ENTROPIC CONFINEMENT AND MODE CONNECTIVITY IN OVERPARAMETERIZED NEURAL NETWORKS

Anonymous authors
Paper under double-blind review

ABSTRACT

Modern neural networks exhibit a striking property: solutions at the bottom of the loss landscape are often connected by low-loss paths, yet optimization dynamics remain confined to one solution and rarely explore intermediate points. We resolve this paradox by identifying entropic barriers arising from the interplay between curvature variations along these paths and noise in optimization dynamics. Empirically, we find that curvature systematically rises away from minima, producing effective forces that bias noisy dynamics back toward the endpoints — even when the loss remains nearly flat. These barriers persist longer than energetic barriers, shaping the late-time localization of solutions in parameter space. Moreover, entropic confinement biases optimization away from poorly generalizing minima, helping to explain why such basins remain inaccessible despite their low training loss. Our results highlight the role of curvature-induced entropic forces in governing both connectivity and confinement in deep learning landscapes.

1 INTRODUCTION

Deep neural networks trained in the overparametrized regime exhibit a number of surprising and counterintuitive properties. One of the most striking is the observation that distinct solutions at the bottom of the training loss landscape are often connected by low-loss paths in parameter space (Garipov et al., 2018; Draxler et al., 2018; Frankle et al., 2020). Such *mode connectivity* results imply that the landscape is far less rugged than once assumed: minima that appear isolated are, in fact, linked by paths of nearly constant loss. At the same time, however, optimization dynamics display a seemingly contradictory behavior. Standard training with stochastic gradient descent (SGD), with or without momentum, converges to a well defined minimum and rarely explores regions of parameter space corresponding to these paths (Baity-Jesi et al., 2019).

We argue that this paradox can be resolved by recognizing the role of *entropic forces* generated by curvature variations along connecting paths. Although the loss may be nearly flat along these paths, the curvature of the landscape typically increases away from found minima, producing effective barriers that bias stochastic dynamics back toward the endpoints. These barriers emerge from the interaction between fluctuations induced by SGD noise and the Hessian spectrum along low-energy paths. In this way, regions of parameter space that are energetically connected become dynamically disconnected.

1.1 RELATED WORK

Our work aims to unify insights from two areas: First, we draw on insights from a body of work that shows SGD has an *implicit bias* towards flatter minima the strength of which increases with smaller minibatch size. This behavior is attributed to the higher noise levels in gradient estimates, which act as a form of implicit regularization preventing convergence to sharp minima and instead favoring wider, flatter basins that generalize better (Keskar et al., 2017). Several works have formalized this intuition: Jastrzebski et al. (2018) and Smith & Le (2018) interpret the minibatch noise as an effective temperature that enables exploration and escape from sharp valleys; Wei & Schwab (2019) and Xie et al. (2021) further show that the resulting dynamics can be modeled as a stochastic process biased toward flat regions. Collectively, these studies demonstrate that curvature and generalization are deeply intertwined with the stochastic geometry of the optimization trajectory. We

leverage these insights to conduct a deeper analysis of *mode connectivity* in neural network training. Garipov et al. (2018) and Draxler et al. (2018) showed the existence of nonlinear paths of low-loss between different minima found by training with different random seeds. Subsequently,Frankle et al. (2020) showed that when the training dynamics of two networks are tied together early in training, the resulting minima found after training is complete are *linearly* connected by paths of low-loss. Follow-up work has analyzed in more depth how linear mode connectivity emerges (Singh et al., 2024; Zhou et al., 2023), and such work has important implications for model merging (Ainsworth et al., 2023; Singh & Jaggi, 2020) and weight-space ensembling (Izmailov et al., 2018; Wortsman et al., 2021; Gagnon-Audet et al., 2023; Wortsman et al., 2022).

