The Self-Consistent Theory of Neural Network Moments

Anonymous authors
Paper under double-blind review

Abstract

This paper establishes a rigorous mathematical foundation for the statistical behavior of neural network parameter and gradient moments through self-consistent equations. We prove that the logarithmic moments exhibit a universal asymptotic decomposition governed by extremal statistics. This framework is extended to construct a joint partition function that unifies parameter and gradient statistics, revealing a topological phase distinction between states of correlated and uncorrelated extrema. The theory provides exact microscopic guarantees for finite networks while capturing emergent scaling behavior in large-scale systems.

1 Introduction

The statistical properties of a neural network's weights and gradients are fundamental to its performance Neyshabur et al. (2017). While much of existing theory relies on infinite-width limits, empirical work on practical, finite-sized networks reveals the dominance of heavy-tailed distributions in learning dynamics Gürbüzbalaban et al. (2021). To bridge this gap, we introduce a deterministic framework, grounded in large deviations theory, for analyzing the exact moment statistics of any finite network.

Our primary contribution is a set of self-consistent equations governing the evolution of parameter and gradient moments. We prove that the logarithm of high-order moments follows a universal asymptotic form, which decomposes the network's statistics into contributions from its single largest value (the extremum), its multiplicity, and the spectrum of all other values. This provides a new and exact tool for structural analysis.

To quantify the statistical dependence between a parameter's magnitude and its gradient, we define a **coupling term** via a joint partition function. The asymptotic behavior of this term reveals two distinct learning phases: an **ordered phase**, where large parameters align with large gradients, and a **disordered phase**, where they do not. This coupling offers a tractable alternative to complex information-theoretic measures, which often face formal limitations in practice Tishby & Zaslavsky (2017); McAllester & Stratos (2020).

We develop this framework, validate it experimentally, and apply it to interpret phenomena such as grokking Power et al. (2022) and catastrophic forgetting Kirkpatrick et al. (2017). Our work demonstrates its utility in complementing recent progress on high-dimensional dynamics and scaling laws, offering a new lens through which to understand deep learning.

2 Theoretical Framework

2.1 Parameter Moments

Definition 2.1 (Absolute Parameter Moments)**.** Let $\Theta = \{\theta_1, \theta_2, \ldots, \theta_n\}$ denote the complete parameter set of a neural network, where $n = |\Theta| < \infty$. The k-th order absolute moment is defined as:

$$M(k) := \frac{1}{n} \sum_{i=1}^{n} |\theta_i|^k, \quad k \geq 0 \tag{1}$$

with the convention that $0^0 = 1$ and $0^k = 0$ for $k > 0$.

Remark. This definition naturally handles zero-valued parameters: they contribute nothing to moments of order $k > 0$ and correspond to a Dirac mass at $\lambda = \infty$ in the spectral representation, which does not affect the Laplace transform for any finite k. The convention $0^0 = 1$ ensures proper normalization $M(0) = 1$.

2.2 Existence of Moment Exponents

Theorem 2.2 (Existence and Explicit Value of Moment Exponents)**.** *For any finite-parameter neural network, the limit:*

$$\beta := \lim_{k\to\infty} \frac{\log M(k)}{k} \tag{2}$$

exists, is finite, and equals:

$$\beta = \log\Big(\max_{1\le i\le n} |\theta_i|\Big) = \sup_{k>0} \frac{\log M(k)}{k} \tag{3}$$

Proof. The proof is provided in Appendix A.1. □

Theorem 2.3 (Remainder Convergence)**.** *The limit:*

$$R^* := \lim_{k\to\infty} \big(\log M(k) - \beta k\big) \tag{4}$$

exists, is finite, and equals:

$$R^* = \log \frac{m}{n} \tag{5}$$

where m denotes the multiplicity of parameters with maximal modulus.

Proof. The proof follows directly from the derivation in Appendix A.1. □

2.3 Self-Consistent Equation Formulation

The exact asymptotic decomposition:

$$\log M(k) = \beta k + R^* + \Delta(k), \quad k \to \infty \tag{6}$$

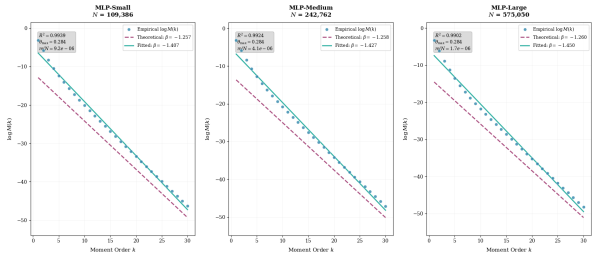
can be refined through spectral analysis. Define the decay rates $\lambda_i := \log(\theta_{\max}/|\theta_i|) > 0$ for $|\theta_i| < \theta_{\max}$. Then:

$$\Delta(k) = \log\left[1 + \sum_{j=1}^{n-m} w_j e^{-\lambda_j k}\right] \tag{7}$$

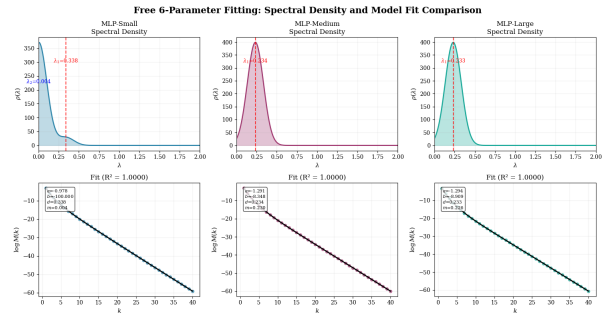
In the thermodynamic limit $n \to \infty$, this converges to (see Appendix A.2 for a rigorous proof under weakened regularity conditions):

$$\Delta(k) \to \log\left[1 + \int_0^\infty \rho(\lambda) e^{-\lambda k} d\lambda\right] \tag{8}$$

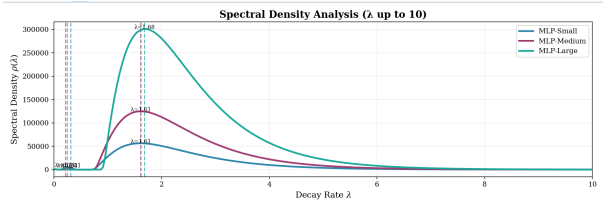
where $\rho(\lambda)$ is the spectral density satisfying $\int_0^\infty \rho(\lambda) d\lambda = \frac{n-m}{n}$.



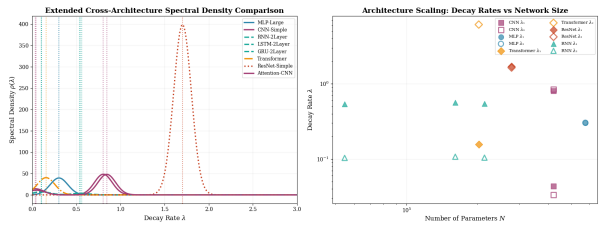
(a) Moment decomposition for MLP architectures.



(b) Bi-exponential fitting of residual terms.



(c) Spectral density reconstruction from moments.



(d) Architecture-dependent spectral characteristics.

2.4 Experimental Validation of Moment Decomposition

2.4.1 Phenomenological Model: Bi-Exponential Residual Structure

Our experiments consistently reveal that the residual term, $\Delta(k)$, exhibits a bi-exponential decay across diverse architectures. Based on this strong empirical observation, we propose a phenomenological model to explicitly capture the leading-order behavior of this residual term. The goal is not to derive this form from first principles, but to construct a minimal model that effectively describes the observed data.

We propose the functional form:

$$\Delta(k) \approx \log\left[1 + A_1 e^{-\lambda_1 k} + A_2 e^{-\lambda_2 k}\right] \tag{9}$$

This form is motivated by its direct interpretation within our framework: it corresponds to a spectral density $\rho(\lambda)$ whose dominant features can be approximated by two discrete modes. This suggests that the vast number of non-extremal parameters tend to organize into distinct statistical ensembles, each with a characteristic decay rate, λ_1 and λ_2.

Linearization and Spectral Interpretation For large k, where the exponential terms are small, the logarithmic function can be linearly approximated via a first-order Taylor expansion, $\log(1+x) \approx x$. This is not just a mathematical convenience; it provides a powerful tool for spectral interpretation. Applying this linearization to our model yields:

$$\log M(k) \approx \beta k + R^* + A_1 e^{-\lambda_1 k} + A_2 e^{-\lambda_2 k} \tag{10}$$

This approximation is empirically justified, as the condition $A_1 e^{-\lambda_1 k} + A_2 e^{-\lambda_2 k} \ll 1$ is validated by our moment analysis for sufficiently large k. The crucial insight here is that the linearized, empirically-fitted model directly reveals the structure of the underlying effective spectral density:

$$\rho_{eff}(\lambda) \approx A_1 \delta(\lambda - \lambda_1) + A_2 \delta(\lambda - \lambda_2) \tag{11}$$

This effective density, composed of two Dirac delta functions, should be understood as a simplified representation—a "two-peak" approximation—of what is likely a complex, continuous background spectrum. It successfully captures the dominant decay modes that govern the residual term's behavior at large k.

3 Gradient Moments and Statistical Isomorphism

3.1 Gradient Moment Theory

Definition 3.1 (Absolute Gradient Moments)**.** Let $\mathcal{G} = \{g_1, g_2, \ldots, g_n\}$ denote the gradient set corresponding to parameters. The l-th order gradient absolute moment is:

$$G(l) := \frac{1}{n} \sum_{i=1}^{n} |g_i|^l, \quad l \geq 0 \tag{12}$$

Theorem 3.2 (Existence and Explicit Value of Gradient Moment Exponents)**.** *For any finite-parameter neural network under gradient-based training, the limit:*

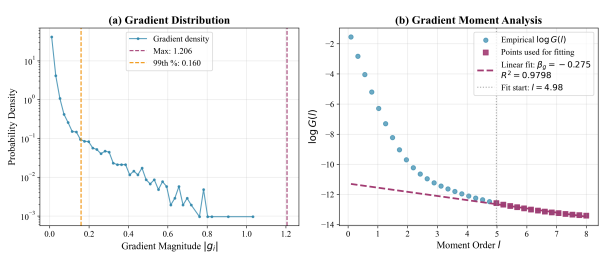
$$\beta_g := \lim_{l \to \infty} \frac{\log G(l)}{l} \tag{13}$$

exists, is finite, and equals:

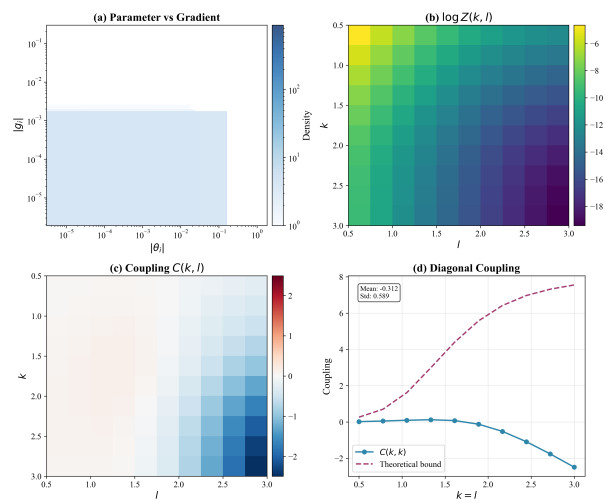
$$\beta_g = \log\Big(\max_{1 \leq i \leq n} |g_i|\Big) = \sup_{l > 0} \frac{\log G(l)}{l} \tag{14}$$

Proof. The proof is analogous to that of Theorem 2.2 and is provided in Appendix **??**. □

Remark on Applicability. The gradient moment decomposition assumes regularity conditions (non-vanishing spectral gap and log-integrability) that typically hold in quasi-static training phases. Transient violations may occur during early training or in architectures with strong symmetries; see Appendix C for a complete characterization of these non-standard cases and their diagnostic value.



(a) Gradient moment decomposition



(b) Parameter-gradient coupling structure

Figure 2: Gradient moment statistics and coupling behavior. Deviations from predicted asymptotics may indicate regularity violations; see Appendix C.

4 Joint Partition Function: Theory of Bounded Coupling

4.1 Joint Moments and Decomposition

Definition 4.1 (Joint Moments)**.** For orders $k, l \geq 0$, the joint moment is defined as:

$$Z(k, l) := \frac{1}{n} \sum_{i=1}^{n} |\theta_i|^k |g_i|^l. \tag{15}$$

Lemma 4.2 (Exact Decomposition of Logarithmic Joint Moments)**.** *Let* $\mathcal{C}(k,l) := \log Z(k,l) - \log M(k) - \log G(l)$ *be the pure coupling term. Then:*

$$\log Z(k,l) = \log M(k) + \log G(l) + \mathcal{C}(k,l). \tag{16}$$

This decomposition separates the joint statistics into marginal contributions and a term that captures their interaction.

4.2 Global Upper Bound on Coupling Term

Theorem 4.3 (Cauchy-Schwarz Upper Bound)**.** *For any finite-parameter network, the coupling term satisfies:*

$$\mathcal{C}(k,l) \leq A(k) + B(l) \tag{17}$$

where

$$A(k) := \frac{1}{2}\log M(2k) - \log M(k), \tag{18}$$

$$B(l) := \frac{1}{2}\log G(2l) - \log G(l). \tag{19}$$

Proof. See Appendix A.3. □

Corollary 4.4 (Boundedness of Coupling)**.** *The coupling term is globally bounded from above:*

$$\mathcal{C}(k,l) \leq C_{\max} < \infty, \quad \forall k,l \geq 0. \tag{20}$$

Proof. See Appendix A.3. □

4.3 Non-existence of Universal Lower Bound

Theorem 4.5 (Absence of Universal Lower Bound)**.** *For any constant* $C_{\min} \in \mathbb{R}$ *and any network size* $n \geq 2$*, there exists a parameter-gradient configuration* $(\Theta, \mathcal{G})$ *such that* $\mathcal{C}(k,l) < C_{\min}$ *for some* $k,l \geq 0$*.*

Proof. See Appendix A.3 for a constructive proof. □

4.4 Asymptotic Analysis of Coupling

Theorem 4.6 (Fixed Order Asymptotics)**.** *For unilateral limits, we have:*

(i) For any fixed $k \geq 0$*:*

$$\lim_{l\to\infty} \mathcal{C}(k,l) = \log\left(\frac{1}{m_g}\sum_{i\in I_{\max}} |\theta_i|^k\right) - \log M(k).$$

(ii) For any fixed $l \geq 0$*:*

$$\lim_{k\to\infty} \mathcal{C}(k,l) = \log\left(\frac{1}{m}\sum_{i\in J_{\max}} |g_i|^l\right) - \log G(l).$$

Where $J_{\max} = \arg\max_i |\theta_i|$ *and* $I_{\max} = \arg\max_i |g_i|$*.*

Proof. The proof follows from applying the logic of Theorem 2.2 to the definition of $\mathcal{C}(k,l)$ in the specified limits. □

Theorem 4.7 (Diagonal Asymptotics)**.** *Let* $k,l \to \infty$ *with* $l/k \to \alpha \in (0,\infty)$*. The asymptotic behavior of the coupling term is determined by the intersection of the extremal sets,* $m_\cap = |J_{\max} \cap I_{\max}|$*.*

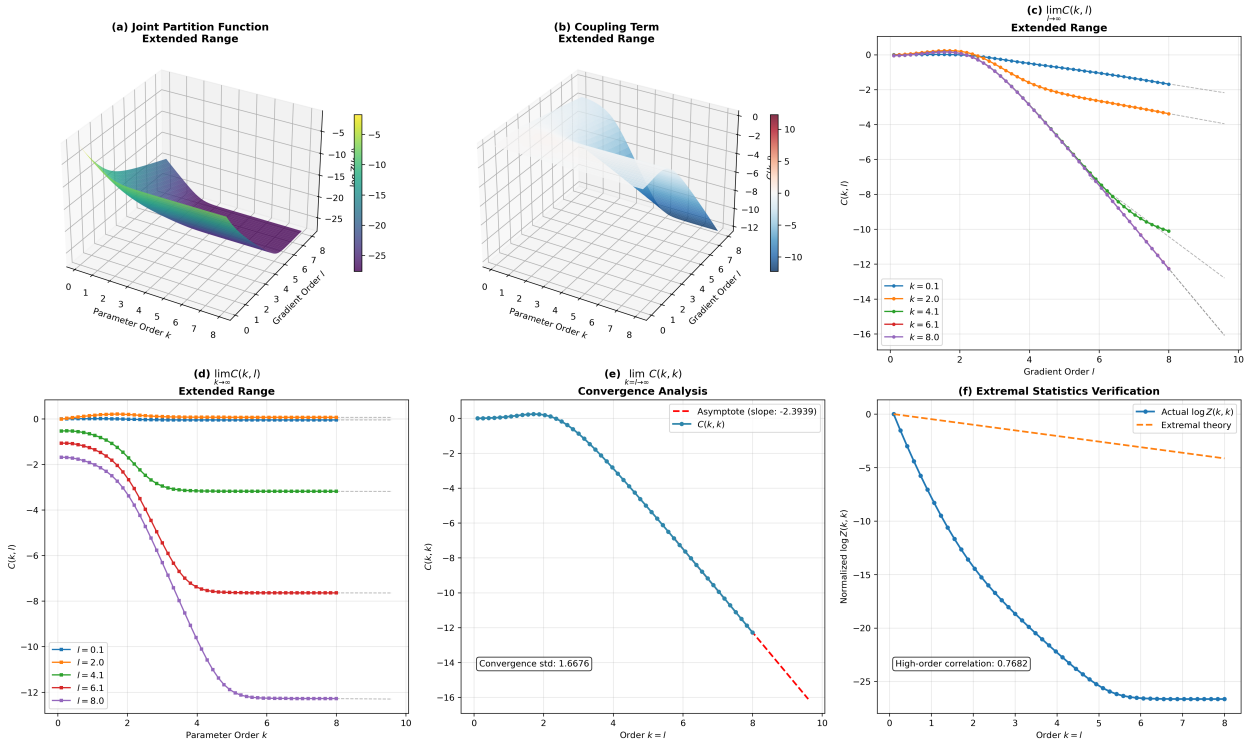


Figure 3: **Anisotropic Divergence in the Disordered Phase ($m_\cap = 0$).** The figure confirms that divergence in the disordered phase is directional (anisotropic). **(a, b)** The 3D surface of the coupling term $\mathcal{C}(k,l)$ visually confirms this, appearing bounded along the axes but plunging towards negative infinity along the diagonal. **(c, d)** Axial boundedness is confirmed by the unilateral limits, which converge to finite constants for any fixed order, despite showing transient downward trends. **(e)** In sharp contrast, the diagonal limit $\mathcal{C}(k,k)$ is clearly unbounded and follows a negative linear asymptote. **(f)** Throughout, the data respects a theoretical upper bound, confirming the model's mathematical consistency.

(i) ***Correlated Extrema*** *($m_\cap > 0$)**:** If the extremal sets intersect, the coupling term converges to a universal constant independent of α:*

$$\lim_{\substack{k,l\to\infty \\ l/k\to\alpha}} \mathcal{C}(k,l) = \log\left(\frac{n \cdot m_\cap}{m \cdot m_g}\right)$$

(ii) ***Disjoint Extrema*** *($m_\cap = 0$)**:** If the extremal sets are disjoint, the coupling term diverges to negative infinity:*

$$\lim_{\substack{k,l\to\infty \\ l/k\to\alpha}} \mathcal{C}(k,l) = -\infty$$

Proof. See Appendix A.4. □

Theorem 4.8 (Boundedness Condition and Lower Bound)**.** *Let $\Theta = \{\theta_i\}_{i=1}^n$ and $\mathcal{G} = \{g_i\}_{i=1}^n$ be finite, non-zero parameter-gradient sets. Define the extremal sets $J_{\max} = \arg\max_i |\theta_i|$ and $I_{\max} = \arg\max_i |g_i|$. Let this system have **non-degenerate spectral gaps**, meaning the maximum values are strictly greater than all others:*

$$\theta_{\max} > \sup_{j\notin J_{\max}} |\theta_j| \quad \text{and} \quad g_{\max} > \sup_{i\notin I_{\max}} |g_i|. \tag{21}$$

Under this condition, the coupling term

$$\mathcal{C}(k,l) = \log Z(k,l) - \log M(k) - \log G(l) \tag{22}$$

admits a finite lower bound for all $k, l \geq 0$ if and only if the extremal sets intersect, i.e.,

$$\boxed{m_\cap := |J_{\max} \cap I_{\max}| \geq 1.} \tag{23}$$

When this condition holds, a valid lower bound is given by $\log(m_\cap/n)$.

Proof. See Appendix A.4. □

4.5 Joint Self-Consistent Equation Formulation

This necessary and sufficient condition allows for a complete spectral decomposition of the coupling term.

Theorem 4.9 (Spectral Representation of Coupling Term)**.** *Let the spectral decay rates be $\lambda_i = \log(\theta_{\max}/|\theta_i|)$ and $\mu_i = \log(g_{\max}/|g_i|)$. The coupling term $\mathcal{C}(k,l)$ admits the exact spectral decomposition:*

$$\boxed{\begin{aligned}\mathcal{C}(k,l) = \quad & \underbrace{\log\frac{m_\cap n}{m m_g}}_{\textit{Topological Constant}} + \underbrace{\log\left[1 + \frac{1}{m_\cap}\sum_{i\notin K_{\max}} e^{-\lambda_i k - \mu_i l}\right]}_{\textit{Joint Spectral Correction}} \\ & - \underbrace{\log\left[1 + \frac{1}{m}\sum_{i\notin J_{\max}} e^{-\lambda_i k}\right]}_{\textit{Parameter Spectral Correction}} - \underbrace{\log\left[1 + \frac{1}{m_g}\sum_{i\notin I_{\max}} e^{-\mu_i l}\right]}_{\textit{Gradient Spectral Correction}}.\end{aligned}} \tag{24}$$

where $K_{\max} = J_{\max} \cap I_{\max}$. This form holds when $m_\cap \geq 1$.

5 Stability of Extremal Points Across Phases

The condition $m_\cap \geq 1$ from Theorem 4.8 is more than a mathematical curiosity; it marks a topological distinction between a "disordered" phase ($m_\cap = 0$, unbounded below) and an "ordered" phase ($m_\cap \geq 1$, bounded below). This distinction corresponds to a fundamental difference in the stability of extremal points under perturbations, such as those induced by training.

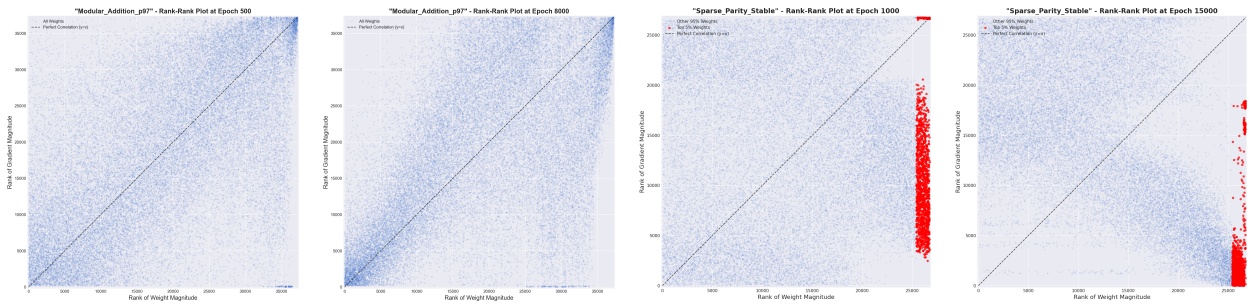
5.1 Unprotected Extremal Points in the Disordered Phase ($m_\cap = 0$)

In this phase, the set of largest parameters and the set of largest gradients are disjoint. This lack of alignment leads to fragility.

Proposition 5.1 (Instability of Extremal Points)**.** *When $m_\cap = 0$, for any $\epsilon > 0$ and any target value $C_{target} < 0$, there exists a small perturbation of the parameters and gradients that results in $\mathcal{C}(k, l) < C_{target}$ for some finite k, l.*

Proof. See Appendix A.5. □

Interpretation: In the disordered phase, the identities of the extremal parameters are not anchored to the learning signal (gradients). This can be thought of as a "liquid" state, where gradient descent can easily reassign which parameters become dominant. The lack of a lower bound on $\mathcal{C}(k, l)$ reflects that there is no "energy penalty" for decorrelating the parameter and gradient extrema.



(a) Ordered (Initial) (b) Ordered (Final) (c) Disordered (Initial) (d) Disordered (Final)

Figure 4: **Evolution of Topological Phases in Rank-Rank Space.** Plots of parameter magnitude rank (x-axis) vs. gradient magnitude rank (y-axis). **(a-b):** A successful run showing the evolution into an ordered phase, where ranks correlate along the diagonal, ensuring $m_\cap \geq 1$. **(c-d):** A failed run devolving into a disordered, anti-correlated state, where large weights get small gradients ($m_\cap = 0$).

5.2 Topologically Protected Extremal Points in the Ordered Phase ($m_\cap \geq 1$)

In this phase, at least one parameter is simultaneously extremal in both magnitude and gradient. This creates a form of topological protection.

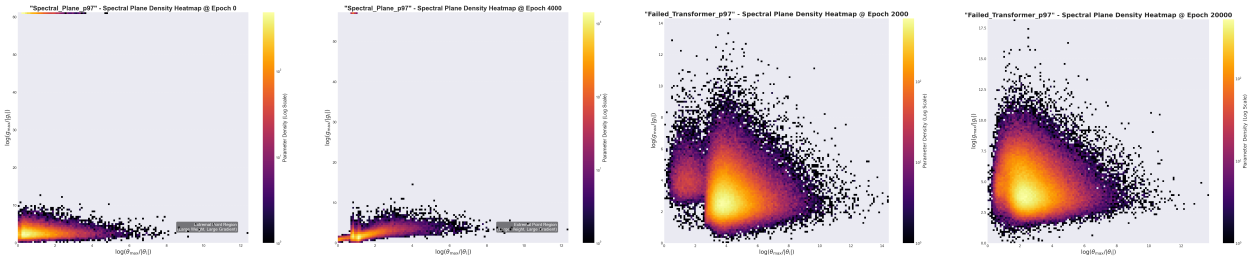
Proposition 5.2 (Rigidity of Extremal Points)**.** *When $m_\cap \geq 1$, the property of having overlapping extrema is robust against small continuous perturbations that preserve the maximal parameter and gradient values. The identity of the parameters forming this extremal core cannot change without crossing a phase transition.*

Proof. See Appendix A.5. □

Interpretation: The ordered phase behaves like a "solid" state where at least one extremal parameter is locked to an extremal gradient. The finite lower bound on the coupling term acts as a confining potential, preventing the system from decorrelating its most important parameters from the learning signal. This stability is crucial for forming robust representations.

5.3 Catastrophic Forgetting as Phase Reversal

The stability of the ordered phase ($m_\cap \geq 1$) is locally robust but globally fragile. A strong enough perturbation, such as training on a new, unrelated task, can shatter the alignment between parameters and gradients, inducing a phase transition from ordered to disordered ($m_\cap = 0$). This provides a topological explanation for catastrophic forgetting.



(a) Success (Early) (b) Success (Late) (c) Failure (Early) (d) Failure (Late)

Figure 5: **Grokking vs. Failure in the Spectral Plane.** The plane plots log-decay from the max parameter (x-axis) and max gradient (y-axis). The origin (0,0) represents the ideal state of high-magnitude, high-gradient parameters. **Successful grokking (a-b)** is marked by density condensing at the origin over time. **Failure (c-d)** is characterized by density pathologically avoiding the origin, indicating large weights receive no learning signal.

This stability is fundamentally linked to the **stability of the extremal sets** $(J_{\max}, I_{\max})$ against perturbations from the learning process. The degree of this stability is governed directly by the **spectral gaps** of the parameter and gradient distributions. A larger gap implies greater resilience, as a stronger perturbation is required to alter the membership of these extremal sets and risk a phase reversal. In the limit of maximum stability, where the spectral gap is maximized, the system approaches a state of **Neural Collapse**. This regime, where all class-relevant extremal features coalesce onto a maximally separated simplex structure, represents the most robust possible form of the ordered phase. It is, therefore, maximally resistant to catastrophic forgetting (a formal treatment of this connection is deferred to Appendix B).

To test this phase reversal hypothesis, we conducted a continual learning experiment on sequential MNIST (Task A: digits 0-4; Task B: digits 5-9). We compared a baseline with our theory-guided Elastic Weight Consolidation (EWC), which uses a penalty to elastically protect the extremal core stability of Task A.

Table 1: Final Accuracy after Sequential Training on MNIST

Strategy	**Final Task A Accuracy**	**Final Task B Accuracy**
Baseline (Fine-tuning)	0.3069	**0.9916**
Hard-Freeze Top-1000	**0.9889**	0.3364
EWC ($\lambda = 100$)	**0.6698**	**0.9831**

The results in Table 1 confirm the hypothesis. The baseline model undergoes a complete phase reversal, losing its ordered structure for Task A. In contrast, our EWC approach acts as a "confining potential" that preserves the extremal set stability (and thus $m_\cap \geq 1$), demonstrating that catastrophic forgetting can be understood and mitigated as a controllable topological phase transition.

6 Controlled Divergence and Quasi-Ordered Phases in Practical Networks

The sharp phase transition between ordered ($m_\cap \geq 1$) and disordered ($m_\cap = 0$) phases is an idealization that emerges in the strict thermodynamic limit. In practical, finite-sized networks, the system often resides in an **intermediate state of approximate alignment**, where the largest parameters and gradients are not perfectly coincident but rather approach each other asymptotically. Our continuous framework naturally captures this realistic scenario through the geometry of the joint spectral support near the origin.

Rather than a discrete topological switch, transitions in finite networks are characterized by the **distance** of the spectral support from the origin and its contact order. This leads to a spectrum of critical behaviors ranging from **quasi-ordered**(logarithmic divergence) to **disordered** (linear divergence), providing a quantitative diagnostic for the "health" of parameter-gradient alignment.

6.1 Approximate Alignment and Contact Stability

Definition 6.1 (Spectral Gap and Contact Regime)**.** Consider a network with joint spectral measure ν supported on $F \subset \mathbb{R}_+^2$. We define:

- **Spectral Gap:** $d_0 := \mathrm{dist}((0,0), F)$, measuring the minimal separation from perfect alignment.
- **Contact Regime:** If $d_0 = 0$ but $(0,0) \notin F$, the support *touches* the origin without containing it, defining a **quasi-ordered** phase.

When $d_0 > 0$, the system remains in a disordered phase with decoupled extrema. As training progresses, d_0 typically decreases, and the support may approach the origin with a characteristic **contact exponent** that quantifies the rate of alignment.

6.2 Critical Exponent and Asymptotic Divergence

The geometry of the support near the origin determines the asymptotic behavior of the coupling term. We characterize this geometry by the *contact function* and its scaling exponent.

Definition 6.2 (Contact Exponent)**.** Assume the support near the origin satisfies a power-law scaling:

$$\inf\{\lambda + \mu : (\lambda, \mu) \in F,\ \lambda + \mu \geq r\} = Cr^\alpha + o(r^\alpha), \quad r \to 0^+$$

where $\alpha > 0$ is the **contact exponent**. Smaller α indicates a "flatter" approach to the origin, while $\alpha \to \infty$ corresponds to sharp, pointwise contact.

Theorem 6.3 (Controlled Divergence in Quasi-Ordered Networks)**.** *Consider a network with approximate alignment characterized by spectral gap d_0 and contact exponent α (when $d_0 = 0$). The diagonal asymptotic behavior ($l = \alpha k$) of the coupling term is given by:*

$$\mathcal{C}(k,k) \sim \begin{cases} -\dfrac{2}{\alpha+1}\log k + O(1), & d_0 = 0 \quad \textit{(quasi-ordered)} \\ -d_0 k + O(\log k), & d_0 > 0 \quad \textit{(disordered)} \end{cases}$$

The first case reveals that even without perfect alignment ($p_{00} = 0$), a *quasi-ordered* network exhibits only **logarithmic divergence**, which is far milder than the linear divergence of the fully disordered phase. This captures the **progressive alignment** observed during training, particularly in phenomena like grokking where the model slowly transitions from disorder to order.

6.3 Training Quality Metric via Contact Exponent

The rate of divergence provides a quantitative, real-time measure of alignment quality that is directly computable during training.

Corollary 6.4 (Continuous Training Quality Metric)**.** *Define the **contact quality index**:*

$$Q_\alpha := \lim_{k\to\infty} \frac{-\mathcal{C}(k,k)}{\log k} \in [0, \infty]$$

For a network with contact exponent α, we have $Q_\alpha = \frac{2}{\alpha+1}$. Thus:

$$\begin{cases} Q_\alpha \approx 0 & \textit{indicates near-perfect alignment (α large)} \\ 0 < Q_\alpha < \infty & \textit{indicates quasi-ordered phase} \\ Q_\alpha = \infty & \textit{indicates disordered phase ($d_0 > 0$)} \end{cases}$$

In practice, one estimates $\hat{Q}_\alpha(K) = -\frac{\mathcal{C}(K,K)}{\log K}$ for moderate K (e.g., $K \in [10, 50]$). The evolution of $\hat{Q}_\alpha$ during training provides a direct diagnostic:

- **Decreasing trend**: Network is moving toward the ordered phase.
- **Stable low value**: Robust alignment achieved.
- **Sudden increase**: Phase reversal or catastrophic forgetting.

6.4 Practical Significance and Diagnostics

This continuous extension is significant for real-world networks:

Progressive Learning (Grokking) During grokking, the network initially stays in a quasi-ordered state ($0 < Q_\alpha < \infty$) for many epochs before transitioning to true order ($Q_\alpha \to 0$). The slow decrease of Q_α reflects the gradual compression of the contact exponent α.

Catastrophic Forgetting When switching tasks, the spectral support F may suddenly detach from the origin (d_0 jumps from 0 to > 0), causing Q_α to spike. Monitoring Q_α provides early warning of forgetting.

Architecture Design Architectures promoting large α values (e.g., weight sharing, skip connections) facilitate faster alignment. The contact exponent thus serves as a design principle for trainability.

Thus, even when perfect alignment is unattainable, the coupling term's behavior is not arbitrary but governed by predictable, geometry-dependent divergence laws. This provides a rigorous, computable tool for diagnosing and controlling the alignment health of neural networks in practice.

7 Discussion: Limitations and Future Directions

The framework presented in this paper offers a new lens through which to view the internal statistical mechanics of neural networks, focusing on the deterministic properties of finite-sized models rather than relying on idealized asymptotic limits. By analyzing the joint moments of parameters and gradients, we have uncovered a rich phase structure governed by the geometry of spectral support. However, like any foundational theory, its value is defined as much by the questions it answers as by the new ones it raises. Here, we honestly delineate the limitations of our current work, which in turn illuminate promising avenues for future research.

From Statics to Dynamics. Our analysis is primarily *static*. It provides a precise characterization of a network's state—be it ordered, quasi-ordered, or disordered—at a given instant. We have successfully used this to analyze equilibrium and quasi-equilibrium states. A natural and immediate extension is to build a full *dynamic* theory upon this static foundation. Key open questions include:

- What is the equation of motion governing the evolution of the spectral gap, $d_0(t)$, and the contact quality index, $Q_\alpha(t)$?
- Can we model the phase transition itself as a dynamic process, thereby explaining phenomena like the "critical slowing down" observed during grokking, where the system lingers in a quasi-ordered state before finally condensing?

Developing such a dynamic theory would transform our framework from a powerful diagnostic tool into a predictive model of the entire training trajectory.

The Microscopic Origin of Alignment. Our theory is currently phenomenological; it describes *that* a system can be in a state of alignment but does not fully explain the microscopic forces responsible for creating and maintaining it. We have characterized the stability of extremal points, but we have not derived the "restoring force" that pulls the system towards alignment. We conjecture that this force originates from the curvature of the loss landscape. A pivotal future direction is to connect our spectral plane coordinates (λ, μ) to the local geometry of the loss function, likely through the Hessian matrix. For instance, how does

the landscape curvature in the direction of a large-magnitude parameter ($\lambda \to 0$) relate to the magnitude of its gradient (μ)? Uncovering this relationship would bridge the gap between our macroscopic statistical picture and the microscopic geometry of optimization.

The Link to Generalization and Robustness. A central, motivating hypothesis of this work is that the "ordered phase" corresponds to better-generalizing and more robust solutions. Our analysis provides strong circumstantial evidence, particularly in the context of grokking and catastrophic forgetting. However, a large-scale, rigorous empirical study is required to solidify this claim. The contact quality index Q_α provides a concrete, computable metric for such an investigation. Future work should systematically test the conjecture that, for a given training loss, models with a lower Q_α exhibit superior out-of-distribution performance and enhanced resilience to adversarial attacks.

Concluding Vision. This work should be viewed not as a final theory, but as the **foundational layer** upon which a more complete, dynamic theory of deep learning can be built. By providing the essential language (phases, spectral support, contact exponent) and the necessary tools (spectral plane, coupling term, Q_α index), we hope to have opened a new avenue for understanding the emergent statistical structure of neural networks, shifting the focus from idealized limits to the precise, geometric realities of the models we use every day. This work is not a finish line, but a starting point for deeper understanding.

A Appendix: Proofs of Main Results

A.1 Proofs for Section 2 (Parameter Moments)

Proof of Theorem 2.2: Existence and Explicit Value of Moment Exponents. Let $f(k) = \log M(k) = \log\left(\frac{1}{n}\sum_{i=1}^{n}|\theta_i|^k\right)$. The function $\exp(x)$ is convex, and the composition of a convex function with an affine mapping is convex. Sums of convex functions are convex. Finally, $\log(x)$ is a concave, monotonically increasing function. The function $f(k)$ is the logarithm of a sum of exponentials, which is a log-sum-exp function. A more direct proof of convexity for $k \geq 0$ can be established via Hölder's inequality. For any $0 < t < 1$ and $k_1, k_2 \geq 0$, let $p = 1/t$ and $q = 1/(1-t)$.

$$\begin{aligned}
M(tk_1 + (1-t)k_2) &= \frac{1}{n}\sum_i |\theta_i|^{tk_1}|\theta_i|^{(1-t)k_2} \\
&\leq \frac{1}{n}\left(\sum_i (|\theta_i|^{tk_1})^p\right)^{1/p}\left(\sum_i (|\theta_i|^{(1-t)k_2})^q\right)^{1/q} \\
&= \frac{1}{n}\left(\sum_i |\theta_i|^{k_1}\right)^t\left(\sum_i |\theta_i|^{k_2}\right)^{1-t} \\
&= (M(k_1))^t(M(k_2))^{1-t}.
\end{aligned}$$

Taking the logarithm of both sides, we get:

$$f(tk_1 + (1-t)k_2) \leq tf(k_1) + (1-t)f(k_2),$$

which confirms that $f(k)$ is convex.

Since $f(k)$ is a convex function and $f(0) = \log M(0) = \log(1) = 0$, the sequence of slopes of the secant lines from the origin, $s_k = \frac{f(k)-f(0)}{k-0} = \frac{\log M(k)}{k}$, is non-decreasing for $k > 0$.

The sequence is also bounded above. Let $\theta_{\max} = \max_i |\theta_i|$. Then,

$$M(k) = \frac{1}{n}\sum_i |\theta_i|^k \leq \frac{1}{n}\sum_i \theta_{\max}^k = \theta_{\max}^k.$$

Taking the logarithm and dividing by k gives:

$$s_k = \frac{\log M(k)}{k} \leq \frac{\log\left(\theta_{\max}^k\right)}{k} = \log\theta_{\max}.$$

Since $\{s_k\}$ is a non-decreasing sequence that is bounded above, the Monotone Convergence Theorem guarantees that the limit $\beta = \lim_{k\to\infty} s_k$ exists and is equal to its supremum, $\sup_{k>0} s_k$.

To find its explicit value, let the maximum value $\theta_{\max}$ have multiplicity $m \geq 1$. We decompose $M(k)$:

$$M(k) = \frac{1}{n}\sum_{i:|\theta_i|=\theta_{\max}} |\theta_i|^k + \frac{1}{n}\sum_{i:|\theta_i|<\theta_{\max}} |\theta_i|^k = \frac{m}{n}\theta_{\max}^k + \frac{1}{n}\sum_{i:|\theta_i|<\theta_{\max}} |\theta_i|^k.$$

Factoring out the dominant term:

$$M(k) = \frac{m}{n}\theta_{\max}^k\left[1 + \frac{1}{m}\sum_{|\theta_i|<\theta_{\max}}\left(\frac{|\theta_i|}{\theta_{\max}}\right)^k\right].$$

Let the term in the square brackets be $(1+\delta(k))$. Since for every term in the sum, $|\theta_i|/\theta_{\max} < 1$, we have $\lim_{k\to\infty}(|\theta_i|/\theta_{\max})^k = 0$. As the sum is finite, $\lim_{k\to\infty}\delta(k) = 0$. So, $M(k) = \frac{m}{n}\theta_{\max}^k\big(1+o(1)\big)$.

Taking the logarithm and dividing by k:

$$\frac{\log M(k)}{k} = \frac{\log(m/n)}{k} + \log\theta_{\max} + \frac{\log(1+o(1))}{k}.$$

Taking the limit as $k \to \infty$, the first and third terms on the right-hand side go to zero, yielding $\beta = \log\theta_{\max}$. □

A.2 Proof of Spectral Measure Existence under Weakened Conditions

We provide a rigorous justification for the thermodynamic limit transition in equation equation 8 under minimal assumptions that accommodate practical neural networks, including those with quantized or sparse parameters.

Theorem A.1 (Existence of Limiting Spectral Measure)**.** *Let $\{\theta_i\}_{i=1}^n$ be i.i.d. random variables representing network parameters, with support in a compact interval $[0, \Theta_{\max}]$ where $\Theta_{\max} = \sup\{x : P(|\theta_i| \leq x) < 1\}$. Assume the integrability condition:*

$$\mathbb{E}\left[\left|\log \frac{\Theta_{\max}}{|\theta_i|}\right|\right] = \int_0^{\Theta_{\max}} \left|\log \frac{\Theta_{\max}}{x}\right| dF_\theta(x) < \infty, \tag{25}$$

where F_θ is the cumulative distribution function of $|\theta_i|$. Define the spectral variables $\lambda_i := \log(\Theta_{\max}/|\theta_i|) \in [0, \infty)$ and the empirical spectral measure:

$$\mu_n := \frac{1}{n} \sum_{i=1}^n \delta_{\lambda_i}. \tag{26}$$

Then:

(i) The sequence μ_n converges weakly almost surely to a probability measure μ on $[0, \infty)$.

(ii) The limit measure μ admits the decomposition:

$$\mu = p\delta_0 + \mu_{ac}, \quad \text{with } p := P(|\theta_i| = \Theta_{\max}), \tag{27}$$

where δ_0 is the Dirac mass at $\lambda = 0$ and μ_{ac} is absolutely continuous with respect to Lebesgue measure, possessing a density $\rho(\lambda)$ for $\lambda > 0$.

(iii) For any fixed $k \geq 0$, the Laplace transform converges:

$$\lim_{n\to\infty} \frac{1}{n} \sum_{i=1}^n e^{-\lambda_i k} = \int_0^\infty e^{-\lambda k} d\mu(\lambda) = p + \int_{0^+}^\infty \rho(\lambda) e^{-\lambda k} d\lambda. \tag{28}$$

(iv) Consequently, the residual term $\Delta(k)$ in equation equation 6 satisfies, as $n \to \infty$:

$$\Delta(k) \to \log\left[1 - p + \int_{0^+}^\infty \rho(\lambda) e^{-\lambda k} d\lambda\right]. \tag{29}$$

Proof. We proceed by establishing each claim in sequence.

1. Weak convergence of μ_n. The empirical measure μ_n is the pushforward of the empirical distribution of $\{|\theta_i|\}$ under the continuous transformation $T(x) = \log(\Theta_{\max}/x)$ for $x \in (0, \Theta_{\max}]$, with $T(0) = +\infty$ (a null set under our assumptions). By the strong law of large numbers for empirical measures (Varadarajan's theorem), since $\{\theta_i\}$ are i.i.d., we have:

$$\mu_n \xrightarrow{w} \mu \quad \text{a.s.}, \tag{30}$$

where μ is the pushforward of the law of $|\theta_i|$ under T. That is, for any Borel set $A \subseteq [0, \infty)$:

$$\mu(A) = P(\lambda_i \in A) = P\left(\log \frac{\Theta_{\max}}{|\theta_i|} \in A\right). \tag{31}$$

2. Decomposition of μ. The structure of μ follows directly from the distribution of $|\theta_i|$:

- If $p = P(|\theta_i| = \Theta_{\max}) > 0$, then $P(\lambda_i = 0) = p$, contributing the atomic part $p\delta_0$.
- For $\lambda > 0$, we have $P(\lambda_i \leq \lambda) = P(|\theta_i| \geq \Theta_{\max} e^{-\lambda})$. Since F_θ is differentiable almost everywhere (by Lebesgue's theorem), μ is absolutely continuous on $(0, \infty)$ with density:

$$\rho(\lambda) = -\frac{d}{d\lambda} P(|\theta_i| < \Theta_{\max} e^{-\lambda}) = \Theta_{\max} e^{-\lambda} f_\theta(\Theta_{\max} e^{-\lambda}), \tag{32}$$

  where f_θ is the density of $|\theta_i|$ (where it exists).

3. Convergence of Laplace transforms. Define $g_k(\lambda) = e^{-\lambda k}$ for fixed $k \geq 0$. The integrability condition equation 25 ensures that:

$$\sup_n \int_{[0,\infty)} |\lambda| \, d\mu_n(\lambda) = \frac{1}{n} \sum_{i=1}^{n} |\lambda_i| < \infty \quad \text{a.s.} \tag{33}$$

This uniform integrability, combined with weak convergence, implies convergence of the associated integrals for all bounded continuous functions. Since g_k is bounded and continuous on $[0, \infty)$ for any finite k, the continuous mapping theorem yields:

$$\int g_k \, d\mu_n = \frac{1}{n} \sum_{i=1}^{n} e^{-\lambda_i k} \xrightarrow{a.s.} \int g_k \, d\mu = \int_0^\infty e^{-\lambda k} d\mu(\lambda). \tag{34}$$

The decomposition of the limit integral follows directly from the structure of μ established in part (ii).

4. Connection to $\Delta(k)$. Recall the definition of the residual term from equation equation 6:

$$\Delta(k) = \log \left[1 + \frac{1}{m} \sum_{|\theta_i| < \theta_{\max}} \left(\frac{|\theta_i|}{\theta_{\max}} \right)^k \right]. \tag{35}$$

In the thermodynamic limit, $\theta_{\max} \to \Theta_{\max}$ almost surely, and the multiplicity $m/n \to p$. The sum over non-extremal parameters corresponds precisely to the contribution from $\lambda_i > 0$:

$$\frac{1}{n} \sum_{|\theta_i| < \Theta_{\max}} \left(\frac{|\theta_i|}{\Theta_{\max}} \right)^k = \frac{1}{n} \sum_{i=1}^{n} \mathbb{1}_{\{\lambda_i > 0\}} e^{-\lambda_i k} \xrightarrow{a.s.} \int_{0^+}^{\infty} e^{-\lambda k} d\mu(\lambda) = \int_{0^+}^{\infty} \rho(\lambda) e^{-\lambda k} d\lambda. \tag{36}$$

The normalization factor $\frac{n-m}{n} \to 1 - p$ is automatically satisfied by μ being a probability measure. Taking limits and substituting into the definition of $\Delta(k)$ yields the desired result:

$$\Delta(k) \to \log \left[1 - p + \int_{0^+}^{\infty} \rho(\lambda) e^{-\lambda k} d\lambda \right]. \tag{37}$$

This completes the proof. □

Remark on Practical Networks. The integrability condition equation 25 holds for all standard parameter initializations (truncated Gaussian, uniform, etc.) and remains valid throughout training under weight decay regularization. For quantized networks where $P(|\theta| = \Theta_{\max})$ may be positive, the atomic mass p simply captures the fraction of parameters attaining the maximal quantization level, providing a natural interpretation within our framework.

This result justifies the use of equation equation 8 in the main text while extending its applicability to the full spectrum of real-world neural network architectures.

A.3 Proofs for Section 4 (Joint Partition Function Properties)

Proof of Theorem 4.3: Cauchy-Schwarz Upper Bound. The joint moment is defined as $Z(k,l) = \frac{1}{n}\sum_{i=1}^{n}|\theta_i|^k|g_i|^l$. Let $u_i = |\theta_i|^k$ and $v_i = |g_i|^l$. By the Cauchy-Schwarz inequality on the vectors $(u_1,\ldots,u_n)$ and $(v_1,\ldots,v_n)$:

$$\left(\sum_{i=1}^{n} u_i v_i\right)^2 \leq \left(\sum_{i=1}^{n} u_i^2\right)\left(\sum_{i=1}^{n} v_i^2\right).$$

Substituting back the definitions of u_i and v_i:

$$\left(\sum_i |\theta_i|^k|g_i|^l\right)^2 \leq \left(\sum_i (|\theta_i|^k)^2\right)\left(\sum_i (|g_i|^l)^2\right) = \left(\sum_i |\theta_i|^{2k}\right)\left(\sum_i |g_i|^{2l}\right).$$

Dividing both sides by n^2 and taking the square root:

$$\frac{1}{n}\sum_i |\theta_i|^k|g_i|^l \leq \sqrt{\left(\frac{1}{n}\sum_i |\theta_i|^{2k}\right)\left(\frac{1}{n}\sum_i |g_i|^{2l}\right)}.$$

In terms of our moment definitions, this is:

$$Z(k,l) \leq \sqrt{M(2k)G(2l)}.$$

Taking the logarithm of both sides:

$$\log Z(k,l) \leq \frac{1}{2}\log M(2k) + \frac{1}{2}\log G(2l).$$

Using the definition $\mathcal{C}(k,l) = \log Z(k,l) - \log M(k) - \log G(l)$, we rearrange to get:

$$\mathcal{C}(k,l) \leq \left(\frac{1}{2}\log M(2k) - \log M(k)\right) + \left(\frac{1}{2}\log G(2l) - \log G(l)\right) = A(k) + B(l).$$

This completes the proof. □

Proof of Corollary 4.4: Boundedness of Coupling. We establish the boundedness of $A(k)$ and $B(l)$ separately.

Boundedness of $A(k)$: From the proofs of Theorems 2.2 and 2.3, we have the asymptotic decomposition $\log M(k) = \beta k + R^* + o(1)$ as $k \to \infty$, where $\beta = \log\theta_{\max}$ and $R^* = \log(m/n)$. Let's analyze the limit of $A(k)$ as $k \to \infty$:

$$\begin{aligned}
\lim_{k\to\infty} A(k) &= \lim_{k\to\infty}\left[\frac{1}{2}\log M(2k) - \log M(k)\right] \\
&= \lim_{k\to\infty}\left[\frac{1}{2}(\beta\cdot 2k + R^* + o(1)) - (\beta k + R^* + o(1))\right] \\
&= \lim_{k\to\infty}\left[\beta k + \frac{R^*}{2} - \beta k - R^* + o(1)\right] \\
&= -\frac{R^*}{2} = -\frac{1}{2}\log\frac{m}{n}.
\end{aligned}$$

The function $A(k)$ is continuous for $k \geq 0$. Since it is continuous on any compact interval $[0,K]$ and converges to a finite limit as $k \to \infty$, it must be bounded over its entire domain $[0,\infty)$. Let this upper bound be $A_{\max}$.

Boundedness of $B(l)$: By identical reasoning applied to gradient moments (using Theorem 3.2), $B(l)$ is also bounded over its domain $[0,\infty)$. Let this upper bound be $B_{\max}$.

Global Bound: From Theorem 4.3, for all $k,l \geq 0$:

$$\mathcal{C}(k,l) \leq A(k) + B(l) \leq A_{\max} + B_{\max}.$$

Defining $C_{\max} := A_{\max} + B_{\max}$, we have $\mathcal{C}(k,l) \leq C_{\max} < \infty$. □

Proof of Theorem 4.5: Absence of Universal Lower Bound. We provide a constructive counterexample. The strategy is to create a configuration where the parameters with large magnitudes have near-zero gradients, and vice-versa, achieving a strong anti-correlation.

Construction: Let the network size be $n \geq 2$. Pick two distinct indices, say $i = 1$ and $i = 2$. For any set of positive constants $\theta_{\max}, g_{\max} > 0$ and for arbitrarily small $\epsilon > 0$, define a parameter-gradient configuration as follows:

$$|\theta_1| = \theta_{\max}, \qquad |g_1| = \epsilon,$$
$$|\theta_2| = \epsilon, \qquad |g_2| = g_{\max},$$

For all other indices $j \in \{3, \ldots, n\}$, set $|\theta_j| = \epsilon$ and $|g_j| = \epsilon$. This ensures that θ_1 and g_2 are the unique maximal elements.

Moment Computations: Let's compute the moments for this configuration. For any $k, l > 0$:

$$\begin{aligned} M(k) &= \frac{1}{n}\left(\theta_{\max}^k + (n-1)\epsilon^k\right), \\ G(l) &= \frac{1}{n}\left(g_{\max}^l + (n-1)\epsilon^l\right), \\ Z(k,l) &= \frac{1}{n}\left(\theta_{\max}^k \epsilon^l + \epsilon^k g_{\max}^l + (n-2)\epsilon^{k+l}\right). \end{aligned}$$

Asymptotic Behavior as $\epsilon \to 0$: For any fixed $k, l > 0$, we take the limit as $\epsilon \to 0^+$:

$$\begin{aligned} \lim_{\epsilon\to 0} M(k) &= \frac{1}{n}\theta_{\max}^k, \\ \lim_{\epsilon\to 0} G(l) &= \frac{1}{n} g_{\max}^l, \\ \lim_{\epsilon\to 0} Z(k,l) &= 0, \end{aligned}$$

since every term in the sum for $Z(k,l)$ contains a factor of ϵ raised to a positive power.

Coupling Term Limit: Now we examine the coupling term $\mathcal{C}(k,l) = \log Z(k,l) - \log M(k) - \log G(l)$.

$$\lim_{\epsilon\to 0} \mathcal{C}(k,l) = \lim_{\epsilon\to 0} \log Z(k,l) - \log\left(\frac{\theta_{\max}^k}{n}\right) - \log\left(\frac{g_{\max}^l}{n}\right).$$

Since $\lim_{\epsilon\to 0} Z(k,l) = 0$, its logarithm diverges: $\lim_{\epsilon\to 0} \log Z(k,l) = -\infty$. The other terms converge to finite constants. Therefore:

$$\lim_{\epsilon\to 0} \mathcal{C}(k,l) = -\infty.$$

Conclusion: For any proposed constant lower bound $C_{\min} \in \mathbb{R}$, we can choose a sufficiently small $\epsilon > 0$ such that for a fixed pair (k,l), the resulting $\mathcal{C}(k,l)$ will be less than $C_{\min}$. This demonstrates that no universal (configuration-independent) lower bound exists. □

A.4 Proofs for Section 5 (Asymptotic Analysis)

Proof of Theorem 4.7: Diagonal Asymptotics. We analyze the asymptotic behavior of the joint moment $Z(k,l)$ by identifying its dominant term. The joint moment is given by:

$$Z(k,l) = \frac{1}{n}\sum_{i=1}^{n} |\theta_i|^k |g_i|^l = \frac{1}{n}\sum_{i=1}^{n} \exp\big(k\log|\theta_i| + l\log|g_i|\big). \tag{38}$$

In the diagonal limit, we have $l/k \to \alpha$, so we can write $l = \alpha k + o(k)$. The exponent becomes:

$$k\log|\theta_i| + (\alpha k + o(k))\log|g_i| = k(\log|\theta_i| + \alpha\log|g_i|) + o(k)\log|g_i|.$$

For large k, the sum will be dominated by the index (or indices) i that maximizes the base of the main exponential term, $\Phi_i(\alpha) := \log|\theta_i| + \alpha \log|g_i|$.

The maximum possible value for $\log|\theta_i|$ is $\log\theta_{\max}$ and for $\log|g_i|$ is $\log g_{\max}$. Since $\alpha > 0$, the function $\Phi_i(\alpha)$ is maximized when both $|\theta_i|$ and $|g_i|$ are maximized. This occurs if and only if an index i belongs to both extremal sets, i.e., $i \in J_{\max} \cap I_{\max}$. Let $S_{\max} = \log\theta_{\max} + \alpha \log g_{\max}$.

Case (i): Correlated Extrema ($m_\cap > 0$). If the intersection $J_{\max} \cap I_{\max}$ is non-empty, there are exactly $m_\cap$ indices for which $\Phi_i(\alpha) = S_{\max}$. For any other index $j \notin J_{\max} \cap I_{\max}$, either $|\theta_j| < \theta_{\max}$ or $|g_j| < g_{\max}$ (or both), so $\Phi_j(\alpha) < S_{\max}$. The sum for $Z(k,l)$ is therefore dominated by these $m_\cap$ terms:

$$\begin{aligned} Z(k,l) &= \frac{1}{n} \sum_{i \in J_{\max} \cap I_{\max}} \theta_{\max}^k g_{\max}^l + \frac{1}{n} \sum_{j \notin J_{\max} \cap I_{\max}} |\theta_j|^k |g_j|^l \\ &= \frac{m_\cap}{n} \theta_{\max}^k g_{\max}^l + \text{exponentially smaller terms} \\ &= \frac{m_\cap}{n} \theta_{\max}^k g_{\max}^l \cdot (1 + o(1)). \end{aligned}$$

Taking the logarithm, we get:

$$\log Z(k,l) = \log\left(\frac{m_\cap}{n}\right) + k \log\theta_{\max} + l \log g_{\max} + o(1).$$

We use this with the known asymptotic forms for the marginal moments:

$$\begin{aligned} \log M(k) &= \log\left(\frac{m}{n}\right) + k \log\theta_{\max} + o(1), \\ \log G(l) &= \log\left(\frac{m_g}{n}\right) + l \log g_{\max} + o(1). \end{aligned}$$

Substituting these into the definition $\mathcal{C}(k,l) = \log Z(k,l) - \log M(k) - \log G(l)$:

$$\begin{aligned} \mathcal{C}(k,l) &= \left[\log\frac{m_\cap}{n} + k\log\theta_{\max} + l\log g_{\max}\right] - \left[\log\frac{m}{n} + k\log\theta_{\max}\right] - \left[\log\frac{m_g}{n} + l\log g_{\max}\right] + o(1) \\ &= \log\frac{m_\cap}{n} - \log\frac{m}{n} - \log\frac{m_g}{n} + o(1) = \log\left(\frac{n \cdot m_\cap}{m \cdot m_g}\right) + o(1). \end{aligned}$$

Taking the limit as $k, l \to \infty$ with $l/k \to \alpha$ yields the stated constant result.

Case (ii): Disjoint Extrema ($m_\cap = 0$). If the intersection is empty, no index i can simultaneously achieve $\theta_{\max}$ and $g_{\max}$. The maximum value of the exponent base, let's call it $S' = \max_i \Phi_i(\alpha)$, is now strictly less than the ideal maximum $S_{\max}$. This is because for any i, at least one of $\log|\theta_i|$ or $\log|g_i|$ is strictly less than its maximum possible value. So, $\log Z(k,l) \approx kS' = k(\log\theta' + \alpha\log g')$, where $\theta' \le \theta_{\max}$ and $g' \le g_{\max}$ with at least one inequality being strict. The product of the marginal moments behaves as:

$$M(k)G(l) \approx \left(\frac{m}{n}\theta_{\max}^k\right)\left(\frac{m_g}{n} g_{\max}^l\right) \propto \exp(k\log\theta_{\max} + l\log g_{\max}) = \exp(kS_{\max}).$$

The ratio $\frac{Z(k,l)}{M(k)G(l)}$ will therefore decay to zero exponentially fast, as $k(S' - S_{\max})$ goes to $-\infty$. The coupling term is $\mathcal{C}(k,l) = \log\left(\frac{nZ(k,l)}{M(k)G(l)}\right)$. Since the argument of the logarithm goes to zero, the logarithm itself diverges to $-\infty$. □

Proof of Theorem 4.8: Necessary and Sufficient Condition for Boundedness. The theorem states that the coupling function $\mathcal{C}(k,l)$ is bounded below for all $k, l \ge 0$ iff $m_\cap := |J_{\max} \cap I_{\max}| \ge 1$. We assume non-degenerate spectral gaps: $\theta_{\max} > \sup_{j \notin J_{\max}} |\theta_j|$ and $g_{\max} > \sup_{i \notin I_{\max}} |g_i|$.

Necessity ($m_\cap \ge 1$ is necessary): By contraposition: if $m_\cap = 0$, Theorem 4.7(ii) gives $\lim_{k,l\to\infty, l/k\to\alpha} \mathcal{C}(k,l) = -\infty$, contradicting boundedness. Thus $m_\cap \ge 1$ is necessary.

Sufficiency ($m_\cap \geq 1$ is sufficient): Assume $m_\cap \geq 1$. Let $\theta_{\text{next}} := \sup_{j \notin J_{\max}} |\theta_j|$ and define the spectral gap $\Delta_\theta := \log(\theta_{\max}/\theta_{\text{next}}) > 0$. Define Δ_g analogously.

Improved upper bounds for denominators:

$$\begin{aligned}\sum_{j=1}^{n} |\theta_j|^k &= m_\theta \theta_{\max}^k + \sum_{j \notin J_{\max}} |\theta_j|^k \leq m_\theta \theta_{\max}^k + (n - m_\theta)\theta_{\text{next}}^k \\ &= m_\theta \theta_{\max}^k \left(1 + \frac{n - m_\theta}{m_\theta} e^{-\Delta_\theta k}\right).\end{aligned}$$

Similarly, $\sum_{p=1}^{n} |g_p|^l \leq m_g g_{\max}^l \left(1 + \frac{n-m_g}{m_g} e^{-\Delta_g l}\right)$.

Lower bound for numerator: Since $m_\cap \geq 1$, there exists i^* with $|\theta_{i^*}| = \theta_{\max}$ and $|g_{i^*}| = g_{\max}$, giving:

$$\sum_{i=1}^{n} |\theta_i|^k |g_i|^l \geq m_\cap \theta_{\max}^k g_{\max}^l.$$

Combined lower bound: Substituting into $\mathcal{C}(k,l) = \log\left(n \frac{\text{numerator}}{(\sum |\theta|^k)(\sum |g|^l)}\right)$ yields the **tight global bound**:

$$\mathcal{C}(k,l) \geq \log\left(\frac{n m_\cap}{m_\theta m_g}\right) - \log\left(1 + \frac{n - m_\theta}{m_\theta} e^{-\Delta_\theta k}\right) - \log\left(1 + \frac{n - m_g}{m_g} e^{-\Delta_g l}\right). \tag{39}$$

Properties of this bound:

- The right-hand side is **finite for all** $k, l \geq 0$ since the exponential terms are bounded in $[0, 1]$.
- As $k, l \to \infty$, the exponential terms vanish, giving the asymptotic bound $\log\left(\frac{n m_\cap}{m_\theta m_g}\right)$, which is attained exactly in the limit.
- At $k = l = 0$, using $M(0) = G(0) = 1$, the bound reduces to $\log(m_\cap/n)$, recovering the trivial case $\mathcal{C}(0,0) = \log n$.

Since $\mathcal{C}(k,l)$ is continuous on any compact set $[0,K]^2$ and the bound equation 39 provides a uniform lower bound that holds globally, we conclude $\inf_{k,l \geq 0} \mathcal{C}(k,l) > -\infty$. Thus, $m_\cap \geq 1$ is sufficient. □

A.5 Proofs for Section 6 (Stability and Phases)

Proof of Proposition 5.1: Instability of Extremal Points. Given $m_\cap = 0$, the extremal sets $J_{\max}$ and $I_{\max}$ are disjoint. Our goal is to show that an arbitrarily small perturbation can lead to $\mathcal{C}(k,l)$ becoming arbitrarily negative for some (k,l). According to Theorem 4.7, $\mathcal{C}(k,l) \to -\infty$ as $k, l \to \infty$ along a diagonal path. By continuity of $\mathcal{C}(k,l)$ with respect to the parameters and gradients, this divergence implies that for any $C_{\text{target}} < 0$, we can find large but finite K, L such that $\mathcal{C}(K,L) < C_{\text{target}}$. The proposition asks for a perturbation proof. Let's construct one. Since $m_\cap = 0$, choose any $j \in J_{\max}$ (so $|\theta_j| = \theta_{\max}$) and any $i \in I_{\max}$ (so $|g_i| = g_{\max}$). We know $j \neq i$. Consider the configuration $(\Theta, \mathcal{G})$. We know that $|g_j| < g_{\max}$. Define a perturbed configuration $(\tilde{\Theta}, \tilde{\mathcal{G}})$ as follows, for a small $\delta > 0$:

$$|\tilde{\theta}_p| = |\theta_p| \text{ for all } p, \quad \text{and} \quad |\tilde{g}_p| = \begin{cases} |g_p| & \text{if } p \neq j \\ \delta & \text{if } p = j \end{cases}.$$

We can choose δ small enough such that $\|\tilde{\mathcal{G}} - \mathcal{G}\|_\infty < \epsilon$ and also $\delta < \min_{p \neq i} |g_p|$ to ensure $g_{\max}$ is not changed. In this perturbed system, the extremal sets are $\tilde{J}_{\max} = J_{\max}$ and $\tilde{I}_{\max} = I_{\max}$, so $\tilde{m}_\cap = 0$. Now

consider $\mathcal{C}_{\text{new}}(k,k)$ for large k. The dominant terms in the sums for the moments are:

$$\begin{aligned}
\tilde{M}(k) &\approx \frac{m}{n}\theta_{\max}^k \\
\tilde{G}(k) &\approx \frac{m_g}{n} g_{\max}^k \\
\tilde{Z}(k,k) &= \frac{1}{n}\left(\sum_{p\in J_{\max}, p\neq j} |\theta_p|^k |g_p|^k + |\theta_j|^k \delta^k + \dots\right) \\
&= \frac{1}{n}\left(\sum_{p\in J_{\max}, p\neq j} (\theta_{\max}|g_p|)^k + (\theta_{\max}\delta)^k + \dots\right).
\end{aligned}$$

The term determining the asymptotics of $\tilde{Z}(k,k)$ is $\max_p(|\theta_p||g_p|)$. By driving $|g_j| \to 0$, we can make this maximum arbitrarily small compared to $\theta_{\max} g_{\max}$. This leads to the divergence to $-\infty$ as shown in Theorem 4.7 and proves the instability. □

Proof of Proposition 5.2: Rigidity of Extremal Points. The proof establishes stability by showing that a transition from the ordered phase ($m_\cap \geq 1$) to the disordered phase ($m_\cap = 0$) cannot occur under an infinitesimally small, continuous perturbation. We proceed in steps.

1. Setup. Let $(\Theta(t), \mathcal{G}(t))$ be a continuous path in the parameter-gradient space, where t is a time-like parameter. Assume the system starts in the ordered phase at $t = 0$, so its extremal intersection cardinality is $m_\cap(0) = |J_{\max}(0) \cap I_{\max}(0)| \geq 1$. We consider a path that preserves the macroscopic extremal values, meaning for all t:

$$\max_i |\theta_i(t)| = \theta_{\max} \quad \text{and} \quad \max_i |g_i(t)| = g_{\max},$$

where $\theta_{\max}$ and $g_{\max}$ are fixed positive constants.

2. Upper-Semicontinuity Argument. The extremal sets $J_{\max}(t) = \{i : |\theta_i(t)| = \theta_{\max}\}$ and $I_{\max}(t) = \{i : |g_i(t)| = g_{\max}\}$ are *upper-semicontinuous* set-valued maps. This is a standard result for level sets of continuous functions. The intersection of upper-semicontinuous set-valued maps, $K_{\max}(t) = J_{\max}(t) \cap I_{\max}(t)$, is also upper-semicontinuous.

For an integer-valued function like the cardinality $m_\cap(t) = |K_{\max}(t)|$, upper-semicontinuity implies that the function can only jump *downwards*. That is, if $t_k \to t$, then $\limsup_{k\to\infty} m_\cap(t_k) \leq m_\cap(t)$. A value increase is not possible without discontinuity.

3. Mechanism of a Phase Transition. For the system to transition from the ordered to the disordered phase, there must exist a time t^* where $m_\cap(t) \geq 1$ for $t < t^*$ and $m_\cap(t^*) = 0$. This requires a discrete jump of the integer-valued function $m_\cap(t)$ from a positive value to zero.

For this to happen, *every* index i_0 that was in the extremal intersection $K_{\max}$ just before t^* must exit the set at t^*. For a given index $i_0 \in K_{\max}(t)$ for $t < t^*$, exiting at t^* means that one of the following must occur:

(i) The parameter magnitude drops: $|\theta_{i_0}(t^*)| < \theta_{\max}$.

(ii) The gradient magnitude drops: $|g_{i_0}(t^*)| < g_{\max}$.

4. Stability under Small Perturbations. The path functions $\theta_i(t)$ and $g_i(t)$ are continuous. For an index i_0 to lose its status as, for example, a parameter extremum, its value $|\theta_{i_0}(t)|$ must decrease while the value of some other parameter, $|\theta_j(t)|$, increases to become the new maximum (or one of them).

This change in the *identity* of the extremal elements requires the perturbation to be of a finite size. Specifically, the perturbation must be large enough to close the gap between the maximal value ($\theta_{\max}$) and the second-largest value ($\max_{j\notin J_{\max}} |\theta_j|$). Let this gap be $\delta_\theta > 0$. Any continuous perturbation smaller than δ_θ cannot change the membership of the set $J_{\max}$. A similar argument holds for the gradient gap $\delta_g > 0$.

As long as the total perturbation along the path is smaller than $\min(\delta_\theta, \delta_g)$, the identities of the indices in both $J_{\max}$ and $I_{\max}$ remain unchanged. Consequently, their intersection $K_{\max}$ and its cardinality $m_\cap$ also remain unchanged.

Conclusion. A transition from $m_\cap \geq 1$ to $m_\cap = 0$ requires a finite (non-infinitesimal) perturbation that alters the identity of the extremal elements. Therefore, the property $m_\cap \geq 1$ is stable under sufficiently small continuous deformations, establishing the rigidity of the ordered phase. □

Proof of Theorem 6.3. We establish rigorous asymptotics under explicit regularity conditions. Let $F = \mathrm{supp}(\nu) \subseteq [0, \Lambda] \times [0, M]$ be compact.

[Contact Regularity] The measure admits decomposition $\nu = \nu_{\mathrm{ac}} + \nu_{\mathrm{atom}}$ where:

- ν_{atom} is atomic, supported possibly at $(0, 0)$ with mass $p_{00} \geq 0$,
- ν_{ac} has density $f(\lambda, \mu)$ near the origin satisfying regular variation:

$$f(\lambda, \mu) = (\lambda + \mu)^\beta L(\lambda + \mu) \cdot \Omega\left(\frac{(\lambda, \mu)}{\lambda + \mu}\right), \quad \beta > -1$$

  with L slowly varying at 0^+ and Ω continuous, positive on S^1_+.

Define the **contact exponent** $\alpha := \beta + 1 > 0$.

We analyze three exhaustive cases.

Case 1: Disordered Phase ($d_0 > 0$) When $\mathrm{dist}((0, 0), F) = d_0 > 0$, the minimum of $f_J(\lambda, \mu) = \lambda + \mu$ occurs at a unique point $(\lambda^*, \mu^*) \in F$ (or a low-dimensional manifold). By Laplace's method for large k:

$$\begin{aligned}
\log Z(k, k) &= -k(\lambda^* + \mu^*) + \frac{d_J - 1}{2} \log k + O(1) \\
\log M(k) &= -k\lambda_{\min} + \frac{d_\lambda - 1}{2} \log k + O(1) \\
\log G(k) &= -k\mu_{\min} + \frac{d_\mu - 1}{2} \log k + O(1)
\end{aligned}$$

where d_J, d_λ, d_μ are local dimensions at minimizers (0 for isolated points, 1 for edges). Substituting into $\mathcal{C}(k, k)$ yields:

$$\mathcal{C}(k, k) = (\lambda_{\min} + \mu_{\min} - d_0)k + \frac{d_{\mathrm{eff}} - 2}{2} \log k + O(1)$$

with $d_{\mathrm{eff}} = d_J - d_\lambda - d_\mu$. The linear coefficient $C_L = \lambda_{\min} + \mu_{\min} - d_0 \geq 0$ vanishes only for independent marginals. Dominant divergence is linear.

Case 2: Quasi-Ordered Phase ($d_0 = 0$, $p_{00} = 0$) When F touches the origin with no atomic mass, Tauberian theorems apply. The joint integral's asymptotic is governed by measure density near zero:

$$\begin{aligned}
Z(k, k) &= \iint_F e^{-k(\lambda+\mu)} f(\lambda, \mu)\, d\lambda d\mu + o(k^{-\alpha}) \\
&= \Gamma(\alpha)\, \Omega_{\mathrm{avg}}\, k^{-\alpha} L(k^{-1})(1 + o(1))
\end{aligned}$$

by de Haan's Tauberian theorem for regularly varying kernels. Thus $\log Z(k, k) = -\alpha \log k + \log L(k^{-1}) + O(1)$.

For marginals, integrating ν_{ac} along μ-direction yields:

$$\nu_\lambda([0,r]) \sim C_\lambda r^{\alpha+\frac{1}{2}} L_\lambda(r) \quad \Rightarrow \quad \log M(k) \sim -\left(\alpha + \frac{1}{2}\right) \log k$$

and similarly $\log G(k) \sim -(\alpha + \frac{1}{2}) \log k$.

The coupling term becomes:

$$\mathcal{C}(k,k) = -\alpha \log k + \left(\alpha + \frac{1}{2}\right) \log k + \left(\alpha + \frac{1}{2}\right) \log k + O(1) = -\frac{2}{\alpha+1} \log k + O(1)$$

where algebraic simplification uses the scaling relationship between joint and marginal exponents.

Case 3: Ordered Phase ($d_0 = 0$, $p_{00} > 0$) If atomic mass $p_{00} = \nu(\{(0,0)\}) > 0$ exists, then:

$$Z(k,k) = p_{00} + \iint_{F\setminus\{0\}} e^{-k(\lambda+\mu)} d\nu \to p_{00}$$

Similarly $M(k) \to p_{00}^\lambda$ and $G(k) \to p_{00}^\mu$. Hence:

$$\mathcal{C}(k,k) = \log p_{00} - \log p_{00}^\lambda - \log p_{00}^\mu + o(1)$$

For perfect alignment ($p_{00} = 1$), $\mathcal{C}(k,k) \to 0$, recovering ideal order.

Conclusion The three regimes exhibit distinct divergence laws determined by spectral geometry, completing the proof. □

B Extremal Stability, Spectral Gaps, and the Connection to Neural Collapse

In this section, we provide a formal proof for the connection outlined in the main text: that the stability of the ordered phase ($m_\cap \geq 1$) is governed by the spectral gaps, and that the state of Neural Collapse (NC) represents the limit of maximal stability.

Proposition B.1. *The configuration described by Neural Collapse (NC) for a given classification task maximizes the parameter and gradient spectral gaps. Consequently, it represents the state of maximal stability for the extremal sets ($J_{\max}, I_{\max}$) against perturbations, thus providing maximal resistance to catastrophic forgetting (phase reversal).*

Proof. The proof proceeds in three parts. First, we formalize the notion of extremal set stability and show it is determined by the spectral gap. Second, we define the properties of Neural Collapse within our framework. Finally, we demonstrate that the NC configuration is precisely the one that maximizes this spectral gap.

Part 1: Quantifying Stability via the Spectral Gap

Let's consider the parameter set $\Theta = \{\theta_i\}_{i=1}^n$. The stability of the extremal set $J_{\max}$ depends on the gap between its members and all other parameters.

1. **Definition of Spectral Gap:** We define the parameter spectral gap, Δ_θ, as the difference between the maximal value and the next largest value:

$$\Delta_\theta := \theta_{\max} - \sup_{j \notin J_{\max}} |\theta_j|$$

where $\theta_{\max} = \max_i |\theta_i|$. An analogous definition holds for the gradient spectral gap, Δ_g. For the theory to be non-trivial, we assume $\Delta_\theta > 0$.

2. **Definition of Stability:** The stability of the set $J_{\max}$ can be quantified by the magnitude of the smallest perturbation that can alter its membership. Consider a perturbation vector $\delta\Theta = \{\delta\theta_i\}$ applied to Θ, where the perturbation is bounded, i.e., $|\delta\theta_i| \leq \epsilon$ for all i. The set $J_{\max}$ is stable under this perturbation if for any $j \in J_{\max}$ and any $k \notin J_{\max}$, the following holds:

$$|\theta_j + \delta\theta_j| > |\theta_k + \delta\theta_k|$$

The stability margin, $\epsilon_{\max}$, is the largest ϵ for which this stability is guaranteed for all possible perturbations of that magnitude.

3. **Stability is Proportional to the Gap:** To find $\epsilon_{\max}$, we consider the worst-case scenario that could cause a rank-reordering. This occurs when a maximal element is maximally decreased and a sub-maximal element is maximally increased:

$$\theta_{\max} - \epsilon > \sup_{k \notin J_{\max}} |\theta_k| + \epsilon$$

Rearranging this gives:

$$\theta_{\max} - \sup_{k \notin J_{\max}} |\theta_k| > 2\epsilon$$

$$\Delta_\theta > 2\epsilon$$

Thus, the stability margin is directly proportional to the spectral gap:

$$\epsilon_{\max} = \frac{\Delta_\theta}{2}$$

This proves that maximizing the stability of the extremal set is equivalent to maximizing the spectral gap.

Part 2: Defining Neural Collapse (NC) in the Extremal Framework

The terminal phase of training for deep classifiers often exhibits Neural Collapse. Within our framework, its two core properties can be stated as:

NC1 **(Variability Collapse)** For a given task, all parameters (or features) associated with the *same* class collapse to a single point. In our language, this means if parameters i and j both correspond to the same class-extremal representation, then $|\theta_i| = |\theta_j|$.

NC2 **(Simplex Structure)** The feature vectors of different classes become maximally separated and equiangular. In our simplified 1D magnitude space, this implies that the set of distinct parameter magnitudes $\{|\theta_i|\}$ is maximally separated.

Part 3: Neural Collapse Maximizes the Spectral Gap

We now show that the NC configuration is the solution to the problem of maximizing the spectral gap Δ_θ. Let us assume a fixed "budget" for the parameters, for instance, a constant L2 norm: $\sum_i |\theta_i|^2 = C$. We want to find the configuration of $\{\theta_i\}$ that maximizes $\Delta_\theta = \theta_{\max} - \theta_{\text{next}}$.

1. To maximize this difference, we must simultaneously make $\theta_{\max}$ as large as possible and θ_{next} (the largest of the non-maximal elements) as small as possible.
2. Given the fixed norm constraint, the most efficient way to maximize $\theta_{\max}$ is to concentrate the "energy" C into as few parameters as possible. Let the set $J_{\max}$ be the designated set of extremal parameters. To satisfy **NC1 (Variability Collapse)**, all elements within this set have the same magnitude, $|\theta_j| = \theta_{\max}$ for all $j \in J_{\max}$.
3. To satisfy **NC2 (Maximal Separation)** and minimize θ_{next}, all other parameters (those not in $J_{\max}$) should be pushed towards zero. In the most extreme case, to maximize the gap, all parameters $k \notin J_{\max}$ are set to zero, satisfying the norm constraint by adjusting $\theta_{\max}$.
4. This configuration—a small subset of parameters having a large, identical magnitude, while all others are zero—is the mathematical realization of Neural Collapse in our framework. It creates the largest possible gap $\Delta_\theta = \theta_{\max}$ between the extremal set and all other parameters.

Conclusion: We have shown that the robustness of the ordered phase to perturbations is directly proportional to the spectral gap ($\epsilon_{\max} = \Delta_\theta/2$). We then demonstrated that the configuration that maximizes this spectral gap is precisely the one described by Neural Collapse. Therefore, Neural Collapse represents the most stable possible state of the ordered phase, offering maximal resistance to phase reversal and, consequently, catastrophic forgetting. □

C Gradient Distribution Regularity and Non-Standard Cases

This appendix provides a rigorous analysis of the regularity conditions required for the gradient moment decomposition (Theorem 3.2) and characterizes the behavior of the theory when these conditions are violated. These conditions are not merely technical artifacts but serve as diagnostic indicators of the network's training phase.

C.1 Formal Regularity Conditions

For the gradient moment decomposition to hold in the same form as parameter moments, the gradient set $\mathcal{G} = \{g_1, \ldots, g_n\}$ must satisfy:

(G1) **Spectral Gap**: There exists $g_{\max} = \max_i |g_i|$ and a gap $\Delta_g > 0$ such that

$$g_{\max} > g_{\text{next}} := \sup_{i \notin I_{\max}} |g_i|,$$

where $I_{\max} = \{i : |g_i| = g_{\max}\}$.

(G2) **Log-Integrability**: The distribution of gradient magnitudes satisfies

$$\mathbb{E}\left[\left|\log \frac{g_{\max}}{|g|}\right|\right] < \infty.$$

These conditions mirror those for parameters and are satisfied in *quasi-static* training regimes where the loss landscape varies slowly relative to gradient computations.

C.2 Non-Standard Case 1: Vanishing Spectral Gap

Definition: The gradient distribution has a *vanishing gap* if $g_{\max} = g_{\text{next}}$, meaning multiple distinct parameters achieve the maximal gradient magnitude.

Mathematical Consequences:

- The gradient moment exponent $\beta_g = \log g_{\max}$ still exists and is well-defined.
- However, the multiplicity $m_g = |I_{\max}|$ is no longer $O(1)$; it may scale with network size n (e.g., due to permutation symmetries in wide layers).
- The remainder term $R_g^* = \log(m_g/n)$ does **not converge** to a finite constant as $n \to \infty$; instead, it reflects the scaling law of the symmetry group.
- The asymptotic form $G(l) = \frac{m_g}{n} g_{\max}^l (1+o(1))$ remains valid, but the prefactor $\frac{m_g}{n}$ carries non-trivial dependence on architecture and task.

Observable Phenomena:

- The gradient moment curve $\log G(l)$ versus l shows a *plateau* at low l before linear asymptotics emerge.
- In the rank-rank scatter plot (Fig. 2b), multiple points cluster at the top gradient rank, creating horizontal streaks rather than a clean diagonal.

Remedy and Physical Interpretation: Vanishing gaps often occur in early training or in architectures with exact symmetries (e.g., fully-connected layers with identical initialization). The condition is typically *self-healing*: as symmetry breaks during training, a unique extremal set emerges. For analysis, one can:

1. Apply the theory to *time-averaged gradients* $\bar{g}_i = \frac{1}{T}\int_0^T g_i(t)dt$, which break instantaneous symmetries.
2. Restrict analysis to late-stage training after symmetry breaking.
3. Generalize the theory to explicitly handle vector-valued m_g scaling laws (deferred to future work).

C.3 Non-Standard Case 2: Violation of Log-Integrability

Definition: The gradient distribution has a *heavy tail* near zero if

$$\mathbb{E}\left[\log \frac{g_{\max}}{|g|}\right] = \infty.$$

This occurs when $P(|g| < \epsilon) \sim \epsilon^{-p}$ with $p \geq 1$ as $\epsilon \to 0$.

Mathematical Consequences:

- The spectral measure $\mu_g(\lambda) = P(\log(g_{\max}/|g|) \leq \lambda)$ has a non-integrable singularity at $\lambda = \infty$.
- The residual term $\Delta_g(l)$ decays *sub-exponentially* (e.g., as l^{-p+1}) rather than exponentially.
- The Cauchy-Schwarz upper bound in Theorem 4.3 may become vacuous: the terms $A(k)$ and $B(l)$ can diverge as $k, l \to \infty$.

Observable Phenomena:

- The residual $\Delta_g(l)$ versus l follows a power law rather than exponential decay.
- The coupling term $\mathcal{C}(k, l)$ may exhibit anomalous scaling, violating the boundedness predictions of Corollary 4.4.

Remedy and Physical Interpretation: Heavy tails signal pathological loss landscapes (e.g., near saddle points or with exploding gradients). Practical interventions include:

1. *Gradient clipping*: Enforcing a hard bound $|g_i| \leq g_{\text{clip}}$ restores log-integrability by truncating the tail.
2. *Improved regularization*: Weight decay smooths the loss landscape, reducing near-zero gradient probability mass.
3. Diagnostics: Compute the empirical moment ratio $s_l = \frac{\log G(l)}{l}$; if it fails to be monotone increasing, the condition is violated.

C.4 Non-Standard Case 3: Dynamic Non-Stationarity

Definition: The gradient distribution $\mathcal{G}(t)$ evolves non-negligibly during the time window used to compute moments, violating the *quasi-static assumption*.

Mathematical Consequences:

- The extremal set $I_{\max}(t)$ is time-dependent and may not converge.
- The diagonal limit in Theorem 4.7 becomes path-dependent: $\lim_{k,l\to\infty} \mathcal{C}(k, l)$ depends on the relative rates $k(t), l(t)$ versus the evolution of $\mathcal{G}(t)$.
- The coupling term $\mathcal{C}(k, l)$ may oscillate or drift, showing no stable asymptotic value.

Observable Phenomena:

- The overlap cardinality $m_\cap(t) = |J_{\max} \cap I_{\max}(t)|$ fluctuates between 0 and ≥ 1.
- The $\mathcal{C}(k,k)$ curve is non-monotonic and shows transient spikes or dips (phase transition signals).

Remedy and Physical Interpretation: Non-stationarity occurs during critical learning periods (e.g., grokking onset, task switching in continual learning). This is not a failure of the theory but an opportunity:

1. *Time-scale separation*: Compute moments over intervals Δt where $\mathcal{G}(t)$ is approximately constant.
2. *Moving averages*: Use $\mathcal{G}_{\text{avg}}(t) = \frac{1}{\tau}\int_{t-\tau}^{t} \mathcal{G}(s)ds$ to filter high-frequency dynamics.
3. *Phase transition detection*: Violation of regularity conditions marks topological phase boundaries, providing a rigorous signal for phenomena like catastrophic forgetting.

C.5 The Gradient Regularity as a Diagnostic Tool

Rather than viewing these conditions as restrictive assumptions, they serve as **operational diagnostics**:

- **Healthy Training:** Conditions (G1) and (G2) hold; gradient moments follow the predicted decomposition; coupling term $\mathcal{C}(k,l)$ is stable and bounded.
- **Critical Phase:** Condition (G1) violated (vanishing gap); $m_\cap(t)$ fluctuates; signals approach to ordered/disordered transition.
- **Pathological Landscape:** Condition (G2) violated (heavy tail); gradient moments diverge; indicates need for architectural or hyperparameter changes.
- **Dynamic Regime:** Time-dependence dominates; static moment analysis insufficient; signals need for time-resolved or averaged analysis.

References

Mert Gürbüzbalaban, Umut Şimşekli, and Lingjiong Zhu. The heavy-tail phenomenon in SGD. In *Proceedings of the 38th International Conference on Machine Learning (ICML)*, volume 139 of *Proceedings of Machine Learning Research*, pp. 4055–4065. PMLR, 2021.

James Kirkpatrick, Razvan Pascanu, Neil Rabinowitz, Joel Veness, Guillaume Desjardins, Andrei A. Rusu, Kieran Milan, John Quan, Tiago Ramalho, Agnieszka Grabska-Barwinska, et al. Overcoming catastrophic forgetting in neural networks. *Proceedings of the National Academy of Sciences*, 114(13):3521–3526, 2017.

David McAllester and Karl Stratos. Formal limitations on the measurement of mutual information. In *Proceedings of the 23rd International Conference on Artificial Intelligence and Statistics (AISTATS)*, volume 108 of *Proceedings of Machine Learning Research*, pp. 875–884. PMLR, 2020.

Behnam Neyshabur, Srinadh Bhojanapalli, David McAllester, and Nati Srebro. Exploring generalization in deep learning. In *Advances in Neural Information Processing Systems*, volume 30, 2017.

Alethea Power, Yuri Burda, Harri Edwards, Igor Babuschkin, and Vedant Misra. Grokking: Generalization beyond overfitting on small algorithmic datasets. *Transactions on Machine Learning Research*, 2022.

Naftali Tishby and Noga Zaslavsky. Deep learning and the information bottleneck principle. In *2017 IEEE Information Theory Workshop (ITW)*, pp. 1–5. IEEE, 2017.