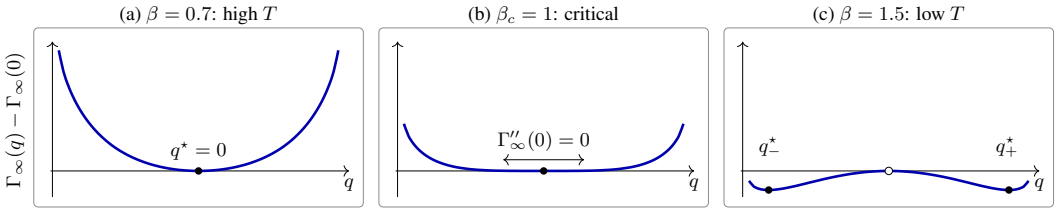


Figure 5: **One-pattern free-energy landscape in the large-d limit.** The curves show the one-pattern Hopfield model free energy and its local minima. Cooling the system turns the minimum at $q^\star = 0$ into two retrieval minima.

The deformation of the energy landscape moves the local minimum. It leads to a phase transition.

C.3 Landscapes, Metastability, and Retrieval Jumps

The one-pattern transition above is continuous: the nonzero minima emerge from $q^\star = 0$ continuously. More complicated retrieval landscapes contain several separated local minima. Specifically, two minima coexist when they have equal free energy. If one is higher, then the minimum is metastable since a barrier prevents movement to the lower state. A system in one basin remains there until that basin loses local stability. It moves only when the energy barrier vanishes, which is called a *spinodal* (Definition 4.3). We illustrate this mechanism in Figure 6.

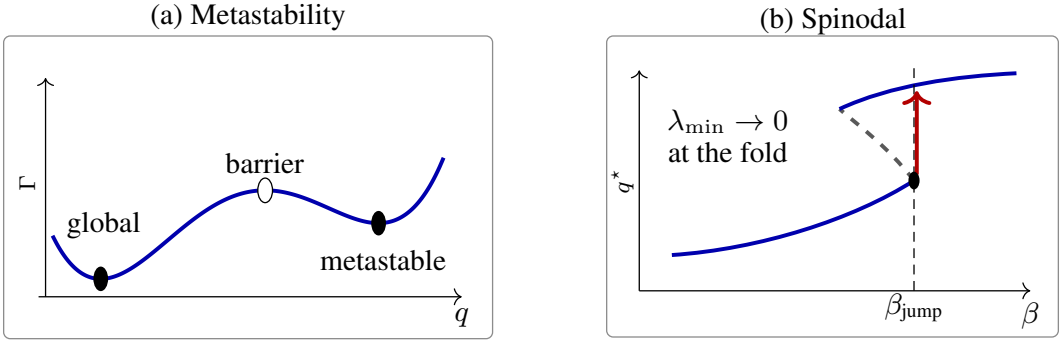


Figure 6: **Phase-transition picture.** **(a)** A metastable minimum is protected by a barrier. **(b)** We plot two smooth branches. In one smooth branch, the minimal Hessian eigenvalue vanishes at a spinodal. This leads to a jump onto another branch.

D Definitions and Proofs in the Main Text

D.1 Basic Tools for Convex Analysis

Here we define several fundamental terminologies from standard convex analysis (Rockafellar, 1997).

Definition D.1 (Proper)**.** A function f is proper if the support domain of f satisfies

$$\mathrm{dom}(f) \coloneqq \{x \in \mathbb{R}^d : f(x) < \infty\} \neq \emptyset, \qquad \text{and} \quad f \neq +\infty.$$

Definition D.2 (Lower Semi-Continuous (lsc))**.** A function f is lower semi-continuous (lsc) if

$$f(x) \leq \lim \inf_{\tau \to \infty} f(x_\tau),$$

for every sequence $(x_i)_\tau \to x_i$.

Definition D.3 (Closed Convex)**.** A function f is closed convex if f is proper, lsc and convex.

Definition D.4 (Essentially Smooth)**.** A function f is essentially smooth if f is differentiable on int dom(f), and $\|\nabla f(x)\| \to \infty$ whenever $x_i \to \bar{x}_i \in \partial(\mathrm{dom}(f))$.

Definition D.5 (Subgradient and Subdifferential)**.** Let $f : \mathbb{R}^d \to [-\infty, +\infty)$ be a proper convex function. A vector $g \in \mathbb{R}^d$ is called a subgradient of f at $x \in \mathrm{dom}(f)$ if

$$f(y) \geq f(x) + \sum_{i=1}^{d} g_i(y_i - x_i) \qquad \forall\, y \in \mathbb{R}^d.$$

The set of all subgradients of f at x is called the subdifferential of f at x, and is denoted by

$$\partial f(x) = \{g \in \mathbb{R}^d : f(y) \geq f(x) + \sum_{i=1}^{d} g_i(y_i - x_i)\ \forall\, y \in \mathbb{R}^d\}.$$

If f is differentiable at x, then the subdifferential reduces to $\partial f(x) = \{\nabla f(x)\}$.

Definition D.6 (Fenchel Conjugation (Rockafellar, 1997)). Consider a function $f : \mathbb{R}^d \to \mathbb{R}$ and an input $x \in \mathbb{R}^d$. Then we define the Fenchel conjugate of $f(x)$ by the following variational formula,

$$f^\star(k) := \sup_{x \in \mathbb{R}^d} \{\sum_{i=1}^d k_i x_i - f(x)\} \in \mathbb{R}\,, \qquad k \in \mathbb{R}^d.$$

Fact 1 (Fenchel Conjugation Properties (Rockafellar, 1997)). Definition D.6 has several properties:

1. **Convexity.** A Fenchel conjugate $f^\star(x)$ is by definition a convex function.
2. **Involutivity.** If a function f is closed convex, then applying Fenchel conjugation twice yields the identity map, i.e., $f^{\star\star} = f$ for all $x \in \mathbb{R}^d$. This is the well-known Fenchel-Moreau Theorem.
3. **Bijectiveness.** If a function f is proper, lsc, strictly convex and essentially smooth, then $\nabla f^\star$: int dom$(f^\star) \leftrightarrow$ int dom(f) is a bijection, where int and dom denote interior and domain.

Definition D.7 (Legendre Transformation). A function f is called a Legendre function if it is proper, lsc, strictly convex and essentially smooth. In other words, Legendre functions satisfy involutivity and bijectiveness. In this case, we refer to the Fenchel conjugate $f^\star$ as the Legendre transformation.

Remark D.1. A Legendre function f satisfies the Fenchel-Moreau involutivity $f^{\star\star} = f$, and its gradient map ∇f is a bijection between the interior domains from Fact 1 automatically.

Remark D.2. Here we adopt the definition of the Legendre transformation from (Rockafellar, 1997). However, in the physics and machine learning literature, for example, in (Krotov & Hopfield, 2020) and (Ramsauer et al., 2020), people denote the same conjugation without verifying the conditions. Specifically, in many cases, the domain is $\mathbb{R}^d$, so $\partial\mathbb{R}^d = \emptyset$ leads to the essential smoothness condition being vacuous. The real issue is the non-bijectiveness of gradient maps. We should understand the transformations in these works as the Fenchel conjugations rather than the true Legendre transforms.

D.2 Proof of Proposition 2.1

Proposition D.1 (Proposition 2.1 Restated: Fenchel-Young Duality Relation (Srednicki, 2007; Blondel et al., 2020)). Let p_J be the Gibbs measure with external source J. From the standard statistical field theory, the induced macroscopic field $\varphi = \mathbb{E}_{p_J}[x]$ reveals the following source-field relation:

$$\varphi = \mathbb{E}_{x \sim p_J}[x] \in \partial W[J] \;\Leftrightarrow\; J \in \partial \Gamma[\varphi]\,,$$

where ∂ is the subdifferential from Definition D.5. If W and Γ are differentiable, then ∂ reduce to gradients ∇. We call them equations of motion (EOMs) in the physics literature. It also refers to the Fenchel-Young duality relation. For φ and J, we call them the classical solutions or saddle points.

Proof. We repeat the proof from (Blondel et al., 2020; Rockafellar, 1997) for completeness. By definition, $\Gamma[\varphi] = \sup_{J \in \mathbb{R}^d}\{\sum_{i=1}^d J_i\varphi_i - W[J]\}$. Therefore, for any J, $\Gamma[\varphi] \geq \sum_{i=1}^d J_i\varphi_i - W[J]$. By the definition of subgradient Definition D.5, we have

$$W[J] \geq W[J^\star] + \sum_{i=1}^d \varphi_i(J_i - J_i^\star), \qquad \forall J \in \mathbb{R}^d\ \forall \varphi \in \partial W[J^\star].$$

Then

$$\sum_{i=1}^d J_i^\star \varphi_i - W[J^\star] \geq \sum_{i=1}^d J_i\varphi_i - W[J] \qquad \forall J \in \mathbb{R}^d.$$

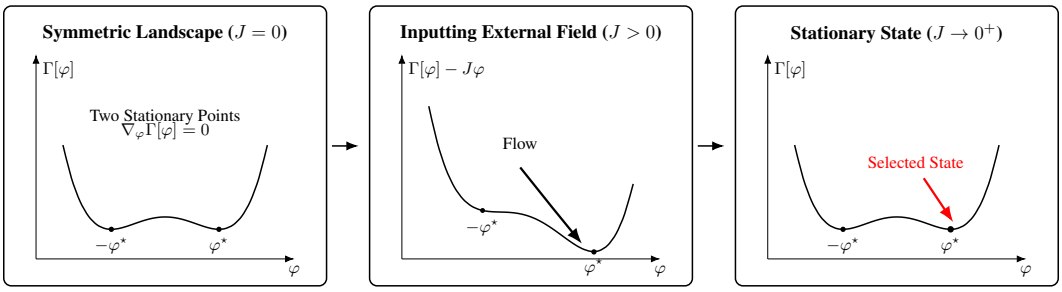


Figure 7: **Selection of a stationary state by an external source.** This figure illustrates how a source J selects one state from multiple degenerate stationary states of the free energy $\Gamma[\varphi]$. **Left.** When $J = 0$, $\Gamma[\varphi]$ is symmetric and has two degenerate local minima $\pm\varphi^\star$. **Middle.** Turning on a small external source $J > 0$ tilts the landscape to $\Gamma[\varphi] - J\varphi$. This breaks the symmetry so that $+\varphi^\star$ becomes preferred. Then ,the system flows toward the selected basin. **Right.** After the selection, we take the limit $J \to 0^+$. Although the explicit source is removed, the system remains in the chosen local minimum $+\varphi^\star$ satisfying $\nabla_\varphi \Gamma[+\varphi^\star] = 0$. This source-selection procedure resolves the degeneracy between degenerate stationary states and specifies the retrieval state.

Hence, $J^\star$ is a maximizer, and therefore $J^\star \in \partial\Gamma[\varphi]$. Conversely, if we assume that $J^\star \in \partial\Gamma[\varphi^\star]$, then

$$\Gamma[\varphi] \geq \Gamma[\varphi^\star] + \sum_{i=1}^{d} J_i^\star(\varphi_i - \varphi_i^\star) \qquad \forall \varphi \in \mathbb{R}^d.$$

In this way,

$$\sum_{i=1}^{d} J_i^\star \varphi_i^\star - \Gamma[\varphi^\star] \geq \sum_{i=1}^{d} J_i^\star \varphi_i - \Gamma[\varphi] \qquad \forall \varphi \in \mathbb{R}^d.$$

Note that $W[J]$ is closed due to the form (2.2). Then by involutivity in Fact 1, we obtain

$$W[J^\star] = W^{\star\star}[J^\star] = \sup_{\varphi \in \mathbb{R}^d} \{\sum_{i=1}^{d} \varphi_i J_i^\star - \Gamma[\varphi]\} \geq \sum_{i=1}^{d} \varphi_i J_i^\star - \Gamma[\varphi] \;\; \forall \varphi \in \mathbb{R}^d.$$

Hence, $\varphi^\star$ is a maximizer, and therefore $\varphi^\star \in \partial W[J^\star]$. After removing the redundant notation $\star$, we arrive the Fenchel-Young duality relation. □

Remark D.3. The non-zero equation, $\nabla_J W(J)$, reflects a symmetry breaking of a free theory due to the external input. For example, consider a spin system. Every spin has a random orientation, $S \in \{-1, +1\}^d$, and there exists a $\mathbb{Z}_2$ symmetry that flips the entire system upside down. Then,

$$\nabla_J W[J]|_{J=0} = \frac{S \exp\{-\beta\mathcal{H}[S] + \sum_{i=1}^{d} J_i S_i\}}{Z[J]}|_{J=0} = \mathbb{E}_{S\sim p(S)}[S] = 0\,, \tag{D.1}$$

where $p(S) \propto \exp\{-\beta\mathcal{H}[S]\}$ up to a normalization factor. This gives the macroscopic variable $\varphi = 0$. However, if we input the external source, $J \neq 0$, then the spin system favors one orientation. This corresponds to a response in a ferromagnet. After this, we approach $J \to 0^+$ as a stationary condition for $\Gamma[\varphi]$. The system stays at one specific local minima. See Figure 7 for the illustration. We shall see that in the sparse modern Hopfield models (Section 3), we always require this response, $\varphi = \nabla_J W[J] \neq 0$, to realize the retrieval states.

Remark D.4. From the Fenchel-Young duality relation, Proposition 2.1, we know that a gradient descent trajectory in J-space has one-to-one mapping to a unique trajectory in φ-space. However, the induced unique trajectory is not a gradient flow. Specifically, suppose we define gradient descent as

$\varphi = \dot{J} = -\nabla_J W[J]$, where $\dot{J} \coloneqq dJ/dt$ if we view update steps as infinitesimal. Then

$$\dot{\varphi} = \nabla_J^2 W[J](-\nabla_J W[J]) = -\varphi \nabla_J^2 W[J] \,,$$

which is not the gradient flow, $\dot{\varphi} = -\nabla_\varphi \Gamma[\varphi]$.

D.3 Tsallis Entropy

In this section, we introduce the Tsallis entropy and its role in statistical mechanics.

Definition D.8 (Tsallis Entropy (Tsallis, 1988))**.** We define the Tsallis entropy, (Tsallis, 1988), of the form

$$\Psi_\alpha[p] \coloneqq \frac{1}{\alpha - 1}(1 - \sum_{\mu=1}^{M} p_\mu^\alpha) \,, \tag{D.2}$$

where $\sum_{\mu=1}^{M} p_\mu = 1$, $\alpha \in \mathbb{R}_+/\{1\}$.

Remark D.5. The maximal Tsallis entropy occurs when the probability function is uniform, i.e., $p_\mu = M^{-1}$ for every μ. Hence, if the system is random, then the entropy becomes maximal.

Proposition D.2 (Types of Entropy (Tsallis, 1988))**.** Specific values of α in Tsallis entropy reduce to other types of entropy.

- If $\alpha \to 1$, it reduces to the Shannon entropy (Shannon, 1948).

$$S[p] \coloneqq -\sum_{\mu=1}^{M} p_\mu \log(p_\mu). \tag{D.3}$$

- If $\alpha \to 2$, it reduces to the Gini entropy (Gini, 1912).

$$G[p] \coloneqq 1 - \sum_{\mu=1}^{M} p_\mu^2. \tag{D.4}$$

- If $\alpha \to \infty$, it is zero.

Proof. We repeat the proof from (Tsallis, 1988) for completeness. We prove them one-by-one as follows.

- Applying the l'Hôpital's rule for $0/0$-form, we have

$$\Psi_{\alpha\to 1}[p] = -\lim_{\alpha\to 1} \sum_{\mu=1}^{M} \log(p_\mu) p_\mu^\alpha = -\sum_{\mu=1}^{M} p_\mu \log(p_\mu) \coloneqq S[p] \,.$$

  This is the form of Shannon entropy.

- For $\alpha \to 2$, the Tsallis entropy function is by definition the Gini entropy.

- Let $p_{\max} \coloneqq \max_\mu \; p_\mu$ and $\mu_{\max}$ be the corresponding maximal index. Then we separate the Tsallis entropy as

$$\Psi_\alpha[p] = \frac{1}{\alpha - 1}[1 - p_{\max}^\alpha (1 + \sum_{\mu \neq \mu_{\max}}^{M} \frac{p_\mu^\alpha}{p_{\max}^\alpha})] \,.$$

After taking $\alpha \to \infty$ and applying the l'Hôpital's rule for ∞/∞-form, we obtain

$$\Psi_{\alpha\to\infty}[p] = -\lim_{\alpha\to\infty} \log(p_{\max}) p_{\max}^{\alpha} ,$$

where $(p_\mu/p_{\max})^\alpha \to 0$ as $\alpha \to \infty$. If $p_{\max} < 1$, then $\Psi_{\alpha\to\infty} = 0$. If $p_{\max} = 1$, then $\Psi_{\alpha\to\infty} \to 0$ due to the logarithm. Hence, $\Psi_{\alpha\to\infty}[p]$ for all p.

This completes the proof. □

Remark D.6. If $\alpha > 0$, then the Tsallis entropy is strictly concave since its second derivative respect to p is negative on the simplex interior. This includes Shannon entropy ($\alpha \to 1$), which is also strictly concave.

Remark D.7. Although the Tsallis entropy is not defined at $\alpha = 1$ from Definition D.8, we use the continuous extension $\Psi_1[p] \coloneqq \lim_{\alpha\to 1} \Psi_\alpha[p]$. It is equal to the Shannon entropy according to Proposition D.2. Therefore, throughout the paper, the $\alpha = 1$ refers to this Shannon/softmax limit.

Proposition D.3 (Fenchel Conjugate of Entropy (Blondel et al., 2020))**.** The Fenchel conjugate of the negative Tsallis entropy is

$$\Omega_\alpha^\star[z] \coloneqq (-\Psi_\alpha)^\star[z] = \sup_{p\in\Delta^M} \{\sum_{\mu=1}^{M} z_\mu p_\mu + \Psi_\alpha[p]\} , \tag{D.5}$$

where $\Delta^M \coloneqq \{p_\mu \in \mathbb{R}_+^M \,|\, \sum_{\mu=1}^M p_\mu = 1\}$ is a $(M-1)$-dimensional unit simplex. Then

- In the Shannon case, $\alpha \to 1$, the negative entropy Ω_α is a Legendre function.
- For $\alpha > 1$, Ω_α satisfies involutivity, but neither Legendre condition nor bijectiveness.
- For $1 > \alpha > 0$, Ω_α is a Legendre function.

Proof. First, we look at the Shannon case.

- Due to (D.3) and Remark D.6, the Shannon entropy is closed and strictly concave. By Fact 1, it satisfies involutivity, $S^{\star\star}[p] = S[p]$. Furthermore, $\|\nabla_p S[p]\|_{p\to 0} = \infty$. Thus, it is essentially smooth. The negative Shannon entropy Ω_α is thereby a Legendre function and satisfies bijectiveness.
- For $\alpha > 1$, the Tsallis entropy is closed and strictly concave on the simplex. Therefore, $\Omega_\alpha = -\Psi_\alpha$ satisfies involutivity condition $\Omega^{\star\star}_{\alpha>1}[p] = \Omega_{\alpha>1}[p]$. However, $\|\nabla_p \Omega_{\alpha>1}[p]\|_{p\to 0}$ does not diverge, so it is not essentially smooth. As a result, it is not a Legendre function nor satisfies bijectiveness.
- For $1 > \alpha > 0$, the Tsallis entropy is strictly concave. Hence, Ω_α is strictly convex. Furthermore, $\partial\Omega_\alpha/\partial p_\mu \to -\infty$ near the boundary, so it satisfies the essential smoothness. Therefore, it is a Legendre function satisfying the invlolutivity and bijectiveness properties.

This completes the proof. □

Remark D.8. We are able to express unit simplex condition in two different ways: $1 = \sum_{\mu=1}^M p_\mu = \int_{\mathcal{X}} p_\mu(x)\, \mathrm{d}\mu(x)$. That is, we transform the x configuration space, $\mathcal{X}$, onto a simplex, Δ^M.

Definition D.9 (α-Deformed Exponential Family and Logarithm (Tsallis, 1988))**.** Let $\alpha > 0$. We define the α-deformed exponential function as

$$\exp_\alpha(x) \coloneqq \begin{cases} \exp(x) & , \text{ if } \alpha = 1 \\ [1 + (1-\alpha)x]_+^{\frac{1}{1-\alpha}} & , \text{ if } \alpha \neq 1 \end{cases} ,$$

and its inverse as

$$\log_\alpha(y) := \begin{cases} \log(y) & , \text{ if } \alpha = 1 \\ \frac{y^{1-\alpha}-1}{1-\alpha} & , \text{ if } \alpha \neq 1 \end{cases},$$

where $x \in \mathbb{R}$, $y \in \mathbb{R}_+$ for well-definedness, and $[z]_+ := \max\{z, 0\}$ for every $z \in \mathbb{R}$.

This α-deformed family has several properties, and some of which differ from the ordinary ones.

Lemma D.1 (Properties of α-Deformed Functions (Tsallis, 1988)). α-deformed functions satisfy

- $\lim_{\alpha\to 1} \exp_\alpha(x) = \exp(x)$ and $\lim_{\alpha\to 1} \log_\alpha(x) = \log(x)$.
- $\exp_\alpha(x)/\log_\alpha(y)$ are monotonic and convex/concave.
- α-deformed exponential/logarithm does not have multiplicity/additivity unless $\alpha = 1$.
- $\frac{d}{dx}\exp_\alpha(x) = [\exp_\alpha(x)]^\alpha$ in the domain $[1 + (1-\alpha)x] > 0$ and $\frac{d}{dy}\log_\alpha(y) = y^{-\alpha}$.

Proof. We repeat the proof from (Tsallis, 1988) for completeness. We prove those properties one-by-one as follows.

- Let $u = 1 - \alpha$. By Definition D.9,

$$\lim_{\alpha\to 1} \exp_\alpha(x) = \lim_{\alpha\to 1} [1 + (1-\alpha)x]_+^{\frac{1}{1-\alpha}} = \lim_{u\to 0} [1 + ux]^{1/u} = \exp(x) .$$

  For the logarithm, we apply the l'Hôpital's rule for $0/0$-form.

$$\lim_{\alpha\to 1} \log_\alpha(y) = \lim_{\alpha\to 1} \frac{y^{1-\alpha} - 1}{1-\alpha} = -\lim_{\alpha\to 1} \log(y) y^{1-\alpha} = \log(y) .$$

- If $[1 + (1-\alpha)x] > 0$, by differentiating $\exp_\alpha(x)$ with respect to x, we have

$$\frac{d}{dx}\exp_\alpha(x) = [1 + (1-\alpha)x]^{\frac{\alpha}{1-\alpha}} > 0 , \quad \frac{d^2}{dx^2}\exp_\alpha(x) = \alpha[1 + (1-\alpha)x]^{\frac{2\alpha-1}{1-\alpha}} > 0 .$$

  Hence, it is strictly increasing (monotonic) and strictly convex. For $\log_\alpha$,

$$\frac{d}{dy}\log_\alpha(y) = y^{-\alpha} > 0 , \quad \frac{d^2}{dy^2}\log_\alpha(y) = -\alpha y^{-\alpha-1} < 0 .$$

  Hence, it is strictly increasing (monotonic) and strictly concave.

- Consider $x_1, x_2 \in \mathbb{R}$.

$$\begin{aligned} \exp_\alpha(x_1)\exp_\alpha(x_2) &= [1 + (1-\alpha)(x_1 + x_2) + (1-\alpha)^2 x_1 x_2]_+^{\frac{1}{1-\alpha}} \\ &= \exp_\alpha(x_1 + x_2) \cdot [1 + \frac{(1-\alpha)^2 x_1 x_2}{1 + (1-\alpha)(x_1 + x_2)}]_+^{\frac{1}{1-\alpha}} . \end{aligned} \tag{D.6}$$

  If $\alpha \to 1$, then $[\cdots]$ in (D.6) is 1. Consider $y_1, y_2 \in \mathbb{R}_+$.

$$\begin{aligned} \log_\alpha(y_1 y_2) &= \frac{y_1^{1-\alpha} y_2^{1-\alpha} - 1}{1-\alpha} = \frac{(y_1^{1-\alpha} - 1) + (y_2^{1-\alpha} - 1) + (y_1^{1-\alpha} - 1)(y_2^{1-\alpha} - 1)}{1-\alpha} \\ &= \log_\alpha(y_1) + \log_\alpha(y_2) + (1-\alpha)\log_\alpha(y_1)\log(y_2) . \end{aligned} \tag{D.7}$$

  If $\alpha \to 1$, the last term in (D.7) is zero.

- For $[1+(1-\alpha)x] > 0$, we have

$$\frac{d}{dx}\exp_\alpha(x) = [1+(1-\alpha)x]^{\frac{1}{1-\alpha}-1} = [\exp_\alpha(x)]^\alpha .$$

For $\log_\alpha(y)$, we obtain the result by differentiation easily. For $\alpha \to 1$, they hold trivially.

This completes the proof. □

Remark D.9. Since $p_\mu^\alpha \log_\alpha(p_\mu) = (1-\alpha)^{-1}(p_\mu - p_\mu^\alpha)$, using $\sum_{\mu=1}^M p_\mu = 1$, we obtain

$$\Psi_\alpha[p] = -\sum_{\mu=1}^M p_\mu^\alpha \log_\alpha(p_\mu). \tag{D.8}$$

D.4 Proof of Theorem 2.1

Theorem D.1 (Theorem 2.1 Restated: Tsallis Entropy and Deformed Free Energy)**.** If we replace the exponential functions and logarithm functions by α-deformed functions, then we have the deformed log-partition function

$$-W_\alpha[J] = \beta \mathbb{E}^{(\alpha)}_{x\sim p_J}[\mathcal{H}[x]] - \frac{\Psi_\alpha[p_J]}{Z_\alpha^{\alpha-1}[J]\int_{\mathcal{X}} p_J^\alpha(x)\,\mathrm{d}\mu(x)} - \sum_{i=1}^d J_i \mathbb{E}^{(\alpha)}_{x\sim p_J}[x_i],$$

where $\mathbb{E}^{(\alpha)}_{x\sim p_J}[\cdot] \coloneqq \int_{\mathcal{X}} p_J^\alpha(x)[\cdot]\mathrm{d}\mu(x) / \int_{\mathcal{X}} p_J^\alpha(x)\mathrm{d}\mu(x)$ is the escort expectation value, and

$$\Gamma_\alpha = \beta \mathbb{E}^{(\alpha)}_{x\sim p_{J^\star}}[\mathcal{H}[x]] - \frac{\Psi_\alpha[p_{J^\star}]}{Z_\alpha^{\alpha-1}[J^\star]\int_{\mathcal{X}} p_{J^\star}^\alpha(x)\,\mathrm{d}\mu(x)}, \qquad \text{with } \varphi = \mathbb{E}^{(\alpha)}_{x\sim p_{J^\star}}[x], \tag{D.9}$$

where $J^\star = \arg\max_J\{\sum_{i=1}^d J_i\varphi_i - W_\alpha[J]\}$. We assume $\mathcal{X}$ is finite-volume to avoid divergence.

Proof. By Definition D.9, we express $W_\alpha[J] = \log_\alpha Z_\alpha[J]$ and

$$p_J(x) = \frac{\exp_\alpha\{-\beta\mathcal{H}[x] + \sum_{i=1}^d J_i x_i\}}{Z_\alpha[J]}\ ,\ Z_\alpha[J] = \int_{\mathcal{X}} \exp_\alpha\{-\beta\mathcal{H}[x] + \sum_{i=1}^d J_i x_i\}\,\mathrm{d}\mu(x)\ .$$

Notice that since $\mathcal{X}$ is finite-volume, $Z_\alpha[J]$ is finite for appropriate J and the following computation is reliable. Then we have

$$\begin{aligned}
\log_\alpha p_J(x) &= \log_\alpha\Big(\frac{\exp_\alpha\{-\beta\mathcal{H}[x] + \sum_{i=1}^d J_i x_i\}}{Z_\alpha[J]}\Big) \\
&= \frac{Z_\alpha^{\alpha-1}[J]\exp_\alpha^{1-\alpha}\{-\beta\mathcal{H}[x] + \sum_{i=1}^d J_i x_i\} - 1}{1-\alpha} && \text{(By Definition D.9)} \\
&= \frac{Z_\alpha^{\alpha-1}[J]\{1+(1-\alpha)(-\beta\mathcal{H}[x] + \sum_{i=1}^d J_i x_i)\} - 1}{1-\alpha} && \text{(By Definition D.9)} \\
&= Z_\alpha^{\alpha-1}[J](-\beta\mathcal{H}[x] + \sum_{i=1}^d J_i x_i) + \frac{Z_\alpha^{\alpha-1}[J] - 1}{1-\alpha} \\
&= Z_\alpha^{\alpha-1}[J](-\beta\mathcal{H}[x] + \sum_{i=1}^d J_i x_i) - Z_\alpha^{\alpha-1}[J]\cdot\Big(\frac{Z_\alpha^{1-\alpha}[J] - 1}{1-\alpha}\Big)
\end{aligned}$$

$$= Z_\alpha^{\alpha-1}[J](-\beta\mathcal{H}[x] + \sum_{i=1}^{d} J_i x_i - W_\alpha[J]).$$

(By Definition D.9 and $W_\alpha[J] = \log_\alpha Z_\alpha[J]$)

By (D.8), we express the Tsaliis entropy as

$$\begin{aligned}\Psi_\alpha[p_J] &= -\int_{\mathcal{X}} p_J^\alpha(x) \log_\alpha(p_J(x))\, \mathrm{d}\mu(x) \\ &= -\int_{\mathcal{X}} p_J^\alpha(x) Z_\alpha^{\alpha-1}[J](-\beta\mathcal{H}[x] + \sum_{i=1}^{d} J_i x_i - W_\alpha[J]). \end{aligned} \tag{D.10}$$

This gives

$$\begin{aligned}-W_\alpha[J] &= \frac{-\Psi_\alpha[p_J] + \beta\int_{\mathcal{X}} p_J^\alpha(x) Z_\alpha^{\alpha-1}[J]\mathcal{H}[x]\,\mathrm{d}\mu(x) - \int_{\mathcal{X}}\sum_{i=1}^{d} p_J^\alpha(x) Z_\alpha^{\alpha-1}[J] J_i x_i}{Z_\alpha^{\alpha-1}[J]\int_{\mathcal{X}} p_J^\alpha(x)\,\mathrm{d}\mu(x)} \\ &= \beta\,\frac{\int_{\mathcal{X}} p_J^\alpha(x)\mathcal{H}[x]\,\mathrm{d}\mu(x)}{\int_{\mathcal{X}} p_J^\alpha(x)\,\mathrm{d}\mu(x)} - \frac{\Psi_\alpha[p_J]}{Z_\alpha^{\alpha-1}[J]\cdot\int_{\mathcal{X}} p_J^\alpha(x)\,\mathrm{d}\mu(x)} - \sum_{i=1}^{d} J_i\,\frac{\int_{\mathcal{X}} p_J^\alpha(x) x_i\,\mathrm{d}\mu(x)}{\int_{\mathcal{X}} p_J^\alpha(x)\,\mathrm{d}\mu(x)}.\end{aligned}$$

The first and last term are the escort expectation values $\mathbb{E}^{(\alpha)}_{x\sim p_J}[\cdot]$. Hence, it is the form of our first equation. After applying the Fenchel conjugation, we have

$$\begin{aligned}\Gamma_\alpha &= \sup_J\{\sum_{i=1}^{d} J_i\varphi_i - W_\alpha[J]\} \\ &= \sup_J\{\sum_{i=1}^{d} J_i(\varphi_i - \mathbb{E}^{(\alpha)}_{x\sim p_J}[x_i]) + \beta\mathbb{E}^{(\alpha)}_{x\sim p_J}[\mathcal{H}[x]] - \frac{\Psi_\alpha[p_J]}{Z_\alpha^{\alpha-1}[J]\cdot\int_{\mathcal{X}} p_J^\alpha(x)\,\mathrm{d}\mu(x)}\} \\ &= \beta\mathbb{E}^{(\alpha)}_{x\sim p_{J^\star}}[\mathcal{H}[x]] - \frac{\Psi_\alpha[p_{J^\star}]}{Z_\alpha^{\alpha-1}[J^\star]\cdot\int_{\mathcal{X}} p_{J^\star}^\alpha(x)\,\mathrm{d}\mu(x)},\end{aligned}$$

where $\varphi = \mathbb{E}^{(\alpha)}_{x\sim p_{J^\star}}[x]$. This proves the theorem. □

Remark D.10. The finite-volume assumption is only to ensure that the α-deformed partition function is well-defined. Specifically, if we consider whole $\mathbb{R}^d$ with $\alpha = 2$ and $a > 0$, then

$$\int_{\mathbb{R}^d} \exp_2\{-a\|x\|_2^2\}\,\mathrm{d}\mu(x) = \int_{\mathbb{R}^d} \frac{\mathrm{d}\mu(x)}{1 + a\|x\|_2^2} = S_{d-1}\int_0^\infty \frac{r^{d-1}}{1+ar^2}\,\mathrm{d}r,$$

where S_{d-1} is the surface area of a unity sphere S^{d-1}. Hence, only if $d < 2$ is the integral finite. This leads to the consideration of a finite-volume domain $\mathcal{X}$. Passing to a finite-volume domain $\mathcal{X} \subset \mathbb{R}^d$ introduces a boundary. It removes the divergence but does not guarantee strict convexity of $W_\alpha[J]$ nor the essential smoothness at the boundary of $\mathrm{dom}W_\alpha$. Hence, according to Definition D.7, the transformation here should be understood as a Fencehl conjugate on $\mathrm{dom}W_\alpha$, not as a Legendre transform necessarily.

D.5 Proof of Proposition 3.1

Proposition D.4 (Proposition 3.1 Restated: Query-Probability Relation)**.** A query state vector relates to a probability function via

$$x_i = \frac{1}{\sqrt{d}}\sum_{\mu=1}^{M} \xi_i^\mu p_\mu + \frac{1}{\beta} J_i.$$

Proof. In the deterministic limit, we have

$$Z_\alpha[J] = \int_{\mathcal{X}} \exp_\alpha\{-\beta\widetilde{\mathcal{H}}_{\text{MHM}}[x,p,\xi] + \sum_{i=1}^{d} J_i x_i\}\, \mathrm{d}\mu(x) = \exp_\alpha\{-\beta\widetilde{\mathcal{H}}_{\text{MHM}}[x,p,\xi] + \sum_{i=1}^{d} J_i x_i\},$$

since we identify x with the mean. The query state vector that maximizes $Z_\alpha[J]$ is the deterministic x itself. Hence, we have

$$\begin{aligned} 0 &= \frac{\partial}{\partial x_i}\{\beta\widetilde{\mathcal{H}}_{\text{MHM}}[x,p,\xi] - \sum_{j=1}^{d} J_j x_j\} \\ &= \beta x_i - \frac{\beta}{\sqrt{d}} \sum_{\mu=1}^{M} \xi_i^\mu p_\mu - J_i. \end{aligned}$$

This proves the proposition. □

D.6 Proof of Theorem 3.1

Proposition D.5 (Theorem 3.1 Restated: Free Energy of MHM)**.** The joint free energy of the MHM is

$$\widetilde{\Gamma}_\alpha[x,p,\xi] = \frac{\beta}{2}\sum_{i=1}^{d} x_i^2 - \beta \sum_{\mu=1}^{M} z_\mu p_\mu.$$

Hence,

$$\Gamma_\alpha[x,\xi] = \frac{\beta}{2}\sum_{i=1}^{d} x_i^2 - \beta \max_{\mu\in[1,M]} z_\mu.$$

Proof. The form of the partition function is

$$Z_\alpha[J,\xi] = \exp_\alpha\{W_\alpha[J,\xi]\} = \int_{\mathcal{X}}\int_{\Delta^M} \exp_\alpha\{-\frac{\beta}{2}\sum_{i=1}^{d} x_i^2 + \beta\sum_{\mu=1}^{M} z_\mu p_\mu + \sum_{i=1}^{d} J_i x_i\}\, \mathrm{d}\mu(p)\mathrm{d}\mu(x)\,.$$

According to the form of the Tsallis entropy (D.8) and (D.10), we have

$$\begin{aligned} \Psi_\alpha[\rho_J] =& -\int_{\mathcal{X}}\int_{\Delta^M} \rho_J^\alpha(x,p)\log_\alpha[\rho_J(x,p)]\, \mathrm{d}\mu(p)\mathrm{d}\mu(x) \\ =& \frac{\beta}{2} Z_\alpha^{\alpha-1}[J,\xi] \sum_{i=1}^{d}\int_{\mathcal{X}}\int_{\Delta^M} \rho_J^\alpha(x,p) x_i^2\, \mathrm{d}\mu(p)\mathrm{d}\mu(x) \\ &- \beta Z_\alpha^{\alpha-1}[J,\xi] \sum_{\mu=1}^{M}\int_{\mathcal{X}}\int_{\Delta^M} \rho_J^\alpha(x,p) p_\mu z_\mu(x)\, \mathrm{d}\mu(p)\mathrm{d}\mu(x) \\ &- Z_\alpha^{\alpha-1}[J,\xi] \sum_{i=1}^{d} J_i \int_{\mathcal{X}}\int_{\Delta^M} \rho_J^\alpha(x,p) x_i\, \mathrm{d}\mu(p)\mathrm{d}\mu(x) \\ &+ Z_\alpha^{\alpha-1}[J,\xi] W_\alpha[J,\xi] \int_{\mathcal{X}}\int_{\Delta^M} \rho_J^\alpha(x,p)\, \mathrm{d}\mu(p)\mathrm{d}\mu(x). \end{aligned}$$

It yields

$$-W_\alpha[J,\xi] = \frac{\beta}{2}\sum_{i=1}^{d}\mathbb{E}^{(\alpha)}_{(x,p)\sim\rho_J}[x_i^2] - \beta\sum_{\mu=1}^{M}\mathbb{E}^{(\alpha)}_{(x,p)\sim\rho_J}[z_\mu p_\mu] - \frac{\Psi_\alpha[\rho_J]}{Z_\alpha^{\alpha-1}[J,\xi]\int_{\mathcal{X}}\int_{\Delta^M}\rho_J^\alpha(x,p)\,\mathrm{d}\mu(p)\mathrm{d}\mu(x)} - \sum_{i=1}^{d}J_i\mathbb{E}^{(\alpha)}_{(x,p)\sim\rho_J}[x_i]. \tag{D.11}$$

Since x and p are fixed, so the probability density function concentrates to an indicator, i.e.,

$$\rho_J(x,p) \to \mathbb{1}_{\{x,\widetilde{\mathbb{E}}^{(\alpha)}_{x\sim p_J}[x]\}}\mathbb{1}_{\{p,\widetilde{\mathbb{E}}^{(\alpha)}_{p\sim p_{\Delta,J}}[p]\}},$$

where

$$\widetilde{\mathbb{E}}^{(\alpha)}_{x\sim p_J}[x] \coloneqq \mathbb{E}^{(\alpha)}_{(x,p)\sim\rho_J}[x] = \frac{\int_{\mathcal{X}}C(x)xp_J^\alpha(x)\,\mathrm{d}\mu(x)}{\int_{\mathcal{X}}C(x)p_J^\alpha(x)\,\mathrm{d}\mu(x)}, \qquad \text{with } C(x) \coloneqq \int_{\Delta^M}\rho_J^\alpha(p|x)\,\mathrm{d}\mu(p), \tag{D.12}$$

and

$$\widetilde{\mathbb{E}}^{(\alpha)}_{p\sim p_{\Delta,J}}[p_\mu] \coloneqq \mathbb{E}^{(\alpha)}_{(x,p)\sim\rho_J}[p_\mu] = \frac{\int_{\Delta^M}C(p)p_\mu p_{\Delta,J}^\alpha(p)\,\mathrm{d}\mu(p)}{\int_{\Delta^M}C(p)p_{\Delta,J}^\alpha(p)\,\mathrm{d}\mu(p)}, \qquad \text{with } C(p) \coloneqq \int_{\mathcal{X}}\rho_J^\alpha(x|p)\,\mathrm{d}\mu(x). \tag{D.13}$$

Then

$$Z_\alpha[J,\xi] \to \exp_\alpha\{-\beta\widetilde{\mathcal{H}}_{\text{MHM}}[\widetilde{\mathbb{E}}^{(\alpha)}_{x\sim p_J}[x],\widetilde{\mathbb{E}}^{(\alpha)}_{p\sim p_{\Delta,J}}[p],\xi] + \sum_{i=1}^{d}J_i\widetilde{\mathbb{E}}^{(\alpha)}_{x\sim p_J}[x_i]\}.$$

Hence, by the Fenchel conjugate, we obtain

$$\begin{aligned}
\widetilde{\Gamma}_\alpha &= \sup_{J\in\mathrm{dom}W_\alpha}\{\sum_{i=1}^{d}J_i\varphi_i - W_\alpha[J,\xi]\} \\
&= \sup_{J\in\mathrm{dom}W_\alpha}\Big\{\sum_{i=1}^{d}J_i(\varphi_i - \widetilde{\mathbb{E}}^{(\alpha)}_{x\sim p_J}[x_i]) + \frac{\beta}{2}\sum_{i=1}^{d}(\widetilde{\mathbb{E}}^{(\alpha)}_{x\sim p_J}[x_i^2]) - \beta\sum_{\mu=1}^{M}\widetilde{\mathbb{E}}^{(\alpha)}_{x\sim p_J}[z_\mu]\widetilde{\mathbb{E}}^{(\alpha)}_{p\sim p_{\Delta,J}}[p_\mu] \\
&\quad - \frac{\Psi_\alpha[\mathbb{1}]}{\exp_\alpha^{\alpha-1}\{-\beta\widetilde{\mathcal{H}}_{\text{MHM}}[\widetilde{\mathbb{E}}^{(\alpha)}_{x\sim p_J}[x],\widetilde{\mathbb{E}}^{(\alpha)}_{p\sim p_{\Delta,J}}[p],\xi] + \sum_{i=1}^{d}J_i\widetilde{\mathbb{E}}^{(\alpha)}_{x\sim p_J}[x_i]\}}\Big\} \\
&= \frac{\beta}{2}\sum_{i=1}^{d}(\widetilde{\mathbb{E}}^{(\alpha)}_{x\sim p_{J^\star}}[x_i^2]) - \beta\sum_{\mu=1}^{M}\widetilde{\mathbb{E}}^{(\alpha)}_{x\sim p_{J^\star}}[z_\mu]\widetilde{\mathbb{E}}^{(\alpha)}_{p\sim p_{\Delta,J^\star}}[p_\mu], \qquad (\text{By } \Psi_\alpha[\mathbb{1}]=0)
\end{aligned}$$

where $J^\star$ is the maximizing source. Since we identify $\widetilde{\mathbb{E}}^{(\alpha)}_{x\sim p_{J^\star}}[x]$ with x, $\widetilde{\mathbb{E}}^{(\alpha)}_{x\sim p_{J^\star}}[x^2]$ with x^2, and $\widetilde{\mathbb{E}}^{(\alpha)}_{p\sim p_{\Delta,J^\star}}[p]$ with p, we have

$$\widetilde{\Gamma}_\alpha[x,p,\xi] = \frac{\beta}{2}\sum_{i=1}^{d}x_i^2 - \beta\sum_{\mu=1}^{M}z_\mu p_\mu.$$

Notice that

$$\inf_{p\in\Delta^M}\{-\sum_{\mu=1}^{M}z_\mu p_\mu\} = -\max_{\mu\in[1,M]}z_\mu.$$

After substituting into the joint free energy to eliminate the dependence of p, we prove the second equation in this proposition. □

D.7 Proof of Proposition 3.3

Proposition D.6 (Proposition 3.3 Revisited: Emergent Attention Mechanism as EOMs in MHMs)**.** Attention emerges as follows.

- If $\alpha \to \infty$, it corresponds to a hardmax function.

$$p_\mu = \text{hardmax}(z_\mu) = \begin{cases} 1\,, & \text{for maximal } z_\mu \\ 0\,, & \text{otherwise} \end{cases}.$$

- If $\alpha = 2$, it corresponds to a sparsemax function (Martins & Astudillo, 2016).

$$p_\mu = \text{Sparsemax}_\beta(z_\mu) = \frac{\beta}{2}[z_\mu - \theta]_+, \qquad \text{with } \frac{\beta}{2}\sum_{\mu=1}^{M}[z_\mu - \theta]_+ = 1.$$

- If $1 < \alpha < 2$, it corresponds to an entmax family (Peters et al., 2019).
- If $\alpha \to 1$, it corresponds to a softmax function (Blondel et al., 2020).

$$p_\mu = \text{Softmax}_\beta(z_\mu) = \frac{\exp\{\beta z_\mu\}}{\sum_{\nu=1}^{M}\exp\{\beta z_\nu\}}.$$

Proof. We include the proof for completeness, adapting the standard argument from (Martins & Astudillo, 2016; Peters et al., 2019; Blondel et al., 2020) to our notation. The training trajectory follows $p_\mu = \arg\max_{p\in\Delta^M}\{\beta\sum_{\mu=1}^{M} z_\mu p_\mu + \Psi_\alpha[p]\}$. We derive the physical meanings of this optimization problem via the Lagrangian function:

$$\frac{\partial}{\partial p_\mu}\{\sum_{\nu=1}^{M} z_\nu p_\nu + \frac{1}{\beta}\cdot\frac{1}{\alpha-1}(1-\sum_{\nu=1}^{M} p_\nu^\alpha) + \theta(1-\sum_{\nu=1}^{M} p_\nu) + \sum_{\nu=1}^{M}\lambda_\nu p_\nu\} = 0,$$

where we use the form of Tsallis entropy, (D.2), and the Lagrangian multipliers, λ and $\lambda_\nu \geq 0$, to include the simplex constraints. This gives the complementary slackness, $\lambda_\mu p_\mu = 0$, and the stationary condition,

$$z_\mu - \frac{\alpha}{\beta(\alpha-1)}p_\mu^{\alpha-1} - \theta + \lambda_\mu = 0 \implies \begin{cases} p_\mu = \{\beta(1-\frac{1}{\alpha})[z_\mu-\theta]_+\}^{\frac{1}{\alpha-1}} & \text{, if } \lambda_\mu = 0 \\ p_\mu = 0 & \text{, if } \lambda_\mu > 0 \end{cases},$$

where $[u]_+ \coloneqq \max(u, 0)$. For $\alpha \to \infty$ and $\alpha = 2$, we obtain the forms of p_μ by inserting α directly. For $\alpha = 1$, we use the Shannon form, (D.3). Hence, for non-zero p_μ, we have

$$z_\mu + \frac{1}{\beta}\frac{\partial}{\partial p_\mu}(-\sum_{\nu=1}^{M} p_\nu \log p_\nu) - \theta = 0.$$

This gives

$$p_\mu = \exp\{\beta(z_\mu - \theta) - 1\}.$$

The normalization condition gives $1 = \sum_{\mu=1}^{M} p_\mu = \exp\{-\beta\theta\}\sum_{\mu=1}^{M}\exp\{\beta z_\mu - 1\}$. This yields $\theta = \beta^{-1}\log\sum_{\mu=1}^{M}\exp\{\beta z_\mu - 1\}$. After plugging this into p_μ, we obtain the softmax function. For $2 > \alpha > 1$, $(\alpha-1)^{-1} > 1$. Hence, p_μ is the entmax family interpolating between softmax and sparsemax. □

D.8 Proof of Proposition 4.1

Proposition D.7 (Proposition 4.1 Restated: Temperature Coexistence Point for Sparse Model)**.** Let a, b be two distinct branches. If $\alpha = 2$, k_a and k_b are the supports of branch a and branch b, respectively, then the coexistence point satisfies

$$\sum_{\mu,\nu\in a}[z^\star_\mu-\theta]_+[z^\star_\nu-\theta]_+K_{\mu\nu}(\beta_{\text{coexist}};\beta_0)=\sum_{\mu,\nu\in b}[z^\star_\mu-\theta]_+[z^\star_\nu-\theta]_+K_{\mu\nu}(\beta_{\text{coexist}};\beta_0),$$

where $K_{\mu\nu}(\beta) \coloneqq \mathbb{1}_{\{\mu,\nu\}} - \frac{\beta}{2}G_{\mu\nu}$. β_{coexist} does not have a closed expression. However, if the distribution is uniform on the simplex and the system satisfies the well-separation condition, i.e.,

$$G_{\mu\nu}=\begin{cases}1 & ,\ \mu=\nu\\ \epsilon\ll 1 & ,\ \mu\neq\nu\end{cases},$$

then for $k_a = 1$ and $k_b = 2$, we have

$$\beta_{\text{coexist}}=\frac{2}{1-\epsilon}.$$

Proof. Consider a system with the minimizer, $p^\star_\mu$. Using that at a stable point, $J_i = 0$ for all $i = 1, ..., d$, we have

$$z^\star_\mu=\frac{1}{\sqrt{d}}\sum_{i=1}^d\xi^\mu_i x^\star_i=\frac{1}{\sqrt{d}}\sum_{i=1}^d\xi^\mu_i(\frac{1}{\sqrt{d}}\sum_{\nu=1}^M\xi^\nu_i p^\star_\nu+\frac{1}{\beta}J_i)=\sum_{\nu=1}^M G_{\mu\nu}p^\star_\nu.$$

Moreover,

$$\sum_{i=1}^d(x^\star_i)^2=\frac{1}{d}\sum_{i=1}^d\sum_{\mu,\nu=1}^M\xi^\mu_i\xi^\nu_i p^\star_\mu p^\star_\nu=\sum_{\mu,\nu=1}^M G_{\mu\nu}p^\star_\mu p^\star_\nu.$$

Suppose the support is k, we denote the indices of the non-zero components as $\widetilde{\mu} = 1, ..., k$. Then we have

$$\begin{aligned}\widetilde{\Gamma}^{(\text{mod})}_\alpha[x^\star,p^\star,\xi]&=\frac{\beta}{2}\sum_{i=1}^d(x^\star_i)^2-\beta\sum_{\widetilde{\mu}=1}^k z^\star_{\widetilde{\mu}}p^\star_{\widetilde{\mu}}-(1-\sum_{\widetilde{\mu}=1}^k(p^\star_{\widetilde{\mu}})^2)\\&=-1-\sum_{\widetilde{\mu},\widetilde{\nu}=1}^k p^\star_{\widetilde{\mu}}(\frac{\beta}{2}G_{\widetilde{\mu}\widetilde{\nu}}-\mathbb{1}_{\{\widetilde{\mu},\widetilde{\nu}\}})p^\star_{\widetilde{\nu}}\end{aligned}$$

By Lemma G.1, we express the minimizer as $p^\star_{\widetilde{\mu}} = \frac{\beta}{2}(z^\star_{\widetilde{\mu}} - \theta)$ for $\alpha = 2$. By matching the free energy on branch a and branch b, we prove the first equation. For the well-separation and uniform distribution conditions, we have

$$\widetilde{\Gamma}^{(\text{mod})}_\alpha[x^\star_a,p^\star,\xi]=-1-\frac{\beta_{\text{coexist}}}{2}\times 1+1=-\frac{\beta_{\text{coexist}}}{2}.$$

For branch b,

$$\widetilde{\Gamma}^{(\text{mod})}_\alpha[x^\star_b,p^\star,\xi]=-1-\begin{pmatrix}\frac{1}{2} & \frac{1}{2}\end{pmatrix}\begin{pmatrix}\frac{\beta_{\text{coexist}}}{2}-1 & \frac{\beta_{\text{coexist}}\epsilon}{2}\\ \frac{\beta_{\text{coexist}}\epsilon}{2} & \frac{\beta_{\text{coexist}}}{2}-1\end{pmatrix}\begin{pmatrix}\frac{1}{2}\\ \frac{1}{2}\end{pmatrix}=-\frac{1}{2}-\frac{\beta_{\text{coexist}}}{4}(1+\epsilon).$$

After equating these two energies, we prove the proposition. □

D.9 Proof of Proposition 4.2

Proposition D.8 (Proposition 4.2 Restated: Temperature Spinodal Point for Sparse Model)**.** Consider $\alpha = 2$ and a branch, a, with support, k. Then the spinodal critical value is

$$\beta_{\text{jump}} = \frac{2}{\lambda_{\max}(PGP)},$$

where $P_{\mu\nu} \coloneqq \mathbb{1}_{\{\mu,\nu\}} - k^{-1}$. If the distribution is uniform on the simplex and the system satisfies the well-separation condition, (4.3), then

$$\beta_{\text{jump}} = \frac{2}{1-\epsilon}.$$

Proof. Since for $\alpha = 2$, the minimizer satisfies $p^\star = \frac{\beta}{2}(z^\star_\mu - \theta)$, where $\mu = 1, ..., k$, we have

$$\frac{\partial p^\star_\mu}{\partial z^\star_\nu} = \frac{\beta}{2}\mathbb{1}_{\{\mu,\nu\}} - \frac{\beta}{2}\frac{\partial}{\partial z^\star_\nu}\left(\frac{\sum_{\rho=1}^k z^\star_\rho - 1}{k}\right) = \frac{\beta}{2}\left(\mathbb{1}_{\{\mu,\nu\}} - \frac{1}{k}\right) = \frac{\beta}{2}P_{\mu\nu}.$$

Then the gradient of the free energy at the local stable point, $x^\star$, is

$$\begin{aligned}\frac{\partial}{\partial x_i}\Gamma^{(\text{mod})}_\alpha[x^\star, \xi] &= \beta x^\star_i - \frac{\partial}{\partial x_i}\sup_{p\in\Delta^M}\{\beta\sum_{\mu=1}^k z_\mu p_\mu + 1 - \sum_{\mu=1}^k p_\mu^2\}|_{x=x^\star} \\ &= \beta x^\star_i - \frac{\beta}{\sqrt{d}}\sum_{\mu=1}^k \xi^\mu_i p^\star_\mu.\end{aligned}$$

The Hessian is

$$\begin{aligned}\frac{\partial^2}{\partial x_i \partial x_j}\Gamma^{(\text{mod})}_\alpha[x^\star, \xi] &= \beta\mathbb{1}_{\{i,j\}} - \frac{\beta}{\sqrt{d}}\sum_{\mu,\nu=1}^k \xi^\mu_i \frac{\partial p^\star_\mu}{\partial z^\star_\nu}\frac{\partial z^\star_\nu}{\partial x_j}|_{x=x^\star} \\ &= \beta\mathbb{1}_{\{i,j\}} - \frac{\beta^2}{2d}\sum_{\mu,\nu=1}^k \xi^\mu_i \xi^\nu_j P_{\mu\nu}. \qquad \text{(D.14)}\end{aligned}$$

The smallest Hessian eigenvalue vanishes when $\frac{\beta_{\text{jump}}}{2d}\lambda_{\max}(\sum_{\mu,\nu=1}^k \xi^\mu_i P_{\mu\nu}\xi^\nu_j) = 1$. Notice that $P_{\mu\nu}$ is a projector, i.e.,

$$\sum_{\nu=1}^k P_{\mu\nu}P_{\nu\rho} = P_{\mu\rho}.$$

Suppose u is the eigenvector such that $\sum_{j=1}^d \sum_{\mu,\nu=1}^k \xi^\mu_i P_{\mu\nu}\xi^\nu_j u_j = \lambda u_i$. We construct

$$C_{\mu\nu} \coloneqq \sum_{\rho,\sigma=1}^k P_{\mu\rho}\left(\sum_{i=1}^d \xi^\rho_i \xi^\sigma_i\right)P_{\sigma\nu} = d\sum_{\rho,\sigma=1}^k P_{\mu\rho}G_{\rho\sigma}P_{\sigma\nu}.$$

In fact, its eigenvector is $v_\mu = \sum_{\nu=1}^k \sum_{i=1}^d P_{\mu\nu}\xi^\nu_i u_i$. In this way,

$$(Cv)_\mu = d\sum_{\nu,\rho,\sigma=1}^k \sum_{\kappa=1}^k \sum_{i=1}^d P_{\mu\rho}G_{\rho\sigma}P_{\sigma\nu}P_{\nu\kappa}\xi^\kappa_i u_i$$

$$
\begin{aligned}
&= \sum_{\rho,\sigma=1}^{k} \sum_{\kappa=1}^{k} \sum_{i,j=1}^{d} P_{\mu\rho} \xi_j^\rho \xi_j^\sigma P_{\sigma\kappa} \xi_i^\kappa u_i && \text{(By } P \text{ being a projector)} \\
&= \sum_{\rho=1}^{k} \sum_{j=1}^{d} P_{\mu\rho} \xi_j^\rho \lambda u_j && \text{(By } \textstyle\sum_{j=1}^{d} \sum_{\mu,\nu=1}^{k} \xi_i^\mu P_{\mu\nu} \xi_j^\nu u_j = \lambda u_i \text{)} \\
&= \lambda v_\mu.
\end{aligned}
$$

Hence, its eigenvalue is the same as which of $\sum_{\mu,\nu=1}^{k} \xi_i^\mu P_{\mu\nu} \xi_j^\nu$. That is,

$$
\lambda_{\max}\Big(\sum_{\mu,\nu=1}^{k} \xi_i^\mu P_{\mu\nu} \xi_j^\nu \Big) = d \cdot \lambda_{\max}\Big(\sum_{\rho,\sigma=1}^{k} P_{\mu\rho} G_{\rho\sigma} P_{\sigma\nu} \Big).
$$

Finally, we obtain

$$
\beta_{\text{jump}} = \frac{2d}{\lambda_{\max}(\sum_{\mu,\nu=1}^{k} \xi_i^\mu P_{\mu\nu} \xi_j^\nu)} = \frac{2}{\lambda_{\max}(\sum_{\rho,\sigma=1}^{k} P_{\mu\rho} G_{\rho\sigma} P_{\sigma\nu})}.
$$

If the system satisfies the well-separation condition, $G_{\mu\nu} = (1-\epsilon)\mathbb{1}_{\{\mu,\nu\}} + \epsilon$, with $\epsilon \ll 1$. Then

$$
\begin{aligned}
\sum_{\rho,\sigma=1}^{k} P_{\mu\rho} G_{\rho\sigma} P_{\sigma\nu} &= (1-\epsilon) \sum_{\rho,\sigma=1}^{k} P_{\mu\rho} \mathbb{1}_{\{\rho,\sigma\}} P_{\sigma\nu} + \epsilon \sum_{\rho,\sigma=1}^{k} P_{\mu\rho} P_{\sigma\nu} \\
&= (1-\epsilon) P_{\mu\nu} && \text{(By } P \text{ being a projector and } P(1\ 1\ \cdots\ 1)^\top = 0\text{)}
\end{aligned}
$$

Let the vector space be $I \oplus I^\perp$, where $I \coloneqq \text{span}\{(1, ..., 1)^\top\}$. Since the eigenvalue is $(1-\epsilon)$ on $I^\perp$ and zero on I, the largest eigenvalue is $(1-\epsilon)$. After substituting into the expression of β_{jump}, we prove the proposition. □

E FLUCTUATED VERSION AND THERMAL ATTENTION

In this section, we revisit the modern Hopfield models but with the fluctuated variables, x, p, and z. First, we study how the temperature influences the fluctuation property. Then we move on to discuss the modern Hopfield models.

E.1 PHYSICS OF TEMPERATURE

In this section, we discuss the relation of φ and x_J first. Then we introduce the role that temperature plays in the perturbative analysis.

Proposition E.1 (Saddle Points as Stationary Points in the Zero-Temperature Limit (Landau & Lifshitz, 1980))**.** Consider the zero-temperature limit, $\beta \to \infty$. If $(-\beta\mathcal{H}[x] + \sum_{i=1}^{d} J_i x_i)$ has a unique maximizer, then the saddle points approach to the stationary points, $\varphi_i = (x_J)_i$.

Proof. We repeat the proof from (Landau & Lifshitz, 1980; Dembo & Zeitouni, 2010) for completeness. We denote the maximizer as x_J. Since it is unique, for any $x_i \neq (x_J)_i$, we have

$$\frac{p_J(x)}{p_J(x_J)} = \exp\{-\beta(\mathcal{H}[x] - \mathcal{H}[x_J]) + \sum_{i=1}^{d} J_i[x_i - (x_J)_i]\} \coloneqq e^{-\beta\Delta + C} ,$$

where $\Delta > 0$, and C is a finite value. Then in the zero-temperature limit, $p_J(x) \to p_J(x_J)$. Thus, $\varphi_i = \mathbb{E}_{p_J}[x_i] \to (x_J)_i$. □

Proposition E.2 (Fluctuation and Loop Expansion (Thouless et al., 1977))**.** The free energy captures noise and fluctuations. The fluctuation term behaves like a data-dependent regularizer. Specifically, if we expand x around φ, then

$$\Gamma[\varphi] = \beta\mathcal{H}[\varphi] - \frac{d}{2}\log(2\pi) + \frac{1}{2}\log\det\left(\beta\nabla_x^2\mathcal{H}[\varphi]\right) + O(\beta^{-1}) \tag{E.1}$$

$$\coloneqq \Gamma_{\text{tree}}[\varphi] + \Gamma_{\text{1-loop}}[\varphi] + \Gamma_{\text{2-loop}}[\varphi] + \cdots . \tag{E.2}$$

This free energy admits higher-order (higher-loop) corrections arising from fluctuations, which we denote as $O(\beta^{-1})$. Specifically, the two-loop is of order β^{-1}, three-loop is of order β^{-2}, and so forth. Conventionally, we ignore the constant term, $d\log(2\pi)/2$, since it is independent of φ and does not affect the dynamics.

Proof. We include the proof for completeness, adapting the standard argument from (Thouless et al., 1977; Zinn-Justin, 2002). By Definition 2.3, we write

$$\begin{aligned}\exp\{-\Gamma[\varphi]\} =&\ \exp\{W[J] - \sum_{i=1}^{d} J_i\varphi_i\} \\ =&\ \int_{\mathcal{X}} \exp\{-\beta\mathcal{H}[x] + \sum_{i=1}^{d} J_i(x_i - \varphi_i)\}\, \mathrm{d}\mu(x) .\end{aligned}$$

Then we expand the right hand side around the stationary point, φ, perturbatively:

$$\begin{aligned}\exp\{-\Gamma[\varphi]\} = \int_{\mathcal{X}} \exp\{&-\beta\mathcal{H}[\varphi] - \beta\sum_{i=1}^{d}\frac{\partial\mathcal{H}}{\partial x_i}|_{x=\varphi}(x_i - \varphi_i) \\ &- \frac{\beta}{2}\sum_{i,j=1}^{d}\frac{\partial^2\mathcal{H}}{\partial x_i\partial x_j}|_{x=\varphi}(x_i - \varphi_i)(x_j - \varphi_j) \\ &- \frac{\beta}{3!}\sum_{i,j,k=1}^{d}\frac{\partial^3\mathcal{H}}{\partial x_i\partial x_j\partial x_k}|_{x=\varphi}(x_i - \varphi_i)(x_j - \varphi_j)(x_k - \varphi_k)\end{aligned}$$

$$+ O(\|x-\varphi\|_2^4)\} \exp\{\sum_{i=1}^d J_i(x_i-\varphi_i)\}\, \mathrm{d}\mu(\mathbf{x})$$

$$= \exp\{-\beta\mathcal{H}[\varphi]\} \int_{\mathcal{X}} \exp\{-\frac{\beta}{2}\sum_{i,j=1}^d (x_i-\varphi_i)(\nabla_x^2\mathcal{H}[\varphi])_{ij}(x_j-\varphi_j)\} \cdot (1-\frac{\beta}{3!}\times$$

$$\sum_{k,\ell,m=1}^d \frac{\partial^3\mathcal{H}}{\partial x_k \partial x_\ell \partial x_m}|_{x=\varphi}(x_k-\varphi_k)(x_\ell-\varphi_\ell)(x_m-\varphi_m) + O(\beta\|x-\varphi\|_2^4))\, \mathrm{d}\mu(x)$$

(By Definition 2.2)

where $\|x'-\varphi\|_2^2 \coloneqq \sum_{i=1}^d (x_i'-\varphi_i)^2$. Firstly,

$$\int_{\mathcal{X}} \exp\{-\frac{\beta}{2}\sum_{i,j=1}^d (x_i-\varphi_i)(\nabla_x^2\mathcal{H}[\varphi])_{ij}(x_j-\varphi_j)\}\, \mathrm{d}\mu(x) = (2\pi)^{d/2}(\det\big(\beta\nabla_x^2\mathcal{H}[\varphi]\big))^{-1/2}\,.$$

(By Gaussian Formula)

Since the odd function vanishes after the integration, the first leading correction is of $O(\|x-\varphi\|_2^4)$. We focus on this integration:

$$\int_{\mathcal{X}} \exp\{-\frac{\beta}{2}\sum_{i,j=1}^d (x_i-\varphi_i)(\nabla_x^2\mathcal{H}[\varphi])_{ij}(x_j-\varphi_j)\} O(\beta\|x-\varphi\|_2^4)\, \mathrm{d}\mu(x)$$

$$= \int_{\mathcal{Y}} \exp\{-\frac{1}{2}\sum_{i,j=1}^d (\nabla_x^2\mathcal{H}[\varphi])_{i,j}y_iy_j\} O(\beta\|y\|_2^4)\cdot(\beta^{-1/2})^4\, \mathrm{d}\mu(y) \qquad (\text{By } x\to y=\sqrt{\beta}(x-\varphi))$$

$$= O(\beta^{-1})\,.$$

Putting everything together, we obtain

$$\exp\{-\Gamma[\varphi]\} = \exp\{-\beta\mathcal{H}[\varphi] - \frac{1}{2}\log\det\big(\beta\nabla_x^2\mathcal{H}[\varphi]\big) + \frac{d}{2}\log(2\pi) + O(\beta^{-1})\}\,.$$

This proves the proposition. □

Proposition E.3 (Low-Temperature Limit and Laplace Approximation (Zinn-Justin, 2002; Dembo & Zeitouni, 2010))**.** If $(-\beta\mathcal{H}[x] + \sum_{i=1}^d J_i x_i)$ has unique maximizer x_J with positive-definite Hessian, then for continuous cases,

$$\frac{1}{\beta}\log\int_{\mathcal{X}} \exp\{-\beta\mathcal{H}[x] + \sum_{i=1}^d J_i x_i\}\, \mathrm{d}\mu(x) = \sup_{x\in\mathcal{X}}\{-\mathcal{H}[x] + \frac{1}{\beta}\sum_{i=1}^d J_i x_i\} + O(\frac{\log\beta}{\beta})\,.$$

For discrete cases,

$$\frac{1}{\beta}\log\sum_{x\in\mathcal{X}} \exp\{-\beta\mathcal{H}[x] + \sum_{i=1} J_i x_i\} = \sup_{x\in\mathcal{X}}\{-\mathcal{H}[x] + \frac{1}{\beta}\sum_{i=1}^d J_i x_i\} + O(\frac{\log|\mathcal{X}|}{\beta})\,. \tag{E.3}$$

Proof. We repeat the proof from (Zinn-Justin, 2002; Dembo & Zeitouni, 2010) for completeness. We start from the continuous case. Since the maximizer is unique, we Taylor expand up to the first leading order as

$$-\beta\mathcal{H}[x] + \sum_{i=1}^d J_i x_i = -\beta\mathcal{H}[\varphi] + \sum_{i=1}^d J_i\varphi_i - \frac{1}{2}\beta[x_i-\varphi_i](\nabla_x^2\mathcal{H}[\varphi])_{ij}[x_j-\varphi_j]\,.$$

(By Definition 2.2)

This yields

$$\begin{aligned}
&\frac{1}{\beta}\log\int_{\mathcal{X}=\mathbb{R}^d}\exp\{-\beta\mathcal{H}[x]+\sum_{i=1}^{d}J_i x_i\}\,\mathrm{d}\mu(x)\\
&=\frac{1}{\beta}\log\left(\exp\{-\beta\mathcal{H}[\varphi]+\sum_{i=1}^{d}J_i\varphi_i\}\right)+\frac{1}{\beta}\log\left((\frac{2\pi}{\beta})^{d/2}(\det\left(\beta\nabla_x^2\mathcal{H}[\varphi]\right))^{-1/2}\right)\\
&=-\mathcal{H}[\varphi]+\frac{1}{\beta}\sum_{i=1}^{d}J_i\varphi_i+\mathcal{O}(\frac{\log\beta}{\beta})\,.
\end{aligned}$$

This proves the first equation. For discrete cases,

$$\begin{aligned}
\max_{x\in\mathcal{X}}\left(-\beta\mathcal{H}[x]+\sum_{i=1}^{d}J_i x_i\right)&\le\log\sum_{x\in\mathcal{X}}\exp\{-\beta\mathcal{H}[x]+\sum_{i=1}^{d}J_i x_i\}\\
&\le\max_{x\in\mathcal{X}}\left(-\beta\mathcal{H}[x]+\sum_{i=1}^{d}J_i x_i\right)+\log|\mathcal{X}|\,.
\end{aligned}$$

After plugging this into (E.3), we complete the proof. □

Remark E.1. In the zero-temperature limit, Proposition E.3 shows that $W[J]/\beta$ is a Fenchel conjugate of the Hamiltonian by Definition D.6.

Remark E.2. From the proof in Proposition E.2 and Proposition E.3, we notice that the perturbation only works in the low-temperature regime. When the temperature increases, the strength of fluctuations increases and the higher-loop corrections starts to grow. There exists a level where the perturbative method breaks down.

E.2 Fluctuated Variables in Modern Hopfield Model

In the following, we start to consider the fluctuated version of MHMs.

Definition E.1 (Statistics-MHM (Fluctuated) Mapping)**.** The mapping between the variables in statistical mechanics and which in the fluctuated version of MHMs is:

- **Microscopic State** $x\in\mathbb{R}^d$: It represents continuous query. It is also the dynamical variable whose energy landscape defines retrieval basins.
- **External Source** $J\in\mathbb{R}^d$: It represents an external bias of the query.
- **Classical Solution** $\varphi\in\mathbb{R}^d$: It represents macroscopic state. Retrieval corresponds to φ_i aligning with some pattern ξ_i^μ.
- **Probability Function** $p\in\Delta^M$: It represents the retrieval distribution over memories.
- **Tsallis Entropy** Ψ_α: It represents a regularizer controlling how sparse the memory weights p are.
- **Conjugate Field** $z\in\mathbb{R}^M$: It represents the query-key similarity between current state and each memory, i.e., $z_\mu\coloneqq\sum_{i=1}^{d}\xi_i^\mu x_i/\sqrt{d}$.

Notice that we distinguish the microscopic state, x, from the macroscopic one, φ, and so are p and z in this case.

Proposition E.4 (Expectation Values and Macroscopic States)**.** Let $\mathcal{X}$ be a finite-volume domain in $\mathbb{R}^d$ and $J\in\mathrm{dom}W_\alpha$. The macroscopic states are determined by the mean of microscopic states. Specifically,

$$\mathbb{E}^{(\alpha)}_{x\sim\rho_J(\cdot|p)}[x_i|p]=\frac{1}{\sqrt{d}}\sum_{\mu=1}^{M}\xi_i^\mu p_\mu+\frac{1}{\beta}J_i+B_i^{\mathcal{X}}(p,J)\coloneqq m_i(p,J)+B_i^{\mathcal{X}}(p,J)\,,\tag{E.4}$$

$$\mathbb{E}^{(\alpha)}_{(x,p)\sim\rho_J}[x_i] = \varphi_i = \frac{1}{\sqrt{d}}\sum_{\mu=1}^{M}\xi_i^\mu \mathbb{E}^{(\alpha)}_{(x,p)\sim\rho_J}[p_\mu] + \frac{1}{\beta}J_i + \mathbb{E}^{(\alpha)}_{(x,p)\sim\rho_J}[B_i^{\mathcal{X}}(p,J)]\,, \tag{E.5}$$

where $\rho_J(x,p)$ is the probability density function in $\mathcal{X}\times\Delta^M$, $p_J(x)$ is which in $\mathcal{X}$, and $p_{\Delta,J}(p)$ is which in Δ^M. Furthermore, there is a boundary correction

$$B_i^{\mathcal{X}}(p,J) \coloneqq \frac{\int_{\mathcal{Y}} y_i \exp_\alpha^\alpha\{-\frac{\beta}{2}(\|y\|_2^2 - \|m\|_2^2)\}\,\mathrm{d}\mu(y)}{\int_{\mathcal{Y}} \exp_\alpha^\alpha\{-\frac{\beta}{2}(\|y\|_2^2 - \|m\|_2^2)\}\,\mathrm{d}\mu(y)}, \qquad \text{with } \mathcal{Y} = \mathcal{X} - m.$$

Proof. We consider x and p follow the distribution $\rho_J(x,p)$ with the joint energy function, (3.1). Then

$$\begin{aligned}
\mathbb{E}^{(\alpha)}_{x\sim\rho_J(\cdot|p)}[x_i|p] &= \frac{\int_{\mathcal{X}} x_i \rho_J^\alpha(x|p)\,\mathrm{d}\mu(x)}{\int_{\mathcal{X}} \rho_J^\alpha(x|p)\,\mathrm{d}\mu(x)} \\
&= \frac{\int_{\mathcal{X}} x_i \exp_\alpha^\alpha\{-\frac{\beta}{2}\sum_{j=1}^d x_j^2 + \frac{\beta}{\sqrt{d}}\sum_{\mu=1}^M\sum_{j=1}^d \xi_j^\mu x_j p_\mu + \sum_{j=1}^d J_j x_j\}\,\mathrm{d}\mu(x)}{\int_{\mathcal{X}} \exp_\alpha^\alpha\{-\frac{\beta}{2}\sum_{j=1}^d x_j^2 + \frac{\beta}{\sqrt{d}}\sum_{\mu=1}^M\sum_{j=1}^d \xi_j^\mu x_j p_\mu + \sum_{j=1}^d J_j x_j\}\,\mathrm{d}\mu(x)} \\
&= \frac{\int_{\mathcal{Y}} (y_i + m_i) \exp_\alpha^\alpha\{\frac{-\beta}{2}\sum_{j=1}^d y_j^2 + \frac{\beta}{2}\sum_{j=1}^d m_j^2\}\,\mathrm{d}\mu(y)}{\int_{\mathcal{Y}} \exp_\alpha^\alpha\{\frac{-\beta}{2}\sum_{j=1}^d y_j^2 + \frac{\beta}{2}\sum_{j=1}^d m_j^2\}\,\mathrm{d}\mu(y)} \\
&= m_i(p,J) + B_i^{\mathcal{X}}(p,J).
\end{aligned}$$

This proves the first equation. In this way,

$$\begin{aligned}
\mathbb{E}^{(\alpha)}_{(x,p)\sim\rho_J}[x_i] &= \frac{\int_{\Delta^M}\int_{\mathcal{X}} x_i \rho_J^\alpha(x,p)\,\mathrm{d}\mu(x)\mathrm{d}\mu(p)}{\int_{\Delta^M}\int_{\mathcal{X}} \rho_J^\alpha(x,p)\,\mathrm{d}\mu(x)\mathrm{d}\mu(p)} \\
&= \frac{\int_{\Delta^M}(\int_{\mathcal{X}} x_i \rho_J^\alpha(x|p)\,\mathrm{d}\mu(x)) p_{\Delta,J}^\alpha(p)\,\mathrm{d}\mu(p)}{\int_{\Delta^M}(\int_{\mathcal{X}} \rho_J^\alpha(x|p)\,\mathrm{d}\mu(x)) p_{\Delta,J}^\alpha(p)\,\mathrm{d}\mu(p)} \qquad (\text{By } p_{\Delta,J}(p)\rho_J(x|p) \coloneqq \rho_J(x,p)) \\
&= \frac{\int_{\Delta^M}(\int_{\mathcal{X}} \rho_J^\alpha(x'|p)\,\mathrm{d}\mu(x')\mathbb{E}^{(\alpha)}_{x\sim\rho_J(\cdot|p)}[x_i|p]) p_{\Delta,J}^\alpha(p)\,\mathrm{d}\mu(p)}{\int_{\Delta^M}(\int_{\mathcal{X}} \rho_J^\alpha(x|p)\,\mathrm{d}\mu(x)) p_{\Delta,J}^\alpha(p)\,\mathrm{d}\mu(p)} \\
&= \frac{1}{\sqrt{d}}\sum_{\mu=1}^M \xi_i^\mu \widetilde{\mathbb{E}}^{(\alpha)}_{p\sim p_{\Delta,J}}[p_\mu] + \frac{1}{\beta}J_i + \widetilde{\mathbb{E}}^{(\alpha)}_{p\sim p_{\Delta,J}}[B_i^{\mathcal{X}}(p,J)],
\end{aligned}$$

where

$$C(p) \coloneqq \int_{\mathcal{X}} \rho_J^\alpha(x|p)\,\mathrm{d}\mu(x), \qquad \text{and } \widetilde{\mathbb{E}}^{(\alpha)}_{p\sim p_{\Delta,J}}[p_\mu] \coloneqq \mathbb{E}^{(\alpha)}_{(x,p)\sim\rho_J}[p_\mu] = \frac{\int_{\Delta^M} C(p) p_\mu p_{\Delta,J}^\alpha(p)\,\mathrm{d}\mu(p)}{\int_{\Delta^M} C(p) p_{\Delta,J}^\alpha(p)\,\mathrm{d}\mu(p)}.$$

This completes the proof. □

Remark E.3. The correction $B_i^{\mathcal{X}}$ vanishes on translation-invariant domains (e.g., $\mathbb{R}^d$ if the integral is normalizable). Furthermore, it is negligible in the large-volume or low-temperature regime when the dominant integrand is localized far from the boundary of $\mathcal{X}$.

E.3 Connection between Free Energy and Tsallis Entropy in Fluctuated Case

In Section 2.2, we relate the log-partition function to the Tsallis entropy via (2.4). Here, we express the log-partition function using the variables in fluctuated version of MHMs, Definition E.1. To make $Z_\alpha[J]$ finite, we assume that the integration region of x is finite volume, i.e., $\mathcal{X} \subset \mathbb{R}^d$ such that $|\mathcal{X}| < +\infty$.

Proposition E.5. Let $\rho_J(x,p)$ and $p_J(x)$ be the probability density function in $\mathcal{X} \times \Delta^M$ and $\mathcal{X}$, respectively. Inspired by Theorem 2.1, we have

$$-W_\alpha[J,\xi] = \frac{\beta}{2}\sum_{i=1}^{d}\mathbb{E}^{(\alpha)}_{(x,p)\sim\rho_J}[x_i^2] - \beta\sum_{\mu=1}^{M}\mathbb{E}^{(\alpha)}_{(x,p)\sim\rho_J}[z_\mu p_\mu] - \frac{\Psi_\alpha[\rho_J]}{Z_\alpha^{\alpha-1}[J,\xi]\int_{\mathcal{X}}\int_{\Delta^M}\rho_J^\alpha(x,p)\,\mathrm{d}\mu(p)\mathrm{d}\mu(x)} - \sum_{i=1}^{d}J_i\mathbb{E}^{(\alpha)}_{(x,p)\sim\rho_J}[x_i]. \tag{E.6}$$

Then the free energy is

$$\widetilde{\Gamma}_\alpha = \frac{\beta}{2}\sum_{i=1}^{d}\mathbb{E}^{(\alpha)}_{(x,p)\sim\rho_{J^\star}}[x_i^2] - \beta\sum_{\mu=1}^{M}\mathbb{E}^{(\alpha)}_{(x,p)\sim\rho_{J^\star}}[z_\mu p_\mu] - \frac{\Psi_\alpha[\rho_{J^\star}]}{Z_\alpha^{\alpha-1}[J^\star,\xi]\int_{\mathcal{X}}\int_{\Delta^M}\rho_{J^\star}^\alpha(x,p)\,\mathrm{d}\mu(p)\mathrm{d}\mu(x)}, \tag{E.7}$$

where $J^\star$ is the maximizing source, i.e., $J^\star = \arg\max_J\{\sum_{i=1}^d J_i\varphi_i - W_\alpha[J,\xi]\}$, and $\varphi = \mathbb{E}^{(\alpha)}_{(x,p)\sim\rho_J}[x]$.

Proof. Following the proof of Theorem 3.1, we have (D.11). This is exactly the first equation we would like to prove. For the free energy, we consider the Fenchel conjugation Definition D.6 of the log-partition function.

$$\begin{aligned}
\widetilde{\Gamma}_\alpha &= \sup_{J\in\mathrm{dom}W_\alpha}\{\sum_{i=1}^{d}J_i\varphi_i - W_\alpha[J,\xi]\} \\
&= \sup_{J\in\mathrm{dom}W_\alpha}\Big\{\sum_{i=1}^{d}J_i(\varphi_i - \mathbb{E}^{(\alpha)}_{(x,p)\sim\rho_J}[x_i]) + \frac{\beta}{2}\sum_{i=1}^{d}\mathbb{E}^{(\alpha)}_{(x,p)\sim\rho_J}[x_i^2] - \beta\sum_{\mu=1}^{M}\mathbb{E}^{(\alpha)}_{(x,p)\sim\rho_J}[z_\mu p_\mu] \\
&\quad - \frac{\Psi_\alpha[\rho_J]}{Z_\alpha^{\alpha-1}[J,\xi]\int_{\mathcal{X}}\int_{\Delta^M}\rho_J^\alpha(x,p)\,\mathrm{d}\mu(p)\mathrm{d}\mu(x)}\Big\} \\
&= \frac{\beta}{2}\sum_{i=1}^{d}\mathbb{E}^{(\alpha)}_{(x,p)\sim\rho_{J^\star}}[x_i^2] - \beta\sum_{\mu=1}^{M}\mathbb{E}^{(\alpha)}_{(x,p)\sim\rho_{J^\star}}[z_\mu p_\mu] \\
&\quad - \frac{\Psi_\alpha[\rho_{J^\star}]}{Z_\alpha^{\alpha-1}[J^\star,\xi]\int_{\mathcal{X}}\int_{\Delta^M}\rho_{J^\star}^\alpha(x,p)\,\mathrm{d}\mu(p)\mathrm{d}\mu(x)},
\end{aligned}$$

where $\varphi = \mathbb{E}^{(\alpha)}_{(x,p)\sim\rho_J}[x]$. This completes the proof. □

Proposition E.6 (MHMs in the Zero-Temperature Limit)**.** In the zero-temperature limit, $\beta\to\infty$, minimizing the free energy becomes minimizing the following quantity:

$$\mathcal{F}[\varphi,\xi] = \frac{1}{2}\sum_{i=1}^{d}\varphi_i^2 - \max_\mu \bar{z}_\mu\ , \tag{E.8}$$

where $\bar{z}_\mu \coloneqq \sum_{i=1}^d \xi_i^\mu\varphi_i/\sqrt{d}$. Notably, this is the free energy in the fixed case, (3.4), up to a scaling , β.

Proof. In the zero-temperature limit, the Gibbs probability density function concentrates to an indicator, i.e.,

$$\rho_J(x,p) \to \mathbb{1}_{\{x,\mathbb{E}^{(\alpha)}_{(x,p)\sim\rho_J}[x]\}}\mathbb{1}_{\{p,\mathbb{E}^{(\alpha)}_{(x,p)\sim\rho_J}[p]\}}\ ,$$

as $\beta \to \infty$. Hence,

$$\mathbb{E}^{(\alpha)}_{(x,p)\sim\rho_J}[x^2] \to (\mathbb{E}^{(\alpha)}_{(x,p)\sim\rho_J}[x])^2 = (\widetilde{\mathbb{E}}^{(\alpha)}_{x\sim p_J}[x])^2 = \varphi^2,$$

and

$$\mathbb{E}_{(x,p)\sim\rho_J}[z_\mu p_\mu] \to \frac{1}{\sqrt{d}}\sum_{i=1}^{d}\xi_i^\mu \mathbb{E}^{(\alpha)}_{(x,p)\sim\rho_J}[x_i]\mathbb{E}^{(\alpha)}_{(x,p)\sim\rho_J}[p_\mu] = \frac{1}{\sqrt{d}}\sum_{i=1}^{d}\xi_i^\mu \varphi_i \widetilde{\mathbb{E}}^{(\alpha)}_{p\sim p_{\Delta,J}}[p_\mu] \coloneqq \bar{z}_\mu \bar{p}_\mu,$$

where $\widetilde{\mathbb{E}}^{(\alpha)}_{x\sim p_J}[x]$ and $\widetilde{\mathbb{E}}^{(\alpha)}_{p\sim p_{\Delta,J}}[p]$ are defined in (D.12) and (D.13), respectively. The free energy (E.7) becomes

$$\Gamma_\alpha \to \frac{\beta}{2}\sum_{i=1}^{d}\varphi_i^2 - \beta\sum_{\mu=1}^{M}\bar{z}_\mu\bar{p}_\mu - \Psi_\alpha[\mathbb{1}]\,.$$

Since $\Psi_\alpha[\mathbb{1}] = 0$, we have

$$\inf_{\bar{p}\in\Delta^M}\{-\sum_{\mu=1}^{M}\bar{z}_\mu\bar{p}_\mu\} = -\max_\mu \bar{z}_\mu\,.$$

Substituting it into the free energy to eliminate the dependence of $\bar{p}_\mu$, we prove this proposition. □

Remark E.4. Proposition E.6 is consistent with Proposition E.2 in the zero-temperature limit. Specifically, $\Gamma_\alpha[\varphi] \to (\beta\mathcal{H}[\varphi] + \text{constant})$ as $\beta \to \infty$. Following the proof of Theorem 3.1, we conclude the same result.

E.4 Retrieval Dynamics in Fluctuated Case and Thermal Attention

In Section 2, we introduce the equation of motions according to Proposition 2.1 and investigate the attention mechanism from the trajectories in associative memory systems. In this section, we consider the fluctuated version, which gives an alternative thermal attention mechanism. Let $\widetilde{W}_\alpha \coloneqq \log_\alpha \int_{\Delta^M} \exp_\alpha\{-\beta\widetilde{\mathcal{H}}_{\text{MHM}} + \sum_{i=1}^{d} J_i x_i\}\,\mathrm{d}\mu(p)$. Notice that in the fixed setting, it reduces to the simpler form, (3.7). Then,

Proposition E.7 (EOMs of Free Energies in MHMs)**.** The EOM of the log-partition function in the x-sector corresponds to the Hopfield retrieval

$$\nabla_J W_\alpha[J,\xi] = \int_{\mathcal{X}}\int_{\Delta^M} x_i \rho_J^\alpha(x,p)\,\mathrm{d}\mu(p)\mathrm{d}\mu(x)\,. \tag{E.9}$$

The EOM of the log-partition function in the p-sector corresponds to the unnormalized escort moment (Naudts, 2004)

$$\nabla_z \widetilde{W}_\alpha[z,J,\xi] = \int_{\Delta^M} \beta p \rho_J^\alpha(p|x)\,\mathrm{d}\mu(p) \coloneqq \beta p^{(\text{esc})}\,. \tag{E.10}$$

Proof. By Definition 2.3, $W_\alpha[J,\xi] = \log_\alpha Z_\alpha[J,\xi]$. After taking the derivative with respect to J, we obtain

$$\begin{aligned}
\frac{\partial}{\partial J_i}W_\alpha[J,\xi] &= \frac{\partial}{\partial J_i}\log_\alpha Z_\alpha[J,\xi] \\
&= Z_\alpha^{-\alpha}[J,\xi]\frac{\partial Z_\alpha[J,\xi]}{\partial J_i} && \text{(By Lemma D.1)} \\
&= Z_\alpha^{-\alpha}[J,\xi]\frac{\partial}{\partial J_i}\int_{\mathcal{X}}\int_{\Delta^M}\exp_\alpha\{-\beta\widetilde{\mathcal{H}}_{\text{MHM}}[x,p,\xi] + \sum_{j=1}^{d}J_j x_j\}\,\mathrm{d}\mu(p)\mathrm{d}\mu(x)
\end{aligned}$$

$$
\begin{aligned}
&= Z_\alpha^{-\alpha}[J,\xi] \int_{\mathcal{X}} \int_{\Delta^M} \exp_\alpha^\alpha\{-\beta\widetilde{\mathcal{H}}_{\text{MHM}}[x,p,\xi] + \sum_{j=1}^d J_j x_j\} \qquad \text{(By Lemma D.1)} \\
&\quad \times \frac{\partial}{\partial J_i}(-\beta\widetilde{\mathcal{H}}_{\text{MHM}}[x,p,\xi] + \sum_{k=1}^d J_k x_k)\, \mathrm{d}\mu(p)\mathrm{d}\mu(x) \\
&= \int_{\mathcal{X}} \int_{\Delta^M} x_i \frac{\exp_\alpha^\alpha\{-\beta\widetilde{\mathcal{H}}_{\text{MHM}}[x,p,\xi] + \sum_{j=1}^d J_j x_j\}}{Z_\alpha^\alpha[J,\xi]}\, \mathrm{d}\mu(p)\mathrm{d}\mu(x) \\
&= \int_{\mathcal{X}} \int_{\Delta^M} x_i \rho_J^\alpha(x,p)\, \mathrm{d}\mu(p)\mathrm{d}\mu(x).
\end{aligned}
$$

For the second equation, we substitute the definition of $\widetilde{W}_\alpha$ directly:

$$
\begin{aligned}
\frac{\partial}{\partial z_\mu}\widetilde{W}_\alpha[z,J,\xi] &= \frac{\partial}{\partial z_\mu} \log_\alpha \int_{\Delta^M} \exp_\alpha\{-\frac{\beta}{2}\sum_{i=1}^d x_i^2 + \beta\sum_{\mu=1}^M z_\mu p_\mu + \sum_{i=1}^d J_i x_i\}\, \mathrm{d}\mu(p) \\
&= \Big(\int_{\Delta^M} \exp_\alpha\{-\frac{\beta}{2}\sum_{i=1}^d x_i^2 + \beta\sum_{\mu=1}^M z_\mu p_\mu + \sum_{i=1}^d J_i x_i\}\, \mathrm{d}\mu(p)\Big)^{-\alpha} \\
&\quad \times \int_{\Delta^M} \beta p_\mu \exp_\alpha^\alpha\{-\frac{\beta}{2}\sum_{i=1}^d x_i^2 + \beta\sum_{\mu=1}^M z_\mu p_\mu + \sum_{i=1}^d J_i x_i\}\, \mathrm{d}\mu(p) \\
&\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad \text{(By Lemma D.1)} \\
&= \int_{\Delta^M} \beta p_\mu \rho_J^\alpha(p|x)\, \mathrm{d}\mu(p)\,.
\end{aligned}
$$

This completes the proof. □

Remark E.5. If $\alpha \to 1$, then everything simplifies to the ordinary Gibbs case. In this way, the Hopfield retrieval becomes

$$
\begin{aligned}
\nabla_J W_{\alpha\to 1}[J,\xi] &= \int_{\mathcal{X}} \int_{\Delta^M} x \rho_J(x,p)\, \mathrm{d}\mu(p)\mathrm{d}\mu(x) \\
&= \mathbb{E}_{(x,p)\sim\rho_J}[x] \\
&= \int_{\mathcal{X}} x p_J(x)\, \mathrm{d}\mu(x) \\
&= \mathbb{E}_{x\sim p_J}[x].
\end{aligned}
$$

Furthermore, the escort moment becomes the conditional expectation value of the probability function for general β.

$$
\nabla_z \widetilde{W}_{\alpha\to 1}[z,J,\xi] = \beta\mathbb{E}_{p\sim\rho_J(\cdot|x)}[p]\,.
$$

Remark E.6. If we do the similar analysis of Proposition 3.3 for the fluctuated case, then we have the thermal attention mechanism. In particular, if $\alpha \to 1$, then the learning trajectory follows the ordinary Gibbs expectation of the conditional distribution function, i.e., $\nabla_z \widetilde{W}_{\alpha\to 1} = \beta\mathbb{E}_{p\sim\rho_J(\cdot|x)}[p]$, due to Remark E.5. If $\alpha > 1$, then the system amplifies high-density regions. This leads to a sharper attention mechanism. If $1 > \alpha > 0$, then it corresponds to a softer attention mechanism.

Remark E.7. From (E.9), we obtain a macroscopic retrieval state depending on the external field at microscopic queries, $\{x_i\}$. It is similar to a spin system in an external magnetic field. After turning on the magnetic field, the probability distribution of spins starts to deform and yields an expectation value $\mathbb{E}[S_i]$, where S_i is the notation of a spin at site i. We call it a Landau response. Hence, the Hopfield retrieval is a Landau response in the x-sector (J-space). For the attention mechanism, it is an inner Landau response in the p-sector (z-space). In this way, we view the query-to-key score,

$z_\mu = \sum_{i=1}^{d} \xi_i^\mu x_i/\sqrt{d}$, as an external field at $\{p_\mu\}$ in Δ^M space. This formalism gives a physical interpretation of the retrieval dynamics.

Lemma E.1 (Linear Function and its Minimizer on Simplex (Rockafellar, 1997))**.** Consider a function $h(p)$ that is linear in p, i.e.,

$$h(p) = \sum_{\mu=1}^{M} a_\mu p_\mu + b,$$

for some constant $a_\mu \in \mathbb{R}^M$. Then $h(p)$ cannot have a unique minimizer $p^\star$ in the interior of the simplex.

Proof. This is a standard convex analysis (Rockafellar, 1997). We repeat the proof for the completeness. Assume that a minimizer lies in the interior of Δ^M. Then $p^\star > 0$. Since $p^\star$ is a minimizer, we have

$$\frac{d}{d\varepsilon} h(p^\star + \varepsilon v)|_{\varepsilon=0} = 0,$$

for all tangent vectors $v \in T_{p^\star}\Delta^M = \{v \in \mathbb{R}^M | \sum_{\mu=1}^{M} v_\mu = 0\} \simeq \mathbb{R}^{M-1}$. Then, after expanding the linear function h, it yields

$$0 = \frac{d}{d\varepsilon}(\sum_{\mu=1}^{M} a_\mu p_\mu^\star + \varepsilon \sum_{\mu=1}^{M} a_\mu v_\mu + b)|_{\varepsilon=0} = \sum_{\mu=1}^{M} a_\mu v_\mu.$$

Since the tangent vector satisfies $\sum_{\mu=1}^{M} v_\mu = 0$, we conclude that $a_\mu = \lambda$ for all μ and some constant λ. Hence,

$$h(p) = \lambda \sum_{\mu=1}^{M} p_\mu + b = \lambda + b,$$

which is constant on Δ^M. This contradicts the assumption that the minimizer $p^\star$ is unique. This proves the lemma. □

Lemma E.2 (Normalization Factor in the Zero-Temperature Limit)**.** Consider the zero-temperature limit $\beta \to \infty$ and $\alpha > 1$. Assume that $\widetilde{\mathcal{H}}_{\text{MHM}} \geq 0$, and $\rho_J(p|x)$ has a unique maximizer $p^\star$ such that $\widetilde{\mathcal{H}}_{\text{MHM}}[x, p^\star, \xi] = 0$. If the p is continuous on the simplex Δ^M, where $M > 1$, then the normalization function has the following limit

$$\lim_{\beta\to\infty} C(x) = \lim_{\beta\to\infty} \int_{\Delta^M} \rho_J^\alpha(p|x)\, \mathrm{d}\mu(p) = \infty.$$

If $\alpha \to 1$ for general β or if p is discrete on the simplex Δ^M for general α, then

$$\lim_{\alpha\to 1} C(x) = \lim_{\beta\to\infty} C^{\text{dis}}(x) = \lim_{\beta\to\infty} \sum_{p\in\Delta^M} (\rho_J^{\text{dis}})^\alpha(p|x) = 1.$$

Proof. We first consider the continuous case. Since $\widetilde{\mathcal{H}}_{\text{MHM}}$ is linear in p, its unique minimizer $p^\star$ must be at a vertex of Δ^M according to Lemma E.1. Hence, we denote

$$p_\mu^\star = \begin{cases} 1, & \text{if } \mu = \mu^\star \\ 0, & \text{if } \mu \neq \mu^\star \end{cases}. \tag{E.11}$$

In this way, for $p \neq p^\star$, we have

$$\widetilde{\mathcal{H}}_{\text{MHM}}[x, p, \xi] = \widetilde{\mathcal{H}}_{\text{MHM}}[x, p, \xi] - \widetilde{\mathcal{H}}_{\text{MHM}}[x, p^\star, \xi] \qquad (\text{By } \widetilde{\mathcal{H}}_{\text{MHM}}[x, p^\star, \xi] = 0)$$

$$
\begin{aligned}
&= \frac{1}{2}\sum_{i=1}^{d} x_i^2 + \sum_{\mu=1}^{M} z_\mu p_\mu - (\frac{1}{2}\sum_{i=1}^{d} x_i^2 + \sum_{\mu=1}^{M} z_\mu p_\mu^\star) \\
&= \sum_{\mu=1}^{M} z_\mu (p_\mu - p_\mu^\star) \\
&= \sum_{\mu=\mu^\star} z_\mu (p_\mu - p_\mu^\star) + \sum_{\substack{\mu=1 \\ \mu\neq\mu^\star}}^{M} z_\mu (p_\mu - p_\mu^\star) \\
&= z_{\mu^\star} p_{\mu^\star} - z_{\mu^\star} p_{\mu^\star}^\star + \sum_{\substack{\mu=1 \\ \mu\neq\mu^\star}}^{M} z_\mu (p_\mu - p_\mu^\star) \\
&= z_{\mu^\star} p_{\mu^\star} - z_{\mu^\star} + \sum_{\substack{\mu=1 \\ \mu\neq\mu^\star}}^{M} z_\mu p_\mu && \text{(By (E.11))} \\
&= \sum_{\substack{\mu=1 \\ \mu\neq\mu^\star}}^{M} (z_\mu - z_{\mu^\star}) p_\mu. && \text{(By } 1 = p_{\mu^\star} + \textstyle\sum_{\mu\neq\mu^\star}^{M} p_\mu\text{)}
\end{aligned}
$$

Since $p^\star$ is a unique minimizer of $\widetilde{\mathcal{H}}_{\text{MHM}}$, $(z_\mu - z_{\mu^\star}) > 0$ for all $\mu \neq \mu^\star$. From Definition D.9, we obtain

$$
\begin{aligned}
\exp_\alpha\{-\beta\widetilde{\mathcal{H}}_{\text{MHM}}[x,p,\xi] + \sum_{i=1}^{d} J_i x_i\} &= \Big\{1 + (\alpha-1)(\beta\widetilde{\mathcal{H}}_{\text{MHM}}[x,p,\xi] - \sum_{i=1}^{d} J_i x_i)\Big\}^{1/(1-\alpha)} \\
&= \Big\{1 + (\alpha-1)\beta \sum_{\mu\neq\mu^\star}^{M} (z_\mu - z_{\mu^\star}) p_\mu - (\alpha-1)\sum_{i=1}^{d} J_i x_i\Big\}^{1/(1-\alpha)}
\end{aligned}
$$

for $\alpha > 1$. Then

$$
\begin{aligned}
C(x) &= \frac{\int_{\Delta^M} \exp_\alpha^\alpha\{-\beta\widetilde{\mathcal{H}}_{\text{MHM}}[x,p,\xi] + \sum_{i=1}^{d} J_i x_i\}\, \mathrm{d}\mu(p)}{(\int_{\Delta^M} \exp_\alpha\{-\beta\widetilde{\mathcal{H}}_{\text{MHM}}[x,p',\xi] + \sum_{i=1}^{d} J_i x_i\}\, \mathrm{d}\mu(p'))^\alpha} \\
&= \frac{\int_{\Delta^M} \Big\{[1-(\alpha-1)\sum_{i=1}^{d} J_i x_i] + (\alpha-1)\beta\sum_{\mu\neq\mu^\star}^{M}(z_\mu - z_{\mu^\star})\Big\}^{\alpha/(1-\alpha)}\, \mathrm{d}\mu(p)}{\Big(\int_{\Delta^M} \Big\{[1-(\alpha-1)\sum_{i=1}^{d} J_i x_i] + (\alpha-1)\beta\sum_{\mu\neq\mu^\star}^{M}(z_\mu - z_{\mu^\star})\Big\}^{1/(1-\alpha)}\, \mathrm{d}\mu(p')\Big)^\alpha}
\end{aligned}
$$

Since the integrands only depend on $\mu \neq \mu^\star$, we change the variable to $u_\mu = \beta q_\mu$, where $q_\mu \coloneqq p_\mu$ with $\mu \neq \mu^\star$. There are $(M-1)$ such coordinates, so $\mathrm{d}\mu(p) \sim \beta^{-(M-1)}\mathrm{d}\mu(u)$. Then, for $\alpha > 1$ and $M > 1$, we have

$$
C(x) \sim \frac{\beta^{-(M-1)}}{\left(\beta^{-(M-1)}\right)^\alpha} = \beta^{(\alpha-1)(M-1)} \to \infty
$$

as $\beta \to \infty$. For $\alpha \to 1$, it is simply $C(x) = \int_{\Delta^M} \rho_J(p|x)\, \mathrm{d}\mu(p) = 1$ for any β. For the discrete case, the conditional distribution is

$$
\rho_J^{\text{dis}}(p|x) = \frac{\exp_\alpha\{-\beta\widetilde{\mathcal{H}}_{\text{MHM}}[x,p,\xi] + \sum_{i=1}^{d} J_i x_i\}}{\sum_{p'\in\Delta^M} \exp_\alpha\{-\beta\widetilde{\mathcal{H}}_{\text{MHM}}[x,p',\xi] + \sum_{i=1}^{d} J_i x_i\}},
$$

with discrete p. Notice that as $\beta \to \infty$, $\exp_\alpha\{-\beta\widetilde{\mathcal{H}}_{\text{MHM}}[x, p, \xi]\}$ is suppressed to zero except for the maximizer $p^\star$, i.e., $\exp_\alpha\{-\beta\widetilde{\mathcal{H}}_{\text{MHM}}[x, p^\star, \xi]\} = 1$. Hence, as $\beta \to \infty$, the denominator of $\rho_J^{\text{dis}}(p|x)$ approaches to

$$\lim_{\beta\to\infty} \sum_{p'\in\Delta^M} \exp_\alpha\{-\beta\widetilde{\mathcal{H}}_{\text{MHM}}[x, p', \xi] + \sum_{i=1}^d J_i x_i\} = \exp_\alpha\{\sum_{i=1}^d J_i x_i\}.$$

Then

$$\begin{aligned}
\lim_{\beta\to\infty} C^{\text{dis}}(x) &= \lim_{\beta\to\infty} \sum_{p\in\Delta^M} (\rho_J^{\text{dis}})^\alpha(p|x) \\
&= \lim_{\beta\to\infty} \Big\{ \sum_{p=p^\star} (\rho_J^{\text{dis}})^\alpha(p|x) + \sum_{\substack{p\in\Delta^M \\ p\neq p^\star}} (\rho_J^{\text{dis}})^\alpha(p|x) \Big\} \\
&= \lim_{\beta\to\infty} \Big\{ \frac{\exp_\alpha^\alpha\{\sum_{i=1}^d J_i x_i\}}{(\exp_\alpha\{\sum_{i=1}^d J_i x_i\})^\alpha} \qquad \big(\text{By } \widetilde{\mathcal{H}}_{\text{MHM}}[x, p^\star, \xi] = 0\big) \\
&\quad + \sum_{\substack{p\in\Delta^M \\ p\neq p^\star}} \frac{\exp_\alpha^\alpha\{-\beta\widetilde{\mathcal{H}}_{\text{MHM}}[x, p, \xi] + \sum_{i=1}^d J_i x_i\}}{(\exp_\alpha\{\sum_{i=1}^d J_i x_i\})^\alpha} \Big\} \\
&= \lim_{\beta\to\infty} \Big\{ 1 + 0 \Big\} \\
&= 1.
\end{aligned}$$

This completes the proof. □

Remark E.8. In practical attention mechanisms, the retrieved memories are a finite set. The attention weights are finite-dimensional vectors. Therefore, the discrete normalization behavior in Lemma E.2 occurs in practice. In contrast, the divergence of $C(x)$ in the continuous case is mainly a theoretical issue.

Proposition E.8 (Emergence of Hard Argmax from Escort Moments)**.** Assume that $\widetilde{\mathcal{H}}_{\text{MHM}} \geq 0$ has a unique minimizer $p^\star \in \Delta^M$. In the zero-temperature limit, $\beta \to \infty$, the following escort variables become deterministic hard argmax selections:

$$\begin{cases} \alpha \to 1 & : p^{(\text{esc})} \\ \alpha > 1, \text{ Continuous} & : p^{(\text{esc})}/C(x) \\ \alpha > 1, \text{ Discrete} & : p^{(\text{esc})} \end{cases} \xrightarrow[\beta\to\infty]{} p^\star = \arg\max_{p\in\Delta^M} \{\sum_{\mu=1}^M z_\mu p_\mu\}. \tag{E.12}$$

Proof. For $\alpha \to 1$, the escort moment reduces to the ordinary Gibbs expectation,

$$p^{(\text{esc})} = \int_{\Delta^M} p\rho_J(p|x)\, \mathrm{d}\mu(p).$$

By Proposition E.3, the Gibbs distribution collapses at the unique minimizer in the zero-temperature limit,

$$\rho_J(p|x) \to \mathbb{1}_{\{p,p^\star\}},$$

as $\beta \to \infty$. Hence, $p^{(\text{esc})} \to p^\star$ as $\beta \to \infty$. Next, we consider $\alpha > 1$ with continuous p. By Lemma E.2, $\lim_{\beta\to\infty} C(x)$ diverges. However, since the escort measure still concentrates at the unique minimizer $p^\star$, we obtain

$$\lim_{\beta\to\infty} \frac{p^{(\text{esc})}}{C(x)} = \lim_{\beta\to\infty} \frac{\int_{\Delta^M} p\rho_J^\alpha(p|x)\, \mathrm{d}\mu(p)}{\int_{\Delta^M} \rho_J^\alpha(p'|x)\, \mathrm{d}\mu(p')} = p^\star.$$

Finally, for $\alpha > 1$ with discrete p, we have $\lim_{\beta\to\infty} C^{\text{dis}}(x) = 1$ according to Lemma E.2. Hence,

$$\lim_{\beta\to\infty} \frac{p^{(\text{esc})}}{C^{\text{dis}}(x)} = \lim_{\beta\to\infty} p^{(\text{esc})} = p^{\star}.$$

This proves the proposition. □

Remark E.9. Proposition E.8 shows that escort-weighted retrieval (E.10) reduces to a deterministic hard argmax selection rule in the zero-temperature limit. This statistical-mechanics framework demonstrates how sparse retrieval behavior is able to emerge from escort concentration in the general fluctuated (stochastic) setting.

F Dynamical Retrieval Stability Analysis

In this section, we investigate the dynamical retrieval stability. Then we discuss how the convergences of x and p relate to each other. To do this, we apply gradient-based optimization and perturb around the fixed point. The stability analysis verifies whether the training trajectory flows back to the fixed-point or not.

Definition F.1 (Retrieval Dynamics)**.** The retrieval update rules are

$$\frac{d}{dt}x_i(t) = -(\nabla_x \Gamma_\alpha^{(\text{mod})}[x(t),\xi])_i \quad \text{(Continuous)}\,, \tag{F.1}$$

$$x_i(t+1) = x_i(t) - \eta(\nabla_x \Gamma_\alpha^{(\text{mod})}[x(t),\xi])_i \quad \text{(Discrete)}\,, \tag{F.2}$$

where η is the learning rate.

In Section 2, we note that $\nabla_\varphi \Gamma_\alpha[\varphi]$ is not bijective. Hence, it converges to a set rather than a point in general. In our setting, we consider retrieval dynamics of the generalized model $\Gamma_\alpha^{(\text{mod})}$ in the query space (x-space), while the memory patterns remain fixed. The projection mapping $z_\mu = \sum_{i=1}^d \xi_i^\mu x_i/\sqrt{d}$ is not injective generally, so distinct $x^\star$ correspond to one specific fixed-point $z^\star$. However, due to $\frac{\beta}{2}\sum_{i=1}^d x_i^2$ term, such projection degeneracy does not imply the degeneracy of the full $\Gamma_\alpha^{(\text{mod})}[x,\xi]$.

By the proposition below, we are capable of studying the stability and refer the stability to the stability modulo the full degeneracy. This interpretation is consistent with the analysis of gradient-based optimization developed in (Lee et al., 2016), where the local Hessian structure governs the stability of gradient flows near critical points. Furthermore, we only apply the gradient-based optimization in x-space. Hence, the stable fixed-point in x-space must correspond to one specific stable fixed-point in z-space.

Proposition F.1 (State-Retrieval Stability Modulo Degeneracy)**.** Let the macroscopic state space be $\mathcal{M}_x$, $\mathcal{K}(x) \coloneqq \ker(\nabla_x^2 \Gamma_\alpha^{(\text{mod})}[x,\xi])$ be the kernel space, and $\mathcal{K}^\perp(x)$ be the orthogonal complement of $\mathcal{K}(x)$. Assume there exists a local stable point $x^\star \in \mathbb{R}^d$ such that

$$\nabla_x \Gamma_\alpha^{(\text{mod})}[x^\star,\xi] = 0\,, \tag{F.3}$$

and

$$\text{Hessian}\ \ \nabla_x^2 \Gamma_\alpha^{(\text{mod})}[x^\star,\xi]\ \ \text{is positive definite on}\ \ \mathcal{K}^\perp(x^\star). \tag{F.4}$$

Let $\mathcal{N}_{x^\star} \subset \mathcal{M}_\varphi$ be the degeneracy space through $x^\star$ such that $\Gamma_\alpha^{(\text{mod})}[x,\xi]$ is constant on it, and $T_{x^\star}\mathcal{N}_{x^\star} = \mathcal{K}(x^\star)$. Then $x^\star$ is a stable local minimizer modulo degeneracy. That is, the gradient flow in the x-space converges to $(x^\star + v)$ for some $v \in \mathcal{K}(x^\star)$. Additionally, for discrete gradient descent, we require

$$0 < \eta < \frac{2}{\lambda_{\max}(\nabla_x^2 \Gamma_\alpha^{(\text{mod})}[x^\star,\xi])|_{\mathcal{K}^\perp(x^\star)}}\,, \tag{F.5}$$

where $\lambda_{\max}(\cdot)|_{\mathcal{K}^\perp(x^\star)}$ is the maximum eigenvalue with the Hessian being on $\mathcal{K}^\perp(x^\star)$.

Proof. Let $x_i - x_i^\star \coloneqq \delta_i \coloneqq \delta_i^\parallel + \delta_i^\perp$ be infinitesimal, where $\delta^\parallel \in \mathcal{K}(x^\star)$ and $\delta^\perp \in \mathcal{K}^\perp(x^\star)$. Then for non-zero $\delta^\perp$,

$$\begin{aligned}\Gamma_\alpha^{(\text{mod})}[x,\xi] &= \Gamma_\alpha^{(\text{mod})}[x^\star,\xi] + \sum_{i=1}^d (\nabla_x \Gamma_\alpha^{(\text{mod})}[x^\star,\xi])_i \delta_i + \frac{1}{2}\sum_{i,j=1}^d (\nabla_x^2 \Gamma_\alpha^{(\text{mod})}[x^\star,\xi])_{ij}\delta_i\delta_j + O(\delta^3) \\ &= \Gamma_\alpha^{(\text{mod})}[x^\star,\xi] + \frac{1}{2}\sum_{i,j=1}^d (\nabla_x^2 \Gamma_\alpha^{(\text{mod})}[x^\star,\xi])_{ij}\delta_i^\perp\delta_j^\perp + O(\delta^3)\,. \qquad \text{(By (F.3) and (F.4))}\end{aligned}$$

Since the Hessian is positive definite on $\mathcal{K}^\perp(x^\star)$, we obtain

$$\Gamma_\alpha^{(\text{mod})}[x,\xi] = \Gamma_\alpha^{(\text{mod})}[x^\star+\delta,\xi] > \Gamma_\alpha^{(\text{mod})}[x^\star,\xi]\ .$$

Hence, the system flows back to the fixed-point, $x^\star$. If the infinitesimal change does not contain the orthogonal complementary component, i.e., $\delta_i = \delta_i^{\parallel}$, then $x^\star+\delta^{\parallel} \in \mathcal{N}_{x^\star}$. By the assumption of $\mathcal{N}_{x^\star}$, $\Gamma_\alpha^{(\text{mod})}[x,\xi]$ remains unchanged on $\mathcal{N}_{x^\star}$. Hence, the higher-order correction vanishes, i.e., $O(\delta^3)=0$. Consider the gradient flow:

$$\begin{aligned}\frac{d}{dt}\{\Gamma_\alpha^{(\text{mod})}[x(t),\xi]-\Gamma_\alpha^{(\text{mod})}[x^\star,\xi]\} &= \sum_{i=1}^{d}(\nabla_x\Gamma_\alpha^{(\text{mod})}[x(t),\xi])_i\frac{dx_i(t)}{dt}\\ &= -\sum_{i=1}^{d}(\nabla_x\Gamma_\alpha^{(\text{mod})}[x(t),\xi])_i^2 \le 0\ ,\end{aligned}$$

where we use (F.1). The gradient flow approaches to $(x^\star+v)$ for some $v\in\mathcal{K}(x^\star)$. For gradient descent,

$$\begin{aligned}\delta_i(t+1) &= x_i(t+1)-x_i^\star\\ &= \delta_i(t)-\eta(\nabla_x\Gamma_\alpha[x(t),\xi])_i \qquad \text{(By (F.2))}\\ &= \sum_{j=1}\{\mathbb{1}_{\{i,j\}}-\eta(\nabla_x^2\Gamma_\alpha^{(\text{mod})}[x^\star,\xi])_{ij}\}\delta_j(t)\ . \qquad \text{(F.6)}\end{aligned}$$

To guarantee that the gradient descent converges for $\mathcal{K}^\perp(x^\star)$, we require that the maximal absolute eigenvalue of $\{\cdots\}$ above should be smaller than one. According to (F.4), the eigenvalues are all positive along $\mathcal{K}^\perp(x^\star)$. Hence,

$$|1-\eta\lambda_{\max}(\nabla_x^2\Gamma_\alpha^{(\text{mod})}[x^\star,\xi])|_{\mathcal{K}^\perp(x^\star)}|<1 \implies 0<\eta<\frac{2}{\lambda_{\max}(\nabla_x^2\Gamma_\alpha^{(\text{mod})}[x^\star,\xi])|_{\mathcal{K}^\perp(x^\star)}}\ .$$

If we do not restrict on $\mathcal{K}^\perp(x^\star)$, then the $\lambda_{\max}(\cdot)=0$ along $\mathcal{K}(x^\star)$. Hence, the gradient descent approaches to $(x^\star+v)$ for some $v\in\mathcal{K}(x^\star)$. This proves (F.5). □

Remark F.1. For symmetry and projection degeneracy, Proposition F.1 holds since symmetries are always global, i.e., the group element does not differ for different φ, and the projection is affine. However, for constrained systems, we require the condition that Γ_α remains the same on the degeneracy manifold of the constrained space. This does not always hold.

Proposition F.2 (State-Retrieval Stability as Query-Key-Retrieval Stability)**.** The dynamical retrieval stability in the macroscopic state space, $\mathcal{M}_x$, corresponds to which in the memory score z-space, $\mathcal{M}_z$. Specifically, let $x^\star$ be the stable fixed-point in $\mathcal{M}_x$, and $\mathcal{D}:=\ker(\xi_i^\mu)\subset\mathbb{R}^d$.

$$x(t)\to x^\star \implies z(t)\to z^\star := \frac{1}{\sqrt{d}}\sum_{i=1}^{d}\xi_i^\mu x_i^\star\ . \qquad \text{(F.7)}$$

Conversely,

$$z(t)\to z^\star \implies \Pi_{\mathcal{D}^\perp}x(t)\to\Pi_{\mathcal{D}^\perp}x^\star\ , \qquad \text{(F.8)}$$

where $\Pi_{\mathcal{D}^\perp}$ is the projection map to $\mathcal{D}^\perp$. Moreover, the convergence rates in z imply convergence rates in x via

$$\|\delta x\|_2 \le \frac{d}{\sigma_{\min}}\|\delta z\|_2\ ,\ \forall\ \delta x\in\mathcal{D}^\perp\ , \qquad \text{(F.9)}$$

where $\sigma_{\min}>0$ denotes the smallest non-zero singular value of (ξ_i^μ).

Proof. Since $\mathbb{R}^d = \mathcal{D} \oplus \mathcal{D}^\perp$, we decompose $x_i = x_i^{\parallel} + x_i^{\perp}$ with $\sum_{i=1}^d \xi_i^\mu x_i^{\parallel} = 0$. For (F.7), it is trivial since $\bar{z}$ is deterministic due to $z_\mu = \sum_{i=1}^d \xi_i^\mu x_i/\sqrt{d}$. For (F.8), we consider $\delta x(t) = \Pi_{\mathcal{D}^\perp}(x(t) - x^\star) \in \mathcal{D}^\perp$. Then

$$\frac{1}{\sqrt{d}} \sum_{i=1}^d \xi_i^\mu \delta x_i(t) = \frac{1}{\sqrt{d}} \sum_{i=1}^d \xi_i^\mu (x_i(t) - x_i^\star) = z_\mu(t) - z^\star \to 0.$$

If $\delta x \in \mathcal{D}$, then $\sum_{i=1}^d \xi_i^\mu \delta x_i(t) = \sum_{i=1}^d \xi_i^\mu \Pi_{\mathcal{D}}(x_i(t) - x_i^\star) = 0$ does not give any information of $x(t)$ and $x^\star$. We restrict to $\mathcal{D}^\perp$ now. Since on $\mathcal{D}^\perp$, (ξ_i^μ) is injective and $\delta x \in \mathcal{D}^\perp$ satisfies

$$\sum_{\mu=1}^M (z_\mu(t) - z_\mu^\star)^2 = \frac{1}{d} \sum_{\mu=1}^M (\sum_{i=1}^d \xi_i^\mu \delta x_i(t))^2 \geq \frac{\sigma_{\min}^2}{d} \sum_{i=1}^d \delta x_i^2(t),$$

we thereby prove (F.9). □

Proposition F.3 (Score-space Attention-Retrieval Stability)**.** Consider the fixed-point in the z-space be $z_\mu^\star$. Then the Jacobian of the update map is

$$\mathcal{J}_{\mu\nu}(z^\star) = -\frac{1}{\beta} \sum_{\rho=1}^M G_{\mu\rho} (\nabla_z^2 \widetilde{\Gamma}_\alpha^{(\text{mod})}[x^\star, p, \xi])_{\rho\nu} \,, \tag{F.10}$$

where $z_\mu^\star = \sum_{i=1}^d x_i^\star \xi_i^\mu / \sqrt{d}$. The dynamical retrieval update is stable at $z_\mu^\star$ if

$$\rho(\mathcal{J}(z^\star)) < 1 \,,$$

where $\rho(\cdot)$ is the spectral radius, i.e., the maximal absolute value of eigenvalues.

Proof. If we perturb around the fixed-point, $\delta z_\mu(t) \coloneqq z_\mu(t) - z_\mu^\star$ gives

$$\begin{aligned} \delta z_\mu(t+1) &= \sum_{\nu=1}^M G_{\mu\nu} p_\nu(z(t)) + b_\mu(J) - \sum_{\nu=1}^M G_{\mu\nu} p_\nu(z^\star) - b_\mu(J) \\ &= \sum_{\nu,\rho=1}^M G_{\mu\nu} \frac{\partial p_\nu}{\partial z_\rho}|_{z=z^\star} \, \delta z_\rho(t) \,. \end{aligned}$$

According to (3.8), we express p in terms of the gradient of $\widetilde{\Gamma}_\alpha^{(\text{mod})}$. In this way,

$$\delta z_\mu(t+1) = -\frac{1}{\beta} \sum_{\nu,\rho=1}^M G_{\mu\nu} (\nabla_z^2 \widetilde{\Gamma}_\alpha^{(\text{mod})}[x^\star, p, \xi])_{\nu\rho} \delta z_\rho(t)$$

By $\delta z_\mu(t+1) = \sum_{\nu=1}^M \mathcal{J}_{\mu\nu} \delta z_\nu(t)$, we obtain the Jacobian, (F.10). If $\rho(\mathcal{J}) < 1$, then $\mathcal{J}^t \to 0$ as $t \to \infty$. Hence, $\delta z_\mu(t) \to 0$, and the system converges to the fixed-point, $z_\mu^\star$. □

Remark F.2. The update map is a non-linear discrete map. Hence, it is neither a gradient flow nor a gradient descent in general. This fixed-point perturbation analysis is consistent to the correspondence between the gradient-based optimization in $\mathcal{M}_x$.

Now we define the Tsallis divergence which helps us characterizes the training dynamics and the convergences. It generalizes KL-divergence and induces dual geometric structures on the probability simplex. (Amari, 1985)

Definition F.2 (Tsallis Divergence (Furuichi et al., 2004))**.** The Tsallis divergence generalizes the Kullback-Leibler (KL) divergence (Kullback & Leibler, 1951). It has the following form:

$$D_\alpha(p\|q) \coloneqq -\sum_{\mu=1}^{M} p_\mu \log_\alpha(\frac{q_\mu}{p_\mu}) = \frac{1}{\alpha-1}(\sum_{\mu=1}^{M} p_\mu^\alpha q_\mu^{1-\alpha} - 1) \,.$$

Remark F.3. If $\alpha \to 1$, the Tsallis divergence reduces to the KL divergence. It follows from Lemma D.1 in which $\log_\alpha \to \log$ as $\alpha \to 1$. Hence,

$$D_{\alpha\to 1}(p\|q) = -\lim_{\alpha\to 1}\sum_{\mu=1}^{M} p_\mu \log_\alpha(\frac{q_\mu}{p_\mu}) = \sum_{\mu=1}^{M} p_\mu \log\left(\frac{p_\mu}{q_\mu}\right) \coloneqq D_{\mathrm{KL}}(p\|q) \,.$$

The dynamical stability of p relates to the attention-retrieval stability via the Tsallis divergence.

Proposition F.4 (Tsallis Divergence and Attention-Retrieval Stability)**.** Let $\alpha > 1$, $p^\star \in \Delta^M$ be the fixed maximizer in the simplex interior satisfying $z_\mu^\star = (Gp)_\mu^\star + b_\mu(J)$. Then there exist constants, $c_1, c_2 > 0$ such that

$$c_1\|z(t) - z^\star\|_2^2 \le D_\alpha(p(t)\|p^\star) \le c_2\|p(t) - p^\star\|_2^2 \,.$$

Proof. Since

$$\frac{\partial}{\partial p_\mu} D_\alpha(p\|q) = \frac{\alpha}{\alpha-1} q_\mu^{1-\alpha} p_\mu^{\alpha-1} \,, \quad \frac{\partial^2}{\partial p_\mu \partial p_\nu} D_\alpha(p\|q) = \alpha q_\mu^{1-\alpha} p_\mu^{\alpha-2} \mathbb{1}_{\{\mu,\nu\}} \,,$$

we have the Taylor expansion

$$\begin{aligned} D_\alpha(p(t)\|p^\star) &= D_\alpha(p^\star\|p^\star) + \frac{\alpha}{\alpha-1}\sum_{\mu=1}^{M}(p_\mu(t) - p_\mu^\star) + \frac{\alpha}{2}\sum_{\mu=1}^{M}\frac{1}{p_\mu^\star}(p_\mu(t) - p_\mu^\star)^2 + O(\|p(t) - p^\star\|_2^3) \\ &= \frac{\alpha}{2}\sum_{\mu=1}^{M}\frac{1}{p_\mu^\star}(p_\mu(t) - p_\mu^\star)^2 + O(\|p(t) - p^\star\|_2^3) \,. \end{aligned}$$

Since $p^\star$ is in the simplex interior, $1 > p_\mu^\star > 0$ for all μ. Hence, there exist $c_1, c_2 > 0$ such that

$$c_1\|p(t) - p^\star\|_2^2 \le D_\alpha(p(t)\|p^\star) \le c_2\|p(t) - p^\star\|_2^2 \,.$$

Moreover,

$$z_\mu(t) - z_\mu^\star = [G(p(t) - p^\star)]_\mu \implies \|z(t) - z^\star\|_2 \le \|G\|_{\mathrm{op}} \cdot \|p(t) - p^\star\|_2 \,.$$

Combining these two inequalities, we prove this proposition. □

Remark F.4. From Proposition F.4, we notice that the convergence in $\mathcal{M}_z$ does not control the convergence of the Tsallis divergence and the convergence in p-space. If the Gram matrix is invertible, then $\|p(t) - p^\star\|_2 \le \|G^{-1}\|_{\mathrm{op}} \cdot \|z(t) - z^\star\|_2$. In this way, the convergence in $\mathcal{M}_x$ controls the convergence in $\mathcal{M}_z$ and the p-space. However, in the real world, the high correlations between memory patterns make $G_{\mu\nu}$ non-invertible. Hence, we do not guarantee the convergence of $p^\star$.

G NUMERICAL STUDIES

In our simulation, we show the retrieval structure of the sparse modern Hopfield model by scanning the inverse temperature, β, and the exponential memory load, $\kappa = \log(M)/d$, in both the sparse regime ($\alpha = 2$) and the softmax regime ($\alpha = 1$).

Firstly, we should introduce the effective support size and sparsity, which help us verify whether the stable points possess spurious information that may lead to retrieval failure.

G.1 EFFECTIVE SUPPORT SIZE AND SPARSITY

To relate entropy functions to the size and the randomness of information concretely, we introduce the effective support size and the sparsity here.

Definition G.1 (Effective Support Size (Hill, 1973))**.** We define the (Tsallis-Rényi) effective support size as

$$K_{\text{eff}}^{(\alpha)}[p] \coloneqq (\sum_{\mu=1}^{M} p_\mu^\alpha)^{\frac{1}{1-\alpha}} ,$$

where $p \in \Delta^M$ satisfying $\sum_{\mu=1}^{M} p_\mu = 1$ and $\alpha \in \mathbb{R}_+/\{1\}$.

Proposition G.1 (Range of Effective Support Size (Hill, 1973))**.** The effective support size satisfies the inequality:

$$1 \leq K_{\text{eff}}^{(\alpha)}[p] \leq M$$

If $K_{\text{eff}}^{(\alpha)}[p] = 1$, then p is a peak. If $K_{\text{eff}}^{(\alpha)}[p] = M$, then p is uniform.

Proof. We repeat the proof from (Hill, 1973) for completeness. We decompose the range of α into $\alpha > 1$ and $1 > \alpha > 0$ and focus on the former case first. For $\alpha > 1$, we have $\sum_{\mu=1}^{M} p_\mu^\alpha \leq 1$ and

$$\begin{aligned}
1 \leq (\sum_{\mu=1}^{M} p_\mu^\alpha)^{\frac{1}{1-\alpha}} &= K_{\text{eff}}^{(\alpha)}[p] \\
&= M^{\frac{1}{1-\alpha}} \cdot (\frac{1}{M} \sum_{\mu=1}^{M} p_\mu^\alpha)^{\frac{1}{1-\alpha}} \\
&\leq M^{\frac{1}{1-\alpha}} \cdot [(\frac{1}{M} \sum_{\mu=1}^{M} p_\mu)^\alpha]^{\frac{1}{1-\alpha}} && \text{(By the Jensen's Inequality)} \\
&= M .
\end{aligned}$$

For $1 > \alpha > 0$, then $1 \leq \sum_{\mu=1}^{M} p_\mu^\alpha$ and the similar derivation holds. We note that it saturates 1 when p is a vertex on the simplex, i.e., $p_{\mu_{\max}} = 1$ and $p_{\mu \neq \mu_{\max}} = 0$, where $\mu_{\max} \coloneqq \arg\max_\mu \ p_\mu$ is non-degenerate. It saturates the Jensen's inequality when $p_\mu = M^{-1}$ for every μ. □

Remark G.1. By Definition D.8,

$$K_{\text{eff}}^{(\alpha)}[p] = (1 - (\alpha - 1)\Psi_\alpha[p])^{\frac{1}{1-\alpha}} ,$$

we differentiate it with respect to Ψ_α. It yields $(K_{\text{eff}}^{(\alpha)}[p])^{\alpha/(1-\alpha)} > 0$. Hence, the effective support size is strictly increasing (monotonic) with respect to the Tsallis entropy. The higher the entropy is, the larger the effective support size is.

Lemma G.1 (Threshold of Probability Function (Martins & Astudillo, 2016; Peters et al., 2019))**.** The minimizer, $p \in \Delta^M$, of $\Gamma_\alpha^{(\text{mod})}[x, p, \xi]$ satisfies

$$p_\mu = (\frac{\alpha-1}{\alpha}\beta[z_\mu - \theta]_+)^{\frac{1}{\alpha-1}} , \tag{G.1}$$

where $[u]_+ \coloneqq \max\{u, 0\}$. We determine θ by the normalization condition, $\sum_{\mu=1}^M p_\mu = 1$. This truncation only appears for $\alpha > 1$. The support of the minimizers is thereby

$$\text{supp}(p) = \{\mu \in [1, M] : z_\mu > \theta\} .$$

We define $|\text{supp}(p)|$ as the exact support size.

Proof. We include the proof for completeness, adapting the standard argument from (Martins & Astudillo, 2016; Peters et al., 2019) to our notation. Since $0 \le p_\mu \le 1$ and $\sum_{\mu=1}^M p_\mu = 1$, we impose these constriants as Lagrangian multiplier terms:

$$\mathcal{L}(p, z, \theta, \lambda) = \beta \sum_{\mu=1}^M z_\mu p_\mu + \frac{1}{\alpha-1}(1 - \sum_{\mu=1}^M p_\mu^\alpha) + \beta\theta(\sum_{\mu=1}^M p_\mu - 1) + \sum_{\mu=1}^M \lambda_\mu p_\mu .$$

Let the optimal saddles that maximizes the Lagrangian function be $\bar{p}$, $\bar{z}$, $\bar{\theta}$, and $\bar{\lambda}$. Then the Karush-Kuhn-Tucker (KKT) conditions are

$$\bar{p}_\mu \ge 0 , \sum_{\mu=1}^M \bar{p}_\mu = 1, \qquad \text{(Primal)}$$

$$\bar{\lambda}_\mu \ge 0, \qquad \text{(Dual)}$$

$$\bar{\lambda}_\mu \bar{p}_\mu = 0, \qquad \text{(Complementary)}$$

$$\frac{\partial \mathcal{L}}{\partial p_\mu}|_{p=\bar{p}, z=\bar{z}, \theta=\bar{\theta}, \lambda=\bar{\lambda}} = \beta\bar{z}_\mu - \frac{\alpha}{\alpha-1}(\bar{p}_\mu)^{\alpha-1} + \beta\bar{\theta} + \bar{\lambda}_\mu = 0, \qquad \text{(Stationary)}$$

for all $\mu \in [1, M]$. By solving the stationary equation, for $\bar{p}_\mu > 0$ and $\alpha > 1$,

$$\bar{p}_\mu = (\frac{\alpha-1}{\alpha}\beta[\bar{z}_\mu + \bar{\theta}])^{\frac{1}{\alpha-1}} .$$

For $\bar{p}_\mu = 0$ and $\alpha > 1$,

$$\bar{z}_\mu + \bar{\theta} + \bar{\lambda}_\mu = 0 \implies \bar{z}_\mu + \bar{\theta} \le 0 .$$

Hence, after replacing $\bar{\theta} \to -\bar{\theta}$, we have

$$p_\mu = (\frac{\alpha-1}{\alpha}\beta[z_\mu - \theta]_+)^{\frac{1}{\alpha-1}} ,$$

where we neglect bars to avoid redundant notations. □

Remark G.2. For $0 < \alpha < 1$, the exponent of (G.1) is negative, so the base should remain positive. In this way, there is no truncation and all minimizers $p_\mu > 0$. For $\alpha \to 1$, the Shannon entropy has the stationary equation,

$$\beta z_\mu - \log(p_\mu) - 1 + \beta\theta + \lambda_\mu = 0 .$$

The minimizer, p_μ, is positive everywhere and does not truncate for finite $(\theta, z_\mu, \lambda_\mu)$.

Definition G.2 (Sparsity (Martins & Astudillo, 2016; Peters et al., 2019)). We say $p \in \Delta^M$ is k-sparse if the exact support size satisfies

$$|\text{supp}(p)| = k \,,$$

where $0 \le k \le M$. The large the k is, the less the sparsity is. In other words, if more points satisfy $p = 0$, then the system is sparser.

Remark G.3. Higher entropy suggests larger effective support size but not higher sparsity. The sparsity depends on the boundary geometry of the entropy function. Lemma G.1 states that only for $\alpha > 1$, there is a truncation and several p_μ are zero. It is consistent with the argument in Proposition D.3. Due to Definition D.4, the Tsallis entropy is not essentially smooth only for $\alpha > 1$. Therefore, for $\alpha > 1$, the optimizer sit on the boundaries of the simplex when the corresponding z_μ is below the threshold. This introduces the sparsity. Otherwise, the divergence of gradient on the boundary forbids the sparsity.

G.2 Settings.

Setup. We consider a system of dimension d with $M = e^{\kappa d}$ normalized stored patterns, $\{\xi^\mu\}_{\mu=1}^M$. To probe the energy landscape, $\Gamma_\alpha^{(\text{mod})}$, we perform continuous scans by varying β or κ. For each scan, we use both up and down sweeps to detect hysteresis and metastability.

1. **Up Sweep**: Initialize from low β (or low κ) and increase.

2. **Down Sweep**: Initialize from high β (or high κ) and decrease.

At each parameter value (β or $|\kappa$), we initialize the optimization from the converged solution at the previous step (warm start). Then we minimize the free energy with respect to x by gradient descent,

$$x(t+1) = x(t) - \eta(t)\nabla_x \Gamma_\alpha^{(\text{mod})}(x(t)), \qquad \eta(t) = \eta_0 2^{-m(t)},$$

where $\eta_0 = 0.5$ and $m(t) := \min\{m \in \mathbb{N}_0 \mid \Gamma_\alpha^{(\text{mod})}(x(t+1)) \le \Gamma_\alpha^{(\text{mod})}(x(t))\}$. This procedure tracks continuous branches of local minima of $\Gamma_\alpha^{(\text{mod})}$. To explore competing solutions, we perform multi-branch continuation.

Data. In our experiment, we consider clustered normalized patterns.

$$\xi_i^\mu = \sqrt{1-\delta}\,\eta_i^{c(\mu)} + \sqrt{\delta}\,\zeta_i^\mu,$$

where $c(\mu) \in \{1, ..., C\}$ assigns pattern μ to a cluster, $\eta_i^{c(\mu)}$ are cluster centers sampled as $\eta^c \sim \mathcal{N}(0, \mathbb{1}_d)$, $\zeta^\mu \sim \mathcal{N}(0, \mathbb{1}_d)$ are i.i.d. normalized noise vectors, and $\delta \in [0, 1]$ indicates the noise strength. This construction introduces off-diagonal correlations in the Gram matrix, which makes the simulation more realistic than the i.i.d. setting. In our simulation, we choose $C = 4$ and $\delta = 0.04$.

Observables. To characterize the coexistence regimes and retrieval success, we measure the following quantities.

1. **Maximum Overlap:** To verify whether a stable point aligns with the stored patterns, we measure the maximum overlap. $z^\star_{\text{max}} = \max_\mu |z^\star_\mu|$.

2. **Exact Support Size:** Each stable point has its own support and exact support size. For example, $z^\star_\mu = \sum_{i=1}^d x_i^\star \xi_i^\mu / \sqrt{d}$ gives a specific value of $p^\star_\mu$ via Lemma G.1, which determines $K_{\text{eff}}^{(\alpha>1)}$. For $\alpha = 2$, we measure $|\text{supp}(p^\star)| = \#\{\mu \in [1, M] : z^\star_\mu > \theta\}$ according to Lemma G.1, where $\#$ indicates the cardinality. For $\alpha = 1$, we measure $\{\mu \in [1, M] : p^\star_\mu > \varepsilon\}$, where we choose

$\varepsilon = 10^{-10}$ in our simulation. Hence, we denote the measured exact support size as

$$|\text{supp}(p^\star)| \coloneqq \begin{cases} \#\{\mu \in [1, M] : p^\star_\mu > 10^{-10}\} & , \alpha = 1 \\ \#\{\mu \in [1, M] : z^\star_\mu > \theta\} & , \alpha = 2 \end{cases}.$$

3. **Fixed Scored-Threshold Support Size:** In temperature analysis in the sparse case, we further define

$$K_\Delta \coloneqq \{\mu \in [1, M] : z^\star_\mu > \Delta\},$$

and we choose $\Delta = 0.01$. Hence, it serves as a diagnostic for counting how many memory components achieve large query-key alignment.

4. **Inverse Participation Ratio (IPR):** It is defined as the effective support size with $\alpha = 2$, i.e., $K^{\text{IPR}}_{\text{eff}} \coloneqq K^{(2)}_{\text{eff}} = (\sum_\mu (p^\star_\mu)^2)^{-1}$. If IPR is large, then there are more components of $p^\star_\mu$ that are non-zero.

5. **Free Energy:** Distinct stable points have distinct free energy in general. To see which is more preferable, i.e., which one has the lowest free energy, we measure $\Gamma^{(\text{mod})}_\alpha[x, \xi]$.

G.3 Temperature Analysis ($\alpha = 2$)

We start from the temperature analysis in the sparse case. Here, we select four informative plots to represent the retrieval (phase) structure in Figure 8.

(a) The up- and down-sweep branches do not coincide, which shows a hysteresis window in β. This indicates that the coexistence of multiple local stable retrieval states. The selected branch depends on initialization, analogous to first-order transitions with supercooling and superheating. As β increases, the maximal overlap $z^\star_{\text{max}}$ grows stepwise and the shaded regions collapse at large β. This indicates a seed-independent saturation. Specifically, it saturates at $z^\star_{\text{max}} \approx 0.0156$. This is the consequence of one pattern retrieval,

$$z^\star_\mu = \frac{1}{\sqrt{d}} \sum_{i=1}^{d} \xi^\mu_i x^\star_i = \frac{1}{\sqrt{d}} \times d \times \frac{1}{\sqrt{d}} \times \frac{1}{d} = \frac{1}{d} = \frac{1}{64} \approx 0.0156.$$

(b) As β increases, the exact support size and IPR decreases. This indicates a transition from distributed to sparse retrieval. The discrepancy of $|\text{supp}(p^\star)|$ and IPR arises since IPR counts the components non-uniformly. However, K_Δ is zero in the high-temperature regime, where noise dominates and no overlap exceeds the threshold $\Delta = 0.01$. This suggests a disordered regime. In the low-temperature regime, all support sizes approach to one, so the system reaches single-pattern retrieval with large overlap. The collapse of the shaded regions again indicates a perfect retrieval and seed-independence.

(c) Initialization at $\beta_0 = 290$. Multi-branch continuation indicates several competing stable retrieval branches in the intermediate-β regime. Their crossings and discontinuous jumps in $z^\star_{\text{max}}$ show that the sparse model admits branch switching between distinct local minima. In this regime, retrieval remains nontrivial. The selected state correspond to either stronger or weaker alignment depending on the branch we are tracking.

(d) Initialization at $\beta_0 = 290$. The branch exhibits coexistence region. As β varies, the global minimizer switches between branches. However, under continuation (warm-start dynamics), the system follows a given local branch rather than the instantaneous global minimizer. Hence, the transition between branches occur before reaching the exact crossing point. This results in discontinuous changes in the selected retrieval state. Hence, we interpret this as a transition within the retrieval phase rather than a transition between retrieval and a disordered phase. This simulation provides a more precise phase structure than earlier analyses, (Amit et al., 1985), (Amit et al., 1987), and (Nishimori, 2001).

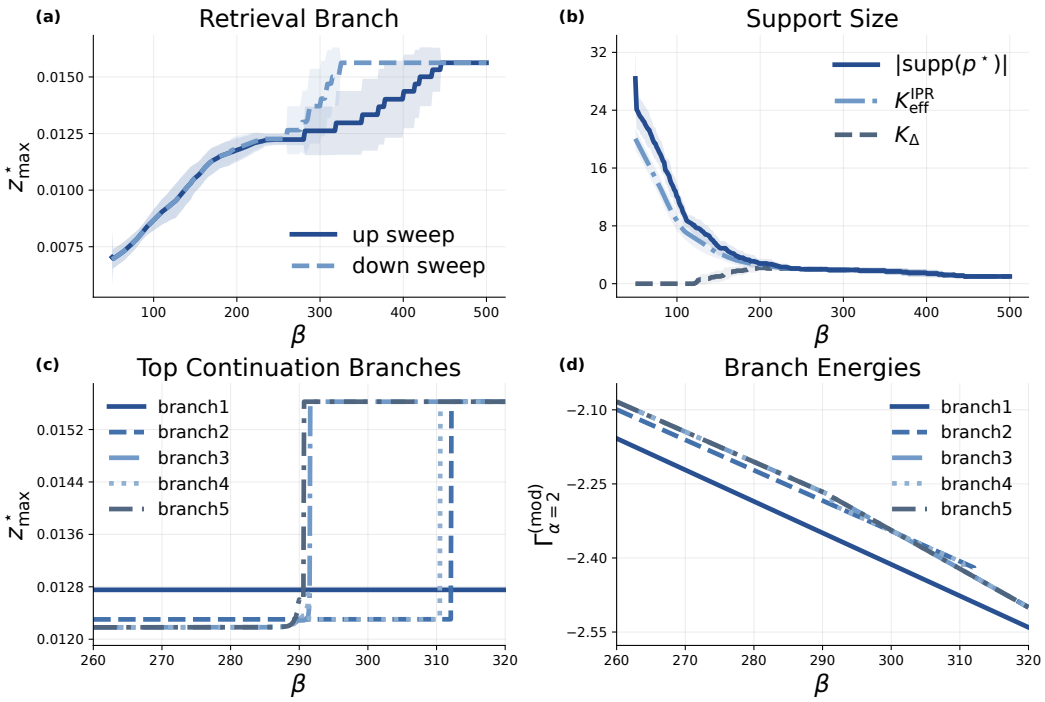


Figure 8: **Temperature analysis of the retrieval structure of MHM in the sparse regime ($\alpha = 2$).** ($M = 100, d = 64$). Results are averaged over 10 random pattern realizations, and shaded regions indicate one standard deviation across different seeds. **(a) Retrieval branches.** The maximal overlap $z^\star_{\max}$ under up- and down-sweeps of β exhibits hysteresis. This indicates the presence of multiple local stable branches (evolution of local stable points). The collapse of shaded regions at high β suggests a seed-independent saturation of $z^\star_{\max}$. **(b) Support size.** We compare the exact support size (number of active components of the probability distribution p), the inverse participation ratio (IPR), and the fixed score-threshold count. As β increases, all support sizes approach to one and their shaded region shrink, reflecting the collapse to a single dominant memory. This demonstrates the sparsity of the support. Low K_Δ in the high-temperature regime indicates a low query-key overlap and the disordered regime. **(c) Top branches.** We track the top five stable branches ranked by $z^\star_{\max}$. Discontinuous jumps indicate the changes in support and reflects non-smooth transitions. **(d) Branch energies.** We show the corresponding effective energies of those top five branches. The energy ordering of these branches changes, indicating the presence of a coexistence region. Under continuation, the system follows a given branch even beyond the point where it is no longer the global minimum. Hence, it explains the hysteresis.

Remark G.4. For the sparse temperature scan, the observed jumps are primarily associated with changes in the active support. Although a spinodal event is able to occur when the minimum Hessian eigenvalue approaches zero, this is not the generic mechanism observed in our scans. In most cases, the discontinuities arise from the non-smooth change of the sparse active set before a smooth branch loses stability via a vanishing Hessian eigenvalue.

Remark G.5. From Section 4, β_{coexist} is not the same as the real jump, β_{jump}, unless they satisfy specific assumptions, e.g., uniform distribution and well-separation conditions. In our experiment, distinct initialization gives distinct selected branches, which are able to remain separable for a specific finite range of β. Hence, β_{coexist} and β_{jump} are different in general. The uniform distribution and well-separation conditions do not always hold.

G.4 Temperature Analysis ($\alpha = 1$)

We provide a temperature analysis in the dense case. Here, we select four informative plots to represent the retrieval (phase) structure in Figure 9.

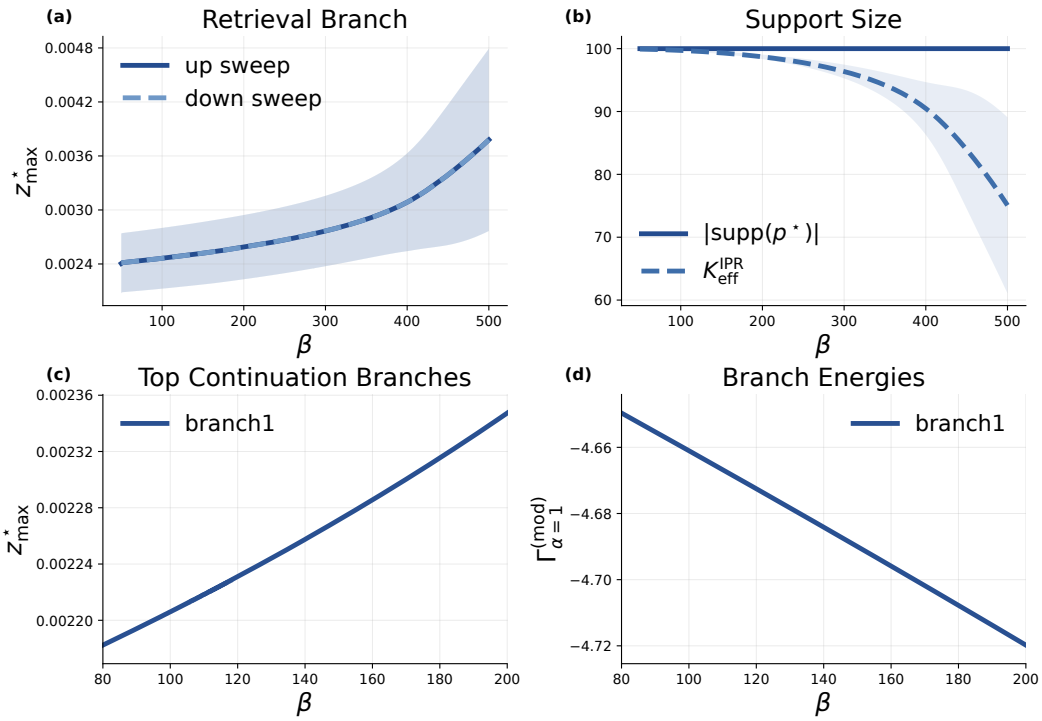


Figure 9: **Temperature analysis of the retrieval structure of MHM in the dense regime** ($\alpha = 1$)**.** ($M = 100, d = 64$). Results are averaged over 10 random pattern realizations, and shaded regions indicate one standard deviation across different seeds. The increase in deviation reflects the sensitivity of the dense regime. **(a) Retrieval branches.** The maximal overlap $z^\star_{\max}$ under up- and down-sweeps of β shows no hysteresis. This indicates the absence of metastability and the presence of a unique stable solution. **(b) Support size.** The exact support remains full, i.e., all components of the probability distribution p are active. In contrast, the inverse participation ratio (IPR) decreases with β monotonically. This indicates that the distribution becomes more concentrated but does not collapse to a single pattern. **(c) Top branches.** We observe only a single stable branch across the entire range of β. This confirms the absence of coexistence state of multiple local stable points. **(d) Branch energies.** The free energy $\Gamma^{(\mathrm{mod})}_{\alpha}$ evoloves without discontinuities. This indicates a smooth deformation of the energy landscape.

(a) The up- and down-sweep branches coincide, which indicates the absence of hysteresis. Hence, the retrieval dynamics admits a unique stable solution for each β. There is no metastable branch or discontinuous transition. The shaded region becomes broader as β increases. Hence, the alignment with stored pattern increases, but the variance increases as well.

(b) The support size remains full for all β. This is consistent with the softmax structure $p^\star_\mu \propto \exp\left(\beta z^\star_\mu\right)$ since it assigns positive weight to every component. As β increases, $K_{\mathrm{eff}}^{\mathrm{IPR}}$ decreases, so the distribution becomes concentrated. However, this does not induce sparsity. The shaded region of $K_{\mathrm{eff}}^{\mathrm{IPR}}$ becomes broader as β increases. This is due to the amplification of small differences from seeds.

(c) Initialization at $\beta_0 = 140$. The maximal overlap $z^\star_{\max}$ grows with β. This indicates a continuous sharpening of alignment with the target pattern. There is no discontinuity or spinodal instability (vanishing Hessian). It is also consistent with the absence of phase-transition-like behavior in the softmax regime.

(d) Initialization at $\beta_0 = 140$. The energy decreases along the branch, and the retrieval state becomes more concentrated. This reflects a continuous sharpening of the energy basin rather than a restructuring of competing minima.

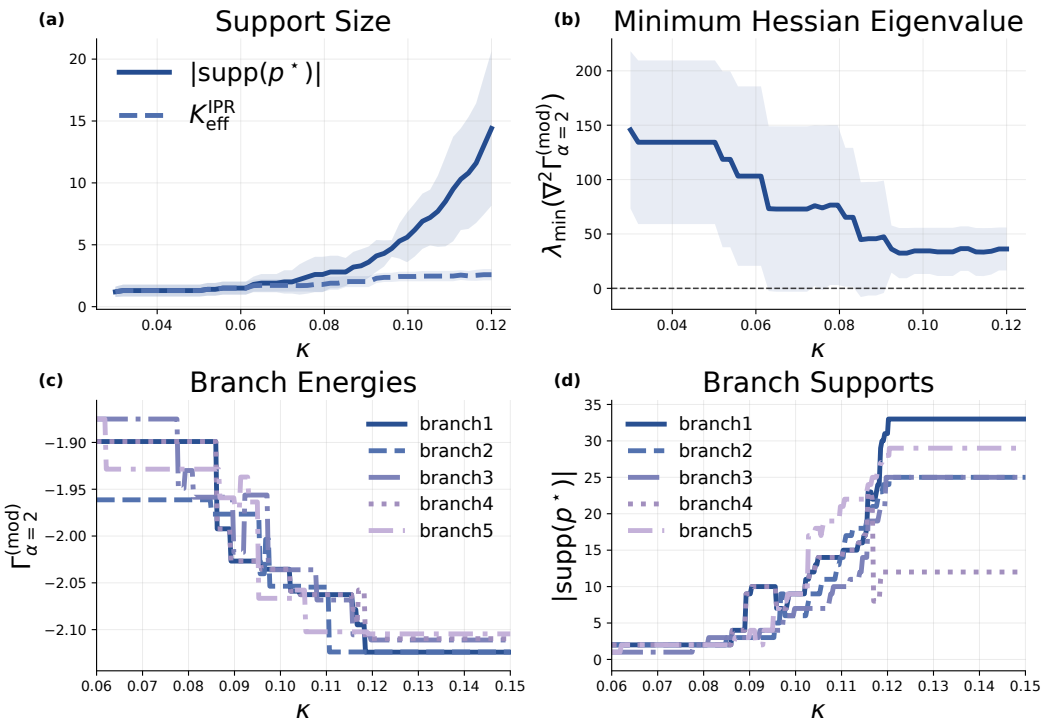


Figure 10: **Exponential memory load analysis for the retrieval structure of MHM in the sparse regime ($\alpha = 2$).** ($M = \exp(\kappa d), d = 48, \beta = 180$). Results are averaged over 10 random pattern realizations, and shaded regions indicate one standard deviation across different seeds. **(a) Support size.** We compare the exact support size $|\mathrm{supp}(p^\star)|$ (number of active components of the probability distribution p) and the inverse participation ratio (IPR). As κ increases, the exact one grows significantly, indicating that more patterns participate. However, the IPR shows that the distribution remains approximately sparse with few dominant components. **(b) Minimum Hessian eigenvalue.** The smallest eigenvalue of the Hessian $\lambda_{\min}(\nabla^2 \Gamma_\alpha^{(\mathrm{mod})})$ is of order $\beta = 180$ initially. As κ increases, more memories enter and lowers the mean eigenvalue. The lower edge of the shaded region crosses zero slightly, suggesting that some realizations induce instability, although the mean curvature remains positive. **(c) Lowest branch energies.** We track the lowest five continuation branches of the free energy. These discontinuous transitions indicates the changes in support rather than smooth deformation. **(d) Branch supports.** The exact support sizes jump to larger values at high memory load. This indicates that single-pattern retrievals break down.

Remark G.6. Prior work, (Koulischer et al., 2023), analyzes temperature-dependent behavior in MHM under softmax attention ($\alpha = 1$). They show that as β increases, the retrieval weights concentrate smoothly. Our results here are consistent with their findings. Besides, we further provides a quantitative characterization of this behavior by tracking the maximal overlap $z^\star_{\max}$, and the support sizes, supp and $K_{\mathrm{eff}}^{(\mathrm{IPR})}$.

G.5 Exponential Memory Load Analysis ($\alpha = 2$)

Here we study the exponential memory load, $M = \exp(\kappa d)$, in the sparse case. We select four informative plots to represent the retrieval (phase) structure in Figure 10.

(a) The system remains in a sparse retrieval regime: although additional overlaps enter the active set as κ increases, the IPR remains close to one. This indicates that a single dominant memory continues to control the retrieval. No delocalization occurs despite the exponential growth in memory size. Only the variance of supp is large since $\mathrm{supp} = \{\mu : z^\star_\mu > \theta\}$ is sensitive to seeds.

(b) The minimal eigenvalue of the Hessian is of order $\beta = 180$ initially. This follows from the single-pattern support limit. According to the Hessian (D.14) of one single pattern, we have

$$\nabla_x^2 \Gamma^{(\mathrm{mod})}[x^\star, \xi] = \beta \mathbb{1}_d - \frac{\beta^2}{2d} \sum_{\mu,\nu=1}^{k} \xi^\mu (\xi^\nu)^\top (\mathbb{1}_{\{\mu,\nu\}} - \frac{1}{k}) \approx \beta \mathbb{1}_d.$$

As κ increases, more memories enter the active set. Hence, the contribution of seeds lowers $\lambda_{\min}(\nabla_x^2 \Gamma_2^{(\mathrm{mod})})$. The shaded region is largest in the crossover regime, where different random pattern realizations enter the multi-support regime at different κ. At larger κ, most seeds enter the multi-support regime, so the variance decreases even though the mean curvature remains lower than in the single-pattern limit. Notice that the lower edge of the shaded region slightly crosses zero. This indicates that some realizations become close to instability, although the mean curvature remains positive.

(c) Initialization at $\kappa_0 = 0.12$. For $\kappa \gtrsim 0.1$, multiple continuation branches emerge and coexist. They correspond to distinct local stable retrieval states. The selected state depends on initialization and the continuation path. This reflects metastability in the energy landscape, i.e., multiple stable points coexist and the dynamics remains trapped in a given basin until stability is lost.

(d) Initialization at $\kappa_0 = 0.12$. For $\kappa \gtrsim 0.1$, several local stable points undergo support changes. For $\kappa \gtrsim 0.12$, no branch remains at supp $= 1$. This indicates that the pattern interference becomes large. Hence, the sparse single-pattern retrieval breaks down, and retrieval involves multiple active components.

G.6 Exponential Memory Load Analysis ($\alpha = 1$)

Here we study the exponential memory load, $M = \exp(\kappa d)$, in the dense case. We select four informative plots to represent the retrieval (phase) structure in Figure 11.

(a) Both the exact support size and IPR grow with κ. This reflects the dense nature of softmax, i.e.,

$$p_\mu^\star = \frac{\exp\{\beta z_\mu^\star\}}{\sum_{\nu=1}^{M} \exp\{\beta z_\nu^\star\}}, \qquad \text{with overlap } \; z_\mu^\star.$$

All patterns have nonzero weight, so no support selection occurs. Hence, retrieval loses selectivity in the high-load regime. The shaded regions are small since $p_\mu^\star > 0$ for softmax functions and supp $\approx M$ for all seeds.

(b) The minimal Hessian eigenvalue remains positive throughout the scan. Hence, there is no spinodal instability. Unlike the sparse case, the minimal eigenvalue increases with κ. This indicates that the energy landscape becomes smoother due to averaging over many patterns. The shaded region is large since the state contains multiple patterns. Hence, the Hessian depends on seeds according to (D.14).

(c) Initialization at $\kappa_0 = 0.08$. Only a single continuation branch appear. The energy decreases with κ monotonically. This indicates a smooth deformation of the solution rather than transitions between competing states.

(d) Initialization at $\kappa_0 = 0.08$. The maximal overlap decreases with κ, which indicates that retrieval becomes delocalized. That is, in the high-load regime, the system forms a mixture over many patterns. The small peak near initialization arises from the biased toward the previous dominant pattern in early iterates. Hence, it reflects an adaptation effect rather than a geometric feature of the energy landscape.

Remark G.7. The curves show the non-smoothness as functions of κ. This is the consequence of the fact that we require M to be an integer in our simulation.

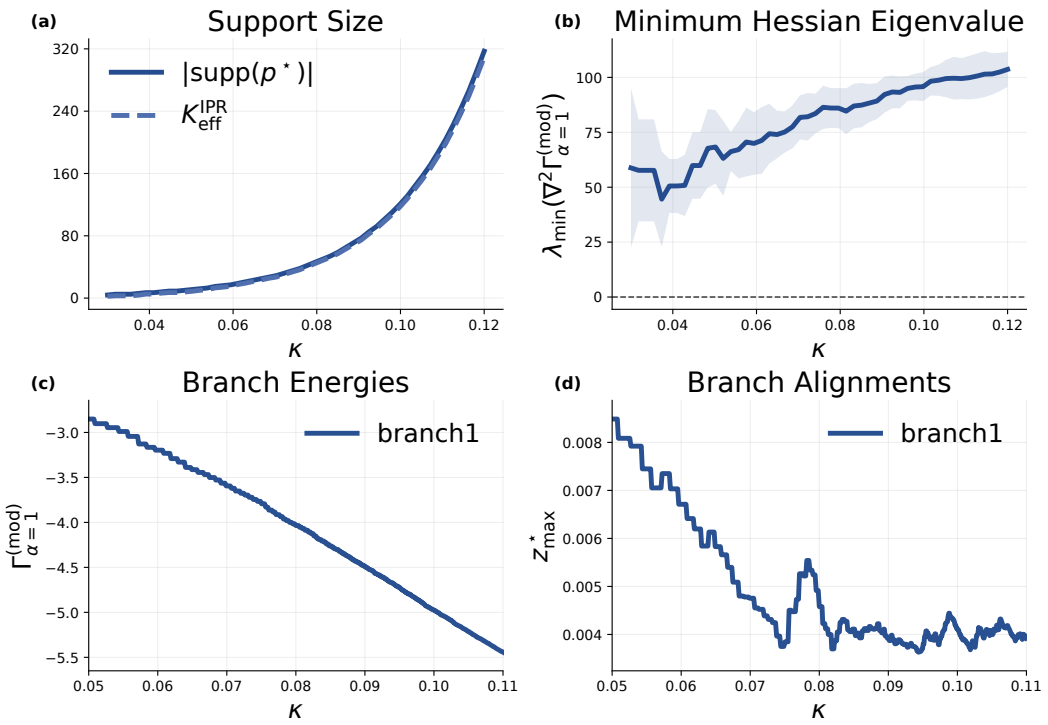


Figure 11: **Exponential memory load analysis for the retrieval structure of MHM in the dense regime ($\alpha = 1$).** ($M = \exp(\kappa d), d = 48, \beta = 180$). Results are averaged over random pattern realizations, and shaded regions indicate one standard deviation across different seeds. **(a) Support size.** We compare the exact support size (number of active components of the probability distribution p) and the inverse participation ratio (IPR). They both grow with κ. This indicates dense activation rather than a sparse support selection. **(b) Minimum Hessian eigenvalue.** The smallest eigenvalue of the Hessian $\lambda_{\min}(\nabla^2\Gamma_\alpha^{(\mathrm{mod})})$ remains positive and seed-dependent, i.e., has nonzero deviation. This indicates that the retrieval states are local stable and have no instability under increasing memory load. **(c) Lowest branch energies.** We observe only a single stable branch across the entire range of κ. This confirms the absence of coexistence state of multiple local stable points. **(d) Branch supports.** The maximal overlap $z^\star_{\max}$ decreases with κ. This indicates a delocalization to a mixed state with multiple patterns. The peak around $\kappa \approx 0.08$ arises from the warm-up initialization at $\kappa_0 = 0.08$.

G.7 Dynamical Stability Analysis for Temperature

In this section, we provide the numerical validation for the theoretical dynamical retrieval stability condition; please see Section F for a detailed discussion and derivations. Specifically, we only analyze the ordinary gradient descent (discrete) algorithm and verify the bound in Proposition F.1.

Setup. We consider the number of pattern be $M = 100$, pattern dimension be $d = 64$, and inverse-temperature be $\beta = 300$ as an example. For simplicity, we apply fixed-step gradient descent

$$x(t+1) = x(t) - \eta \nabla_x \Gamma_\alpha^{(\mathrm{mod})}(x(t)),$$

where η is a constant. Let $\eta_{\mathrm{crit}} = 2/\lambda_{\max}(\nabla^2\Gamma_\alpha^{(mod)}(x^\star))$ be the theoretical critical learning rate threshold, $\lambda_{\max}(\cdot)$ be the maximal eigenvalue, and $x^\star$ be a fixed local stable point. We investigate the convergence through iteration numbers, t. We define the initial condition as $x(t=0) = x^\star + Au_{\mathrm{soft}}$, where $A = 10^{-2}$ is a small amplitude and u_{soft} is the eigen-direction of the Hessian with the smallest eigenvalue. Notice that we derive u_{soft} numerically, so it is not the exact eigenvector but exhibits tiny component in other directions.

Data. We use the same pattern set $\{\xi^\mu\}_{\mu=1}^M$ as in the previous experiments, where each pattern $\xi^\mu \in \mathbb{R}^d$ is normalized. We generate this collection through the method in Section G.2.

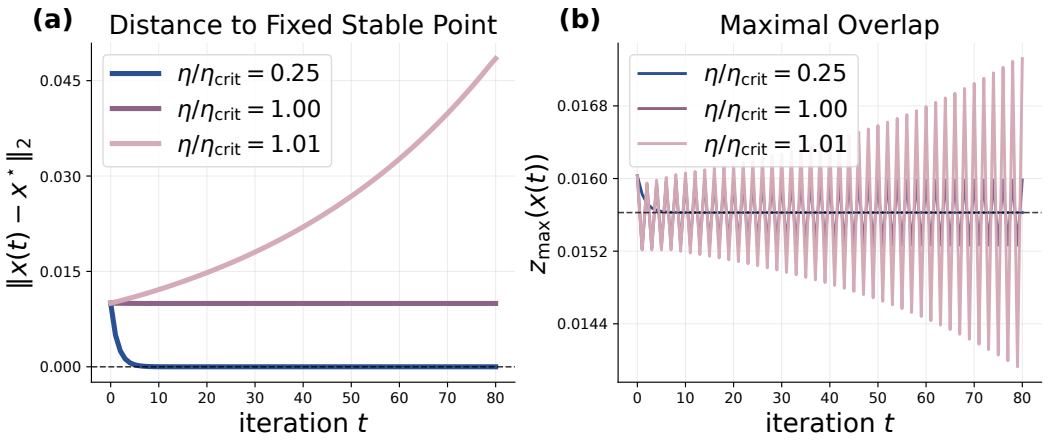


Figure 12: **Dynamical retrieval stability of MHM gradient descent in the sparse regime ($\alpha = 2$).** ($M = 100, d = 64$). We study the gradient descent update $x(t+1) = x(t) - \eta \nabla \Gamma_\alpha^{(\text{mod})}(x(t))$. We denote the critical learning rate as $\eta_{\text{crit}} = 2/\lambda_{\max}(\nabla^2 \Gamma_\alpha^{(\text{mod})}(x^\star))$ with $\lambda_{\max}$ the largest eigenvalue at a fixed stable point $x^\star$. **(a) Distance to the fixed point.** For $\eta < \eta_{\text{crit}}$, the training converges to the fixed point. Otherwise, the dynamics is unstable or even diverges. **Maximal overlap.** For $\eta < \eta_{\text{crit}}$, the training converges to the fixed value of overlap $z_{\max} \approx 0.0156$. It is the saturation value of one-pattern perfect retrieval in Figure 8 (a). Otherwise, the dynamics is unstable or even diverges.

Observables. We track two quantities along the gradient-descent trajectory: the distance to the fixed point, $\|x(t) - x^\star\|_2$ and the maximal overlap, $z_{\max}(x(t)) := \max_\mu z_\mu(x(t))$.

Results. In the sparse regime ($\alpha = 2$), we demonstrate the result in Figure 12.

(a) For $\eta < \eta_{\text{crit}}$, the distance (blue curve) approaches to zero asymptotically, indicating that the training $x(t)$ converges to the fixed stable point, $x^\star$. At $\eta \approx \eta_{\text{crit}}$, the convergence becomes marginal. For $\eta > \eta_{\text{crit}}$, the training $x(t)$ deviates from the local point as iteration increases. This indicates a dynamical instability. Therefore, this plot provides the evidence that η_{crit} separates the convergent regime.

(b) The maximal overlap $z_{\max}(x(t))$ exhibits the same behavior. For $\eta < \eta_{\text{crit}}$, $z_{\max}(x(t))$ converges to the fixed value $z_{\max}(x^\star)$. This indicates stable retrieval of the dominant memory pattern. Notably, the convergent value is $z_{\max}(x^\star) \approx 1/d \approx 0.0156$, which is the one-pattern retrieval in Figure 8 (a). At $\eta \approx \eta_{\text{crit}}$, it exhibits persistent oscillations around $z_{\max}(x^\star)$. For $\eta > \eta_{\text{crit}}$, the oscillations grow, indicating a dynamical instability in retrieval.

In the dense regime ($\alpha = 1$), we demonstrate the result in Figure 13.

(a) For $\eta < \eta_{\text{crit}}$, the iteration converges to the stable fixed point $x^\star$ rapidly. Larger the learning rate is, more rapid the convergence is. For $\eta > \eta_{\text{crit}}$, the distance declines first, and then grows up after some number of iteration. This phenomenon arises from unstable modes being small due to our initialization $x(t=0)$. In particular, from (F.6), we have $\delta_i(t+1) = (1 - \eta\lambda_i)\delta_i(t)$, where $\delta_i(t) \equiv x_i(t) - x_i^\star$, and λ_i is the i-th eigenvalue of the Hessian $\nabla_x^2 \Gamma_\alpha^{(\text{mod})}$. The convergence condition is

$$|1 - \eta\lambda_i| < 1, \qquad \text{or } 0 < \lambda_i < 2/\eta.$$

Hence, even if $\eta > \eta_{\text{crit}}$, some modes $x_i(t)$ still converge to the fixed point $x_i^\star$ if the eigenvalues satisfy $\lambda_i < 2/\eta$. Initially, we consider the dominant mode be which with $\lambda_{\min}$. Hence, it

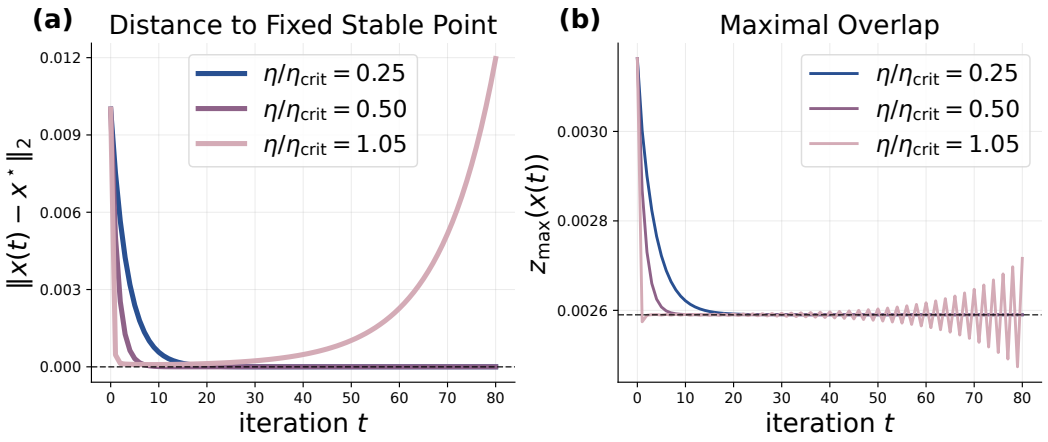


Figure 13: **Dynamical retrieval stability of MHM gradient descent in the dense regime ($\alpha = 1$).** ($M = 100, d = 64$). We study the gradient descent update $x(t+1) = x(t) - \eta \nabla \Gamma_\alpha^{(\text{mod})}(x(t))$. We denote the critical learning rate as $\eta_{\text{crit}} = 2/\lambda_{\max}(\nabla^2 \Gamma_\alpha^{(\text{mod})}(x^\star))$ with $\lambda_{\max}$ the largest eigenvalue at a fixed stable point $x^\star$. **(a) Distance to the fixed point.** For $\eta < \eta_{\text{crit}}$, the training converges to the fixed point. For $\eta > \eta_{\text{crit}}$, a divergence follows after a decay. This reflects that the iterate close to $x^\star$ initially. After this, the unstable eigen-directions start to dominate. We provide the precise explanation in this section. **(b) Maximal overlap.** For $\eta < \eta_{\text{crit}}$, the training converges to the fixed value of overlap $z_{\max}$. For $\eta > \eta_{\text{crit}}$, a growing oscillations follow after an initial convergence. This is similar for the behavior in the distance plot.

converges to the fixed point in the early stage. However, for $\lambda_i > 2/\eta$, $\delta_i(t)$ grows with t. Hence, the unstable modes start to dominate, and the trajectory diverges.

(b) For $\eta < \eta_{\text{crit}}$, the iteration converges to the stable fixed value $z_{\max}(x^\star)$. Larger the learning rate is, more rapid the convergence is. For $\eta > \eta_{\text{crit}}$, it exhibits the same behavior as in distance analysis. The dynamics converges first, then it diverges after some number of iteration. This arises from unstable modes being small due to our initialization $x(t = 0)$ as well.

H ADDITIONAL EXPERIMENTS

In this section, we analyze the empirical retrieval behaviors in realistic end-to-end Transformer-architectures. Furthermore, we validate the special coexistence point formula Proposition 4.1 and spinodal formula Proposition 4.2 numerically.

H.1 SUPPORT AND CAPACITY–RELIABILITY TRADE-OFF

We study how increasing the number of memories affects retrieval reliability and effective support (IPR) in a trained Transformer. We compare three attention maps, $\alpha \in \{1, 1.5, 2\}$ (softmax, entmax-1.5, and sparsemax), respectively.

Setup. The data generation pipeline is

$$u_Q \xrightarrow[\text{fixed}]{A_Q} q^{\text{in}}, \qquad \text{and } u_\mu \xrightarrow[\text{fixed}]{A_K} k_\mu^{\text{in}}.$$

We generate latent query and key variables and map them to observed inputs. After that, we input q^{in} and k_μ^{in} into a multi-head cross-attention Transformer, yielding the output $\widehat{y} \in \mathbb{R}^{d_v}$.

Data. Each example contains a query $q^{\text{in}} \in \mathbb{R}^{d_{\text{in}}}$, a set of memory keys $k_\mu^{\text{in}} \in \mathbb{R}^{d_{\text{in}}}$, and independent values $v_\mu \sim \mathcal{N}(0, \mathbb{1}_{d_v})$ with $d_v = 32$. The latent dimension is $d_{\text{lat}} = 16$, the input dimension is $d_{\text{in}} = 24$, and the number of latent clusters is $C = 4$. For each example, we draw normalized cluster centers independently

$$c_r = \frac{\widetilde{c}_r}{\|\widetilde{c}_r\|_2}. \qquad \widetilde{c}_r \sim \mathcal{N}(0, \mathbb{1}_{d_{\text{lat}}}), \qquad r \in \{1, \cdots, C\}.$$

Memory μ is assigned a cluster index $a(\mu)$, and its latent key is

$$u_\mu = \frac{\sqrt{1-\delta}\, c_{a(\mu)} + \sqrt{\delta}\, \epsilon_\mu}{\|\sqrt{1-\delta}\, c_{a(\mu)} + \sqrt{\delta}\, \epsilon_\mu\|_2}, \qquad \epsilon_\mu \sim \mathcal{N}(0, \mathbb{1}_{d_{\text{lat}}}).$$

We choose $\delta = 0.12$ to control the latent key variance. The observed query and keys are then

$$q^{\text{in}} = A_Q u_Q + \sigma_{\text{obs}} \eta_Q, \qquad k_\mu^{\text{in}} = A_K u_\mu + \sigma_K \eta_\mu,$$

where $A_Q, A_K \in \mathbb{R}^{d_{\text{in}} \times d_{\text{lat}}}$ with the orthogonalization $A_Q^\top A_Q = A_K^\top A_K = \mathbb{1}_{\text{lat}}$, $\eta_Q, \eta_\mu \sim \mathcal{N}(0, \mathbb{1}_{\text{in}})$, $\sigma_{\text{obs}} = 0.03$, and $\sigma_K = 0.05$. Here we divide u_Q generation into two cases, the selective retrieval case and the uniform-aggregation case.

1. **Selective retrieval.** In this case, u_Q retrieves one particular memory. Consider that we sample a target memory index t from $\{1, ..., M\}$. Therefore, we denote u_t as the latent key located at the sampled target slot t. We then generate the latent query by adding noise

   $$u_Q = \frac{u_t + \sigma_Q \epsilon_Q}{\|u_t + \sigma_Q \epsilon_Q\|_2}, \qquad \epsilon_Q \sim \mathcal{N}(0, \mathbb{1}_{d_{\text{lat}}}).$$

   During training, the query-noise level cycles through $\sigma_Q \in \{0.05, 0.10, 0.15\}$ and the number of memories cycles through $M = \{8, 16, 32, 64\}$ simultaneously. During evaluation, the query noise is fixed at $\sigma_Q = 0.10$ while M is increasing.

2. **Uniform-aggregation.** In this case, we sample a target cluster $r^\star$. Then the latent query is

   $$u_Q = \frac{c_{r^\star} + \sigma_Q \epsilon_Q}{\|c_{r^\star} + \sigma_Q \epsilon_Q\|_2}, \qquad \epsilon_Q \sim \mathcal{N}(0, \mathbb{1}_{d_{\text{lat}}}).$$

   We denote the relevant set as $R = \{\mu : a(\mu) = r^\star\}$. Since there are $C = 4$ clusters, we choose $|R| = M/4$ in this case.

Transformer. Inside the Transformer, we apply one-layer multi-head cross-attention. Specifically, we conduct the first step via

$$q_h = W_h^Q \mathrm{LN}_Q\Big(W_Q^{\text{in}} q^{\text{in}} + r_Q\Big), \qquad k_{h,\mu} = W_h^K \mathrm{LN}_K\Big(W_K^{\text{in}} k_\mu^{\text{in}} + r_K\Big),$$

$$v_{h,\mu} = W_h^V \mathrm{LN}_V\Big(W_V^{\text{in}} v_\mu + r_V\Big),$$

where $W_Q^{\text{in}}, W_K^{\text{in}} \in \mathbb{R}^{d_{\text{mod}} \times d_{\text{in}}}$, $W_V^{\text{in}} \in \mathbb{R}^{d_{\text{mod}} \times d_v}$, $W_h^Q, W_h^K, W_h^V \in \mathbb{R}^{d_h \times d_{\text{mod}}}$, LN denotes the LayerNorm, and $h = \{1, \cdots, H\}$. Then the multi-head output is followed by two residual updates

$$x_1 = W_Q^{\text{in}} q^{\text{in}} + r_Q + W^O \mathrm{Concat}\Big(\sum_{\mu=1}^{M} p_{1,\mu} v_{1,\mu}, \cdots, \sum_{\mu=1}^{M} p_{H,\mu} v_{H,\mu}\Big) \in \mathbb{R}^{d_{\text{mod}}},$$

and

$$x_2 = x_1 + \mathrm{FFN}\Big(\mathrm{LN}_{\text{FFN}}(x_1)\Big) \in \mathbb{R}^{d_{\text{mod}}},$$

where $W^O \in \mathbb{R}^{d_{\text{mod}} \times d_{\text{mod}}}$, $p_{h,\mu}$ is the attention with the form Proposition 3.3, and $\mathrm{FFN}(x) := W_2^{\text{FFN}} \mathrm{GELU}(W_1^{\text{FFN}} x)$. Finally, the final predicted output is

$$\widehat{y} = W^{\text{out}} \mathrm{LN}_{\text{final}}(x_2) \in \mathbb{R}^{d_v},$$

where $W^{\text{out}} \in \mathbb{R}^{d_v \times d_{\text{mod}}}$. We choose the architectural hyperparameters as

- Model dimension $d_{\text{mod}} = d_h \times H = 96$.
- 4 attention heads ($H = 4$) with head dimension $d_h = 24$.
- FFN with hidden dimension 192, i.e., $W_1^{\text{FFN}} \in \mathbb{R}^{192 \times 96}$, and $W_2^{\text{FFN}} \in \mathbb{R}^{96 \times 192}$.

In the Training stage, we choose 2500 AdamW steps, batch size $B = 64$, learning rate $\eta = 3 \times 10^{-4}$, weight decay 10^{-4}, MSE loss, $\beta = 1$, $M \in \{8, 16, 32, 64\}$. Specifically, the MSE we evaluate is

$$\text{MSE} = \frac{1}{B} \sum_{b=1}^{B} \Big\| \widehat{y}^{(b)} - y_{\text{target}}^{(b)} \Big\|_2^2,$$

where $y_{\text{target}}^{(b)} = y_{t_b}$ for the selective retrieval case, and $y_{\text{target}}^{(b)} = \sum_{\mu \in R_b} y_\mu^{(b)} / |R_b|$ for the uniform-aggregation case.

In inference stage, we fix $\beta = 1$ and the learned parameters, e.g., $W_h^Q, W^O, r_Q, \cdots$. Then we increase the memory load from $M = 8$ to $M = 4096$ to measure how good the prediction $\widehat{y}$ is.

Observables. We measure

- **Selective retrieval: Value top-1 accuracy.** We define the value top-1 accuracy as

$$A_{\text{value}} := \frac{1}{B} \sum_{b=1}^{B} \mathbb{1}\Big\{ \arg\min_{\mu} \big\| \widehat{y}^{(b)} - y_\mu^{(b)} \big\|_2^2 = t_b \Big\}.$$

  This measures end-to-end success, rather than simply measures whether attention puts its largest weight on the correct key.

- **Selective retrieval: Support size (IPR).** As introduced in Section G.2, we measure the support via the inverse partition ratio (IPR). However, since we consider batch size and multi-head here,

we further average over them, i.e.,

$$K_{\text{eff}}^{\text{IPR}} := \frac{1}{HB} \sum_{b=1}^{B} \sum_{h=1}^{H} \frac{1}{\sum_{\mu=1}^{M} p_{b,h,\mu}^2}.$$

- **Uniform-aggregation: Error/zero-output baseline.** The target is $y_{\text{target}} = \sum_{\mu \in R} y_\mu / |R|$, where $y_\mu \sim \mathcal{N}(0, \mathbb{1}_{d_v})$. Hence, the zero-output baseline is $E_0 := \mathbb{E}\|y_{\text{target}}\|_2^2 = d_v/|R|$. In this way, we define the error/zero-output baseline E_{rel} as

$$E_{\text{rel}} := \frac{\text{MSE}}{E_0} = \frac{|R|}{Bd_v} \sum_{b=1}^{B} \left\| \widehat{y}^{(b)} - \frac{1}{|R_b|} \sum_{\mu \in R_b} y_\mu^{(b)} \right\|_2^2.$$

- **Uniform-Aggregation: IPR / $|R|$.** We measure the normalized IPR with the size of the target set

$$\rho_{\text{IPR}} := \frac{K_{\text{eff}}^{\text{IPR}}}{|R|} = \frac{1}{HB|R|} \sum_{b=1}^{B} \sum_{h=1}^{H} \frac{1}{\sum_{\mu=1}^{M} p_{b,h,\mu}^2}.$$

Results. We present the experiments as follows.

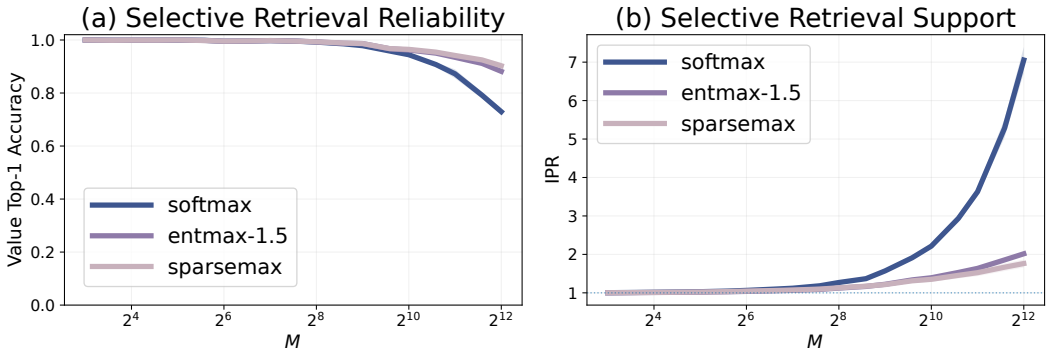


Figure 14: **Capacity-reliability trade-off for selective retrieval.** We evaluate the trained Transformer at fixed $\beta_{\text{train}} = 1$ while increasing M. **(a).** We measure the value top-1 accuracy to characterize the retrieval reliability. Sparsemax retain the highest accuracy. Furthermore, their accuracy decreases as M grows, indicating the capacity-reliability trade-off. **(b).** Softmax support remains full, while sparsemax use the least supports and is the most support-efficient.

From Figure 14 (a), the sparsemax arrives the best value top-1 accuracy, while the softmax arrives the worst one. Furthermore, both of their accuracy decrease as M increases. This indicates that high memory load leads to lower reliability, and reveals the capacity-reliability trade-off. From Figure 14 (b), the support of softmax remains full, while which of sparsemax is low. This suggests that sparsemax is support-efficient. Those plots conclude that for a selective task, sparsemax has not only high-accuracy but also high support-efficiency.

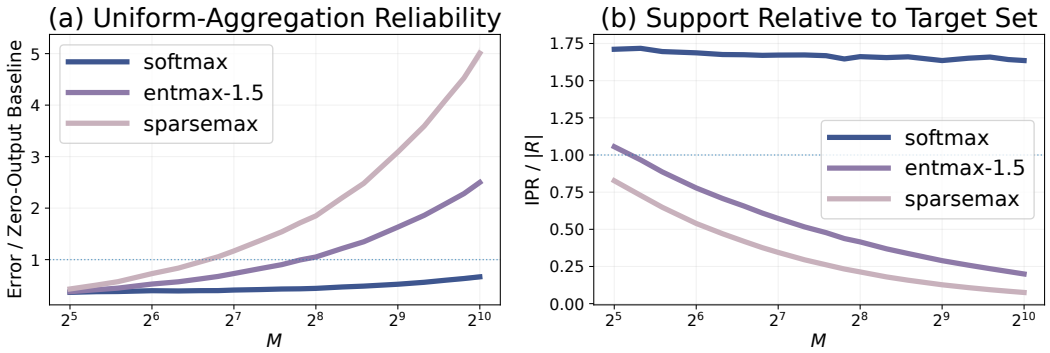


Figure 15: **Capacity-reliability trade-off for uniform-aggregated retrieval.** We evaluate the trained Transformer at fixed $\beta_{\text{train}} = 1$ while increasing M. **(a).** We measure the ratio of the output error and the zero-output baseline to characterize the retrieval reliability. Sparsemax exhibits the worst reliability. Furthermore, their retrieval reliabilities become worse as M grows, indicating the capacity-reliability trade-off. **(b).** Softmax support remains full, while sparsemax use the least supports and is the most support-efficient.

Compared to Figure 14, Figure 15 (a) shows that softmax arrives the highest reliability among three attentions and the capacity-reliability trade-off as well. From Figure 15 (b), we still observe that the sparsemax uses the least aggregated support size, and hence the most support-efficient. In conclusion, although a large α leads to support-efficiency, the reliability depends on the task: sparsemax predicts a selective retrieval the best, while softmax predicts a uniform-aggregation retrieval the best.

H.2 Support, Temperature, and Reliability

In this section, we investigate the temperature effect on effective support size (IPR) and retrieval reliability in a realistic Transformer architecture numerically. We compare three attention maps, $\alpha \in \{1, 1.5, 2\}$ (softmax, entmax-1.5, and sparsemax), respectively.

Setup. We adopt the same setup and Transformer architecture as in Section H.1, i.e., one-layer four-head cross-attention Transformer with MSE training objective. The only different setup is that we isolate β by fixing M at $M = 128$ with query-noise level $\sigma_Q = 0.10$ in the inference stage. We vary the inverse-temperature over $\beta \in \{2^{-2}, 2^{-1}, 2^0, \cdots, 2^9\}$. Furthermore, we compare two protocals:

- **Frozen-β.** In this case, we first train each model at $\beta_{\text{train}} = 1$. After training, we freeze every learned parameter (W_h^Q, W^O, r_Q, and so forth). At inference, we observe how reliable the output performance is at each $\beta \in \{2^{-2}, 2^{-1}, 2^0, \cdots, 2^9\}$.
- **Matched-β.** In this case, we train a separate Transformer for every β. That is, for different β, we have different trained parameters W_h^Q, W^O, r_Q, and so forth.

Observables. We adopt the same observables as which in Section H.1.

Results. We present the experiments as follows.

From Figure 16 (a) and (b), we observe that sparsemax has highest accuracy and support-efficiency for frozen-β case in a selective retrieval task. Furthermore, both of the accuracy and IPR saturates at high β, arriving a single-pattern reliable retrieval for sparsemax. For softmax, the support is full, but the IPR shows that its retrieval becomes strongly concentrated. For the uniform-aggregation task, Figure 16 (c) and (d) show that all three frozen models perform far worse than the zero-output baseline in the high-β regime. That is, since we train the parameters only at $\beta = 1$, the Transformer exhibits greater selectivity, but creates a mismatch for multi-memory aggregation.

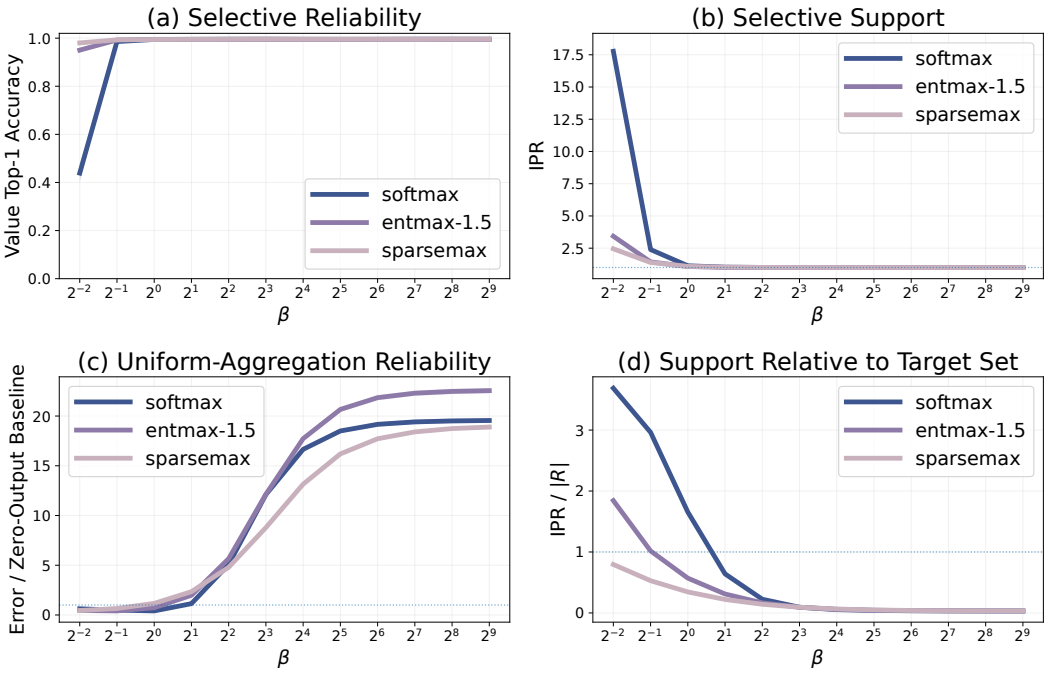


Figure 16: **Temperature effect on frozen-parameter Transformer.** We train the Transformer parameters at $\beta_{\text{train}} = 1$, then we fix those parameters, set $M = 128$ and vary β during inference. **(a).** The three attention families both exhibit high accuracy for a selective retrieval task. **(b).** For selective retrieval, the sparsemax uses the least effective support size (IPR), which shows the support-efficiency and reliability of high-β sparsemax. **(c).** For a uniform-aggregation task, the three attention families have worse error relative to the zero-output baseline as β increases. **(d).** The IPR relative to target-set size decline as β grows, which indicates that the retrieval becomes selective. This leads to a mismatch for multi-memory aggregation.

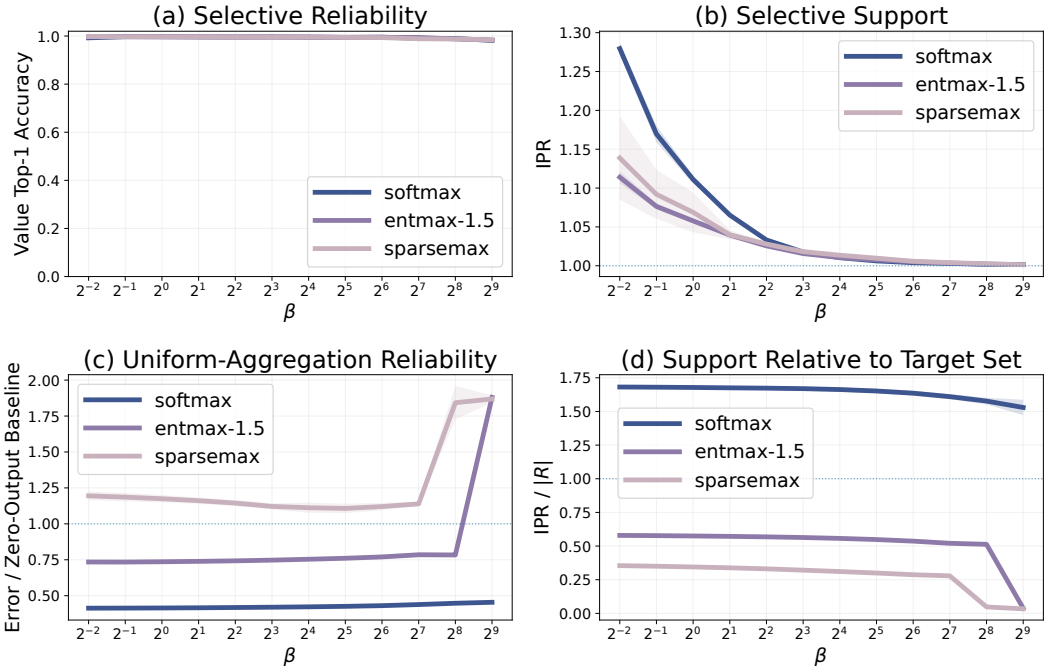


Figure 17: **Temperature effect on matched-parameter Transformer.** We train the Transformer parameters at each β, then we evaluate with $M = 128$ during inference. **(a).** All three attention families maintain high accuracy in a selective retrieval task. **(b).** Sparsemax and entmax-1.5 are support-efficient in a selective retrieval task. **(c).** Softmax has the highest reliability among the three in a uniform-aggregation task. **(d).** Softmax uses numerous memories, which indicates a poor support-efficiency.

From Figure 17 (a) and (b), all three attention families maintain perfect accuracy in a selective retrieval task. Furthermore, sparsemax and entmax-1.5 are more support-efficient that softmax. For the uniform-aggregation task. Figure 17 (c) and (d) show that softmax achieves the highest reliability but its effective support remains broader than the relevant set. Again, this indicates that high β increases the selectivity, but achieves poor performance for an aggregation task.

H.3 Numerical Validation of Coexistence and Spinodal Formula

In this section, we validate the coexistence formula in Proposition 4.1 and the spinodal formula in Proposition 4.2 for $\alpha = 2$ numerically.

Setup. We test the special two-key formulas under the uniform distribution and well-separation assumptions of Proposition 4.1 and Proposition 4.2. Then we construct a second dataset to test the full spinodal formula on larger supports.

Data. We construct two experiments as follows.

- **Two-key systems.** We consider two unit-norm keys $\widetilde{k}_1$ and $\widetilde{k}_2$ satisfying $\widetilde{k}_1^\top \widetilde{k}_2 = \epsilon$. Therefore, we set

$$\widetilde{k}_1 = e_1, \qquad \widetilde{k}_2 = \epsilon e_1 + \sqrt{1-\epsilon^2}\, e_2$$

  Then we apply a random orthogonal transformation Q_r to both keys

$$k_1 = Q_r \widetilde{k}_1, \qquad k_2 = Q_r \widetilde{k}_2,$$

  where the Gram matrix is preserved under this rotation, but removes the coordinate dependence. We choose the state dimension as $d = 16$, correlation as $\epsilon \in \{0, 0.05, 0.10 \cdot \cdot, 0.85\}$ and use 5 independent rotation. Hence, this gives $5 \times 18 = 90$ two-key systems.

- **General k-support systems.** In this experiment, we use $k \in \{3, 4, 5, 6, 8\}$ each with 10 independently generated systems and ambient dimension $d = 20$. For each key index $\mu \in \{0, \cdots, k-1\}$, we construct the feature vector

$$\widetilde{k}_\mu = \Big(b, \Big\{a_\ell \cos\Big(\frac{2\pi\ell\mu}{k}\Big), a_\ell \sin\Big(\frac{2\pi\ell\mu}{k}\Big)\Big\}_{\ell=1}^{m}\Big), \qquad m = \lfloor \frac{k-1}{2} \rfloor,$$

  where we sample $b \sim \mathrm{Unif}(1.5, 3.0)$, $a_\ell \sim \mathrm{Unif}(0.15, 0.80)$ independently. Then we normalize very vector and apply a random orthogonal rotation to k_μ.

Observables. We measure

- **Two-key: Special coexistence formula.** Under a uniform distribution and well-separation condition, the coexistence point occurs at

$$\beta_{\text{coexist}} = \frac{2}{1-\epsilon}.$$

  This is the special colsed-form result of Proposition 4.1.

- **Two-key: Numerical energy crossing.** We define the free energy difference

$$\Delta\Gamma(\beta) = \Gamma_{\alpha=2}^{(\text{mod})}(x_{\text{loc}}; \beta) - \Gamma_{\alpha=2}^{(\text{mod})}(x_{\text{mix}}; \beta),$$

  where $x_{\text{loc}} := k_1$ and $x_{\text{mix}} := (k_1 + k_2)/2$. Then we compute the coexistence point from $\Delta\Gamma(\widehat{\beta}_{\text{coexist}}) = 0$ numerically.

- **Two-key: Relative error.** We define the coexistence relative error as

$$E_{\text{coexist}} := \frac{\left|\widehat{\beta}_{\text{coexist}} - \beta_{\text{coexist}}\right|}{\beta_{\text{coexist}}}.$$

For the spinodal formula, we have $\beta_{\text{sp}} = 2/(1-\epsilon)$ in the two-key systems. We obtain the empirical spinodal point by solving $\widehat{c}_h(\widehat{\beta}_{\text{sp}}) = 0$ numerically, where

$$\widehat{c}_h(\beta) = \frac{\Gamma^{(\text{mod})}_{\alpha=2}(x^\star + hu;\beta) + \Gamma^{(\text{mod})}_{\alpha=2}(x^\star - hu;\beta) - 2\Gamma^{(\text{mod})}_{\alpha=2}(x^\star;\beta)}{h^2},$$

$h = 2 \times 10^{-5}$, and $u := (k_1 - k_2)/\|k_1 - k_2\|_2$. Then the spinodal relative error is defined as

$$E_{\text{sp}} := \frac{\left|\widehat{\beta}_{\text{sp}} - \beta_{\text{sp}}\right|}{\beta_{\text{sp}}}.$$

- **General k-support: Spinodal formula.** From Proposition 4.2, the predicted critical spinodal point is

$$\beta_{\text{sp}} = \frac{2}{\lambda_{\max}(PGP)},$$

where $P_{\mu\nu} = \mathbb{1}_{\{\mu,\nu\}} - k^{-1}$, and $G_{\mu\nu} = k_\mu^\top k_\nu / d$. We compute the empirical spinodal point $\widehat{\beta}_{\text{sp}}$ from the equation $\widehat{c}_h(\widehat{\beta}_{\text{sp}}) = 0$ numerically.

Results. We present the experiments as follows.

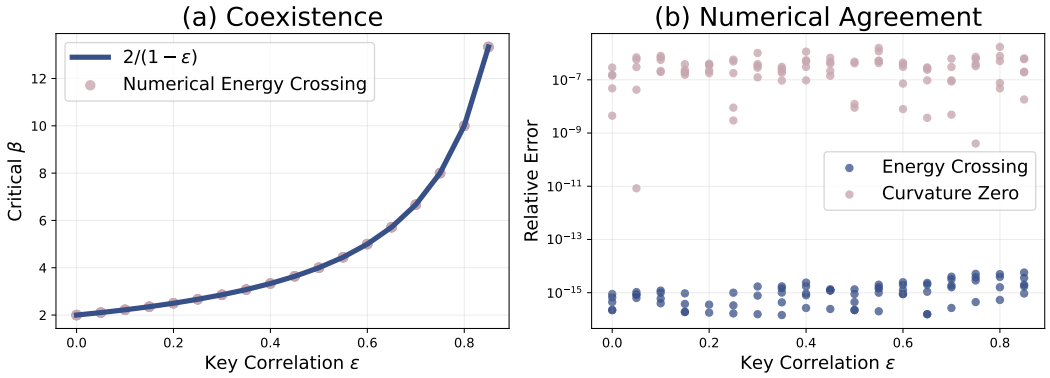


Figure 18: **Two-key coexistence and spinodal validation.** We consider two unit keys with correlation ϵ. **(a).** The solid blue curve presents $\beta_{\text{coexist}} = 2/(1-\epsilon)$. The numerical points are the root of the free energy difference $\Delta\Gamma(\widehat{\beta}_{\text{coexist}}) = 0$. The numerical energy crossings coincide with the theoretical curve over the full correlation range. **(b).** We plot the relative errors of energy crossing (coexistence) and curvature zero (spinodal). The low values validate the theoretical fromulas in Proposition 4.1 and Proposition 4.2.

From Figure 18 (a), the blue curve represents the theoretical formula $\beta_{\text{coexist}} = 2/(1-\epsilon)$. The numerical energy crossing points lie on the theoretical curve. This validates the special coexistence formula. For Figure 18 (b), we plot the relative errors of energy crossing (coexistence) and curvature zero (spinodal). The maximal relative error of the coexistence formula is 5.73×10^{-15}, and which of the spinodal formula is 1.72×10^{-6}. They both validate the theoretical formulas in Proposition 4.1 and Proposition 4.2 under the uniform distribution and well-separation conditions.

In Figure 19, we compare the theoretical spinodal formula Proposition 4.2 and the empirical spinodal values. Across all 50 general k-support systems, the predicted β_{sp} aligns with the empirical one closely, with the maximal relative error 1.92×10^{-6}. Hence, the agreement supports not only the

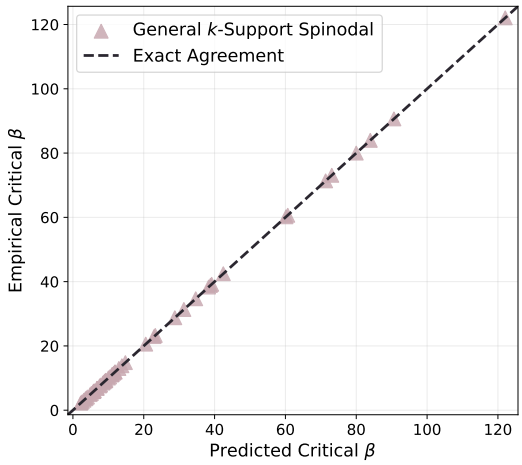


Figure 19: **General k-support spinodal validation.** We test the general k-support spinodal formula $\beta_{\text{jump}} = 2/\lambda_{\max}(PGP)$. We consider 50 systems with unit keys. The numerical points agree with the theoretical values, which validates Proposition 4.2.

qualitative existence of the transitions, but also the quantitative location of the predicted critical inverse-temperatures.