1.2 Contributions

Our main contributions are as follows:

- We show empirically that the curvature along minimum-loss paths between minima generically increases away from the endpoints.
- We argue that such a "bump" in the curvature leads to an *entropic barrier*, and that such entropic barriers lead to confinement of solutions even when the loss is near zero.
- We show that despite the existence of low-loss connecting paths between solutions, entropic forces dynamically confine models to specific regions of parameter space.
- We show that entropic barriers between minima persist longer than energetic barriers, when considering models that shared the first k epochs of training, suggesting that both energetic *and* entropic forces are responsible for the the final region of parameter space that a model ends up in.

2 Background: Entropic Forces and Curvature

It is well established in statistical physics that the dynamics of a system are governed not only by energetic forces—derived from gradients of an energy or potential—but also by entropic forces, arising from thermal fluctuations. In the context of neural networks, the energy landscape is defined by the training loss, and its gradient directs the deterministic component of learning. However, the stochasticity introduced by finite learning rates and minibatch sampling induces an effective temperature, making entropic contributions to the dynamics non-negligible. As a result, optimization can be biased toward broader, flatter regions of the landscape—not because they are lower in loss, but because they occupy a larger volume in parameter space. We illustrate this idea in a simple toy model. Consider a Brownian particle evolving in a two-dimensional potential

$$\dot{\boldsymbol{x}} = -\nabla V(\boldsymbol{x}) + \boldsymbol{\xi}(t), \qquad V(x,y) = \frac{1}{2} g(y) x^2, \tag{1}$$

where $\boldsymbol{\xi}$ is white delta correlated Gaussian noise with variance $2T$. In the analogy to deep learning, V plays the role of the task loss function, while the noise ξ arises from SGD noise due to minibatching and finite learning rate. To leading order one can identify $T \propto \eta/B$ (Mandt et al., 2017; Smith et al., 2020; Liu et al., 2021), though the precise relationship depends on local curvature and other details of the loss landscape (Ziyin et al., 2021). In this simple motivating example we assume ξ is white and Gaussian, even though in deep networks the SGD noise is neither perfectly white nor perfectly Gaussian. In this analogy y corresponds to soft modes (directions with nearly flat curvature where the loss hardly changes), and x represents the stiff modes (directions associated with large eigenvalues of the Hessian).

For fixed y, the distribution of x is Gaussian with variance $\mathbb{E}\left[x^2\right] = g(y)^{-1}$. If x relaxes on a faster timescale than y, which is the case if the curvature along the x direction is much higher than the curvature along the y direction, averaging yields effective dynamics for y:

$$\dot{y} = -g_y(y)\langle x^2 \rangle + \xi = -T\frac{g_y(y)}{g(y)} + \xi, \qquad g(y) > 0 \tag{2}$$

where $g(y)$ is an arbitrary positive function of the coordinate y. Equivalently, the marginalized distribution is

$$P(y) \propto \exp\Big[-V_{\text{eff}}(y)/T \Big], \qquad V_{\text{eff}}(y) = -T \ln g(y). \tag{3}$$

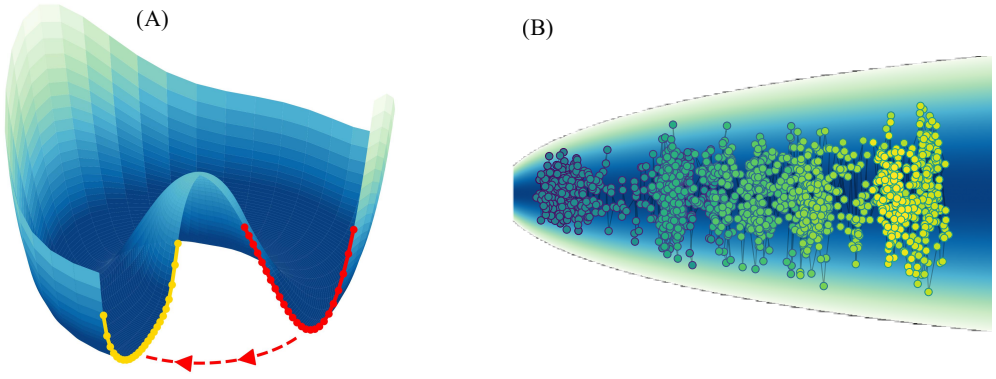

Figure 1: **Curvature produces an entropic force.** (A) Illustration of a potential $V(r, \theta)$ with a circular minimum at $r = 1$, where the curvature varies with angle. At zero temperature ($T = 0$), the angular distribution is uniform, $P(\theta) = 1/(2\pi)$. At finite temperature, thermal fluctuations bias the system toward flatter regions (yellow) rather than sharper ones (red). (B) Example of a Brownian particle diffusing along the ridge of a loss landscape, lighter colors correspond to larger times . Entropic forces generated by fluctuations push the particle toward flatter directions, effectively favoring broader regions of the landscape.

In Figure 1 we show two example of potentials with a gradient in the curvature which leads to an effective entorpic force pushing the system towards flatter regions. This reveals the key mechanism: regions with smaller curvature $g(y)$ (flatter directions in x) contribute larger entropy and are statistically favored, even if the original energy $V(x, y)$ is minimized elsewhere. In effect, curvature generates an entropic force that biases dynamics toward flatter regions of the landscape. We call these forces entropic because they are proportional to the effective temperature T, i.e., they vanish in the absence of noise. In the case of deep neural networks, we therefore expect these forces to grow stronger as the minibatch size decreases, since smaller minibatches correspond to higher effective temperature. We note also that depending on the interplay between entropy and energy, it is possible for entropic forces to be the stronger than energetic forces, leading to a scenario where entropy causes optimization to *climb* the loss. We see an example of this phenomenon in a deep network in Section 4.1.

This minimal example is far from capturing the true dynamics at the bottom of the loss landscape of real deep neural networks. However, it illustrates how stochasticity couples with curvature to favor flatter minima. Even at near-zero training loss, such entropic barriers can confine solutions to specific regions of parameter space. In Section 4.1, we show empirically that entropic forces arising from loss landscape curvature of real networks trained on natural images lead to qualitatively similar behavior.

3 Methods

We are interested in understanding whether models remain confined to well-defined regions of parameter space, even when low-loss paths connect distinct solutions, and if so what mechanisms create such confinement. To investigate this, we train image classification models on CIFAR-10 using Wide ResNet and ResNet architectures. We chose this setup as emblematic of mode connectivity. We obtain a set of distinct minima by training from different random seeds for parameter initialization and data ordering. In addition, we deliberately construct minima that overfit while still being global minima of the loss. We then proceed to find the constant loss non-linear path connecting these minima and evaluate the curvature of the loss landscape along such paths.

3.1 Training Details

Unless otherwise specified, all experiments are conducted on Wide ResNet-16-4 (Zagoruyko & Komodakis, 2016) trained on the CIFAR-10 dataset (Krizhevsky, 2009). Following standard practice (Zagoruyko & Komodakis, 2016), we use stochastic gradient descent (SGD) with momentum $\beta =$

0.9, weight decay $w = 5 \times 10^{-4}$, and an initial learning rate of $\eta = 0.1$. Models are trained for 200 epochs with a batch size of 256, and the learning rate is reduced by a factor of 5 at 30%, 60%, 80%, and 90% of the total training epochs. We apply mild data augmentation consisting of random horizontal flips and random crops with 4-pixel padding followed by resizing to the original 32×32 resolution.

3.2 Minimum energy paths

To explore the structure of the loss landscape between different solutions, we identify low-loss connecting paths using the *Automatic Nudged Elastic Band* (AutoNEB) algorithm introduced by Draxler et al. (2018). In brief, the algorithm initializes k intermediate *pivots* along the straight-line path between two minima and optimizes their positions such that they evolve as if connected by elastic springs, while minimizing the loss orthogonal to the path.

Since the loss along the straight segments between pivots may still be high, AutoNEB dynamically adds new pivots whenever the loss along a segment exceeds a predefined threshold. This adaptive refinement ensures a smooth, low-loss path is found. Following Draxler et al. (2018), we refer to such paths as minimum energy paths (MEPs), by analogy with physical systems. In the following, the relative position along the MEP is reported in terms of pivot index, normalized by the total number of pivots. Note that this parametrization does not reflect the actual metric distance along the path, as the pivot density is non-uniform: AutoNEB adaptively inserts additional pivots in regions where the loss landscape is sharper, requiring finer segmentation to accurately follow the MEP.

Unless otherwise specified, all MEPs shown in the paper are computed using a sequence of refinement cycles with decreasing learning rates. Specifically, we run four cycles each with the following parameters: $(0.1,\ 10)$, $(5 \times 10^{-2},\ 5)$, $(10^{-2},\ 5)$, and $(10^{-3},\ 5)$, where each tuple denotes (learning rate, number of epochs).

3.3 Curvature Measures

A natural measure of the curvature of the loss landscape is the Hessian of the loss function, defined as $\mathcal{H} \equiv \nabla^2_\theta \mathcal{L}(\theta)$. More precisely, it is the *spectrum* of the Hessian that captures the local geometry of the landscape. However, if the model has N parameters, then $\mathcal{H} \in \mathbb{R}^{N \times N}$, making it intractable to compute or store explicitly for modern networks. Instead, we use three independent summary statistics of the Hessian spectrum, each providing a tractable yet informative proxy for curvature.

We estimate the *maximum eigenvalue* of the Hessian, $\lambda_{\max}(\mathcal{H})$, using the *power iteration method* (see, e.g., Yao et al. (2020)). Crucially, this method requires only Hessian–vector products, which can be computed efficiently via automatic differentiation in $\mathcal{O}(N)$ time. The update rule for the power method is:

$$v^{(n+1)} = \frac{\mathcal{H}v^{(n)}}{\|\mathcal{H}v^{(n)}\|}, \qquad \text{where} \quad \mathcal{H}v = \sum_\beta \frac{\partial^2 \mathcal{L}(\theta)}{\partial\theta_\alpha \partial\theta_\beta} v_\beta. \tag{4}$$

After a few iterations, $v^{(n)}$ converges to the dominant eigenvector, and $\lambda_{\max} \approx \|\mathcal{H}v^{(n)}\|$.

We estimate the *trace* of the Hessian and part of its spectrum using its connection to the Fisher Information Matrix near a minimum. Specifically, when $\theta^\star$ is a local minimum and the model is well-calibrated, the Hessian can be approximated by the Fisher Information Matrix:

$$\mathcal{F}(\theta^\star) \equiv \mathbb{E}_{(x,y)\sim D}\left[s_\theta(x,y) s_\theta^\top(x,y)\right]\Big|_{\theta^\star} \qquad s_\theta(x,y) \equiv \nabla_\theta \log p_\theta(y \mid x) \tag{5}$$

where $s_\theta(x,y)$ is the score (**?**). This equivalence is discussed further in the Appendix A.2. From this expression, we can compute the trace of the Fisher—and hence approximate the trace of the Hessian—by summing the diagonal elements of the outer product $s_\theta s_\theta^\top$.

As a third measure, we compute the Fisher matrix on a small random subset of the training dataset of size E, and perform singular value decomposition (SVD) on the resulting score matrix, which has shape $N \times (CE)$, where N is the number of parameters and C the number of classes. This procedure yields an estimate of the leading components of the curvature spectrum, allowing us to capture the dominant eigenmodes efficiently without requiring full-batch computation or explicit construction of the full Fisher or Hessian matrices.

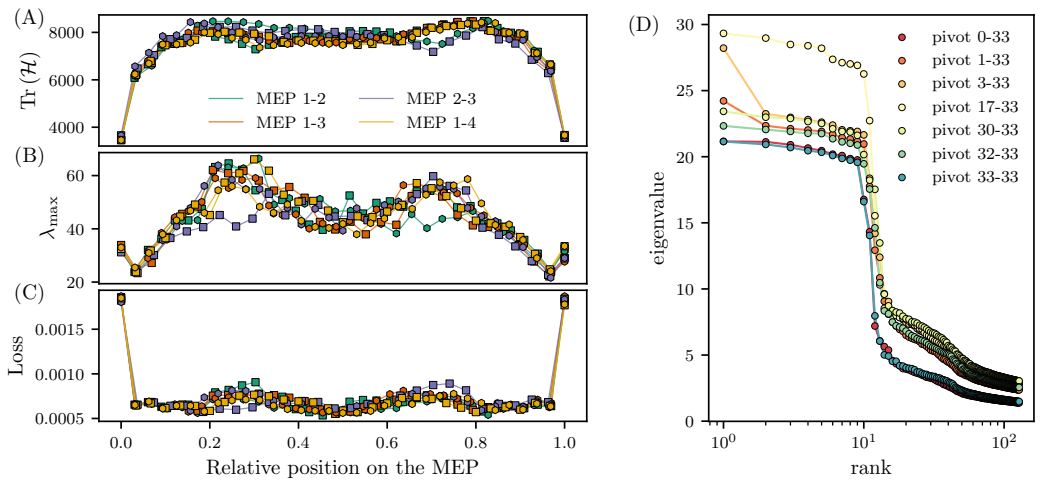

Figure 2: Entropy induces barriers between minima. **(A)** Cross entropy loss as in equation**??** along MEPs connecting different pairs of regular minima. **(B, C)** Curvature along minimum energy paths (MEPs) connecting different minima, measured via the maximum eigenvalue of the Hessian (**B**) and the trace of the Hessian (**C**). Markers indicate pivot points; different colors correspond to different pairs of minima, and marker shapes denote different AutoNEB realizations. **(D)** Spectrum of the Hessian along MEP 1–2, estimated via singular value decomposition (SVD) of the score matrix computed on $E = 1024$ training examples. As we move into the interior of the MEP, the entire spectrum shifts upward, reflecting a global increase in curvature along the path.

3.3.1 A Note on Reparameterization

Dinh et al. (2017) showed that symmetries in the architecture of networks allow deep networks to be re-parameterized without changing the function computed by the network. While this observation casts some doubt on the causal relationship between flatness and generalization (and in fact there are empirical measurements that show this relationship is not absolute (Kaur et al., 2023)), we note that when considering SGD optimization dynamics it is still the Hessian that governs the dynamics of the system. Particularly, for any symmetry T_α that leaves the function computed by the network the same, we have that $\boldsymbol{\nabla}_\alpha L(T_\alpha \theta) = 0$, and so symmetries do not have any affect on optimization dynamics.

4 Results

4.1 Entropic Confinement

In Figure 2, we show the loss (**A**) and the curvature—quantified by the maximum eigenvalue $\lambda_{\max}(\mathcal{H})$ (**B**) and the trace $\mathrm{Tr}(\mathcal{H})$ (**C**) of the Hessian—along MEPs connecting different pairs of minima of Wide ResNet-16-4. Interestingly, the loss along the MEP is often lower than at the endpoints. This behavior likely arises because each pivot is pulled downward both by the loss gradient (locally minimizing energy) and by the elastic coupling to neighboring pivots. Despite the absence of loss barriers along the MEPs, we observe a sharp rise in curvature along the MEP[1], measured either via $\lambda_{\max}$ or the Hessian trace. The curvature decreases only near the endpoint minima. As argued in Section 2, such variations in curvature generate entropic forces that bias optimization toward flatter regions, even in the absence of explicit loss barriers. One might argue that the increase in sharpness along the MEP is simply a consequence of the decreasing loss or an effective reduction in

[1]A small dip is visible near the endpoints in Figure 2(B), where the estimated maximum eigenvalue of the Hessian briefly decreases. We believe this artifact is due to the estimation procedure: computing the Hessian away from an exact minimum introduces a correction proportional to the norm of the gradient. This effect is stronger at the ends of the MEP, where the loss is slightly higher. Interestingly, this dip is not present in the estimates based on the Fisher Information Matrix (panel (**C**) and (**D**)).

learning rate, especially given prior work suggesting a relationship between sharpness and learning rate (Cohen et al., 2021). We argue that this is not the case: while the loss drops between the first and second pivots, it then remains approximately constant along the rest of the MEP. In contrast, both sharpness metrics—maximum eigenvalue and trace of the Hessian—continue to rise. This indicates that the observed increase in curvature is not merely a byproduct of lower loss or implicit regularization, but rather reflects a genuine change in the geometry of the optimization landscape.

To directly observe these entropic effects, we initialize models at specific points along a given MEP and study how stochastic gradient descent pushes them along the path. We use a variant of SGD that projects updates back onto the nearest linear segment of the MEP, ensuring that dynamics remain constrained to the path (see Section A.1 for details). As shown in Figure 3(A), when a model is initialized at relative position 0.7 along the linear path between the first and second pivots of MEP 1–2, it is pushed backward toward the flatter endpoint of the path. As expected, the strength of this entropic drift increases for smaller batch sizes, where the stochastic fluctuations are stronger. We note that entropic forces drive the optimization back towards the first pivot despite the fact that the loss actually increases along the optimization trajectory, illustrating a scenario where entropic force is stronger than energetic force.

In Figure 3(B), we quantify this behavior by measuring the relaxation time t_{relax}—the time it takes for the model to return to the first pivot. We find that t_{relax} decreases with increasing batch size, consistent with the intuition that entropic forces scale with the effective temperature of the dynamics. Assuming that the effective temperature is inversely proportional to the relaxation time and fitting the relation $T \sim B^{-\rho}$, we estimate $\rho = 1.4 \pm 0.1$, which is significantly larger than the naive prediction $\rho = 1$. This deviation may be attributed to the non-Gaussian nature of minibatch noise and to nonlinear effects in the dynamics.

The same mechanism applies to models initialized deeper along the MEP. In Figure 3(C), we initialize the model at the fourth pivot and observe relaxation back toward the first pivot under the same projected dynamics. The system follows the MEP segment by segment, switching to the next closest segment as it progresses. Discontinuities in the trajectory correspond to these segment transitions, reflecting the piecewise-linear nature of the projected updates.

The increase of the curvature along the MEP adds nuance to the idea that the loss landscape consists of one large "valley" containing all the parameter configurations with low-loss: Although minima in such a valley may be connected energetically, our experiments suggest that such a valley is broken up into disconnected regions by entropic barriers.

4.2 Linear mode connectivity

Although we have shown evidence that entropic forces separate the zero-loss region of parameter space into dynamically confined regions, we have not yet addressed how and when these confined regions are chosen along the course of training. In this section we will take some steps towards answering that question through the lens of linear mode connectivity. Following the methods of Frankle et al. (2020), we train M networks with a *shared* data order up until epoch k, which we will call the *splitting epoch.* After epoch k, each of the M networks sees an *independent* ordering of the data and can then potentially move away from its "siblings," the other $M-1$ networks. The sibling networks are then trained until convergence. All networks trained in this section use the ResNet-20 architecture (He et al., 2015).

The crucial observation in Frankle et al. (2020) is that once k becomes sufficiently large, the sibling networks become connected by linear paths of low-loss, implying that they converge to the same region of parameter space. Interestingly, k does not have to be very large compared to the number of epochs required for convergence before linear mode connectivity is observed. In Figure 4, we reproduce these experiments, and measure the curvature along the linear, low-loss paths between converged siblings. Similarly to the nonlinear case (Section 4.1), we see a bump in the curvature along these paths, peaking around $\alpha = 0.5$. However, we also notice that these entropic barriers persist for larger values of k than their energetic counterparts, implying that entropic forces are responsible for the final stages of the model's localization to a region of parameter space.

Although we have shown evidence that entropic forces separate the zero-loss region of parameter space into dynamically confined regions, we have not yet addressed how and when these confined

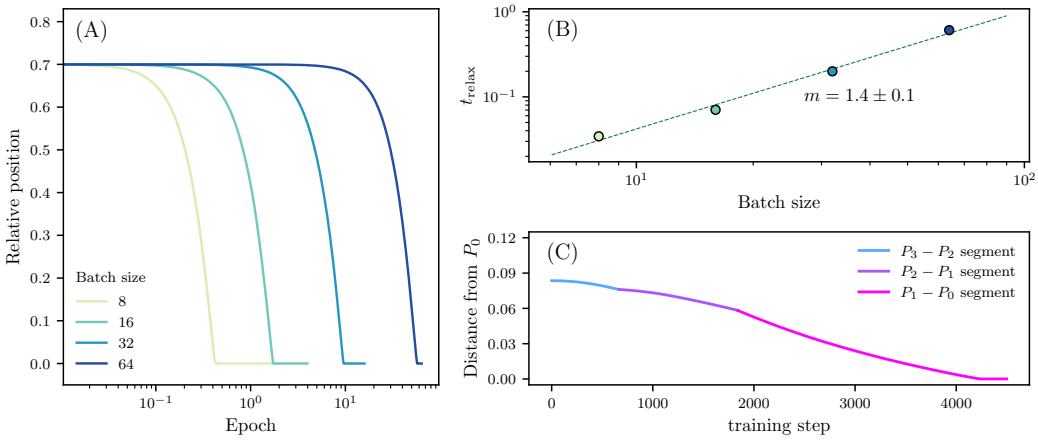

Figure 3: Relaxation dynamics induced by entropic forces. (**A**) A model is initialized at relative position 0.7 along the MEP between two minima (between the first and second pivots). Using projected SGD constrained to the path (see Section A.1), the model drifts toward the endpoint at relative position 0.0 due to entropic pressure. (**B**) The relaxation time t_{relax}, defined as the time required to return to the first pivot, decreases with batch size. This scaling supports the interpretation of entropic forces as temperature-dependent, with an effective temperature that grows as batch size decreases. (**C**) A model is initialized at the fourth pivot P_3 and relaxes along the MEP toward the endpoint at P_0. Different colors indicate which segment of the piecewise-linear MEP the model is currently on, and sharp transitions in the trajectory correspond to changes in the nearest segment under projection.

regions are chosen along the course of training. In this section we will take some steps towards answering that question through the lens of linear mode connectivity. Following the methods of Frankle et al. (2020), we train M networks with a *shared* data order up until epoch k, which we will call the *splitting epoch.* After epoch k, each of the M networks sees an *independent* ordering of the data and can then potentially move away from its "siblings," the other $M - 1$ networks. The sibling networks are then trained until convergence.

The crucial observation in Frankle et al. (2020) is that once k becomes sufficiently large, the sibling networks become connected by linear paths of low-loss, implying that they converge to the same region of parameter space. Interestingly, k does not have to be very large compared to the number of epochs required for convergence before linear mode connectivity is observed. In Figure 4, we reproduce these experiments, and measure the curvature along the linear, low-loss paths between converged siblings. Similarly to the nonlinear case (Section 4.1), we see a bump in the curvature along these paths. However, we also notice that these entropic barriers persist for larger values of k than their energetic counterparts, implying that entropic forces contribute to the final stages of the model's localization to a region of parameter space. To see this, we plot the *instability* as a function of k (Figure 4D). The instability measures the fractional change in the metric (loss or curvature) along the linear path. For small values of k, the loss has the larger instability, while for larger values of k the curvature exhibits greater instability.

5 Discussion

Entropic Confinement. Our results provide new insight into the global geometry of the loss landscape. While prior work has emphasized that minima are often connected by low-loss paths, forming a single broad "valley" of solutions (Garipov et al., 2018; Frankle et al., 2020), our findings reveal that these paths are not dynamically flat (Figure 2). Instead, they exhibit systematic increases in curvature away from their endpoints, producing localized "bumps" in sharpness. This observation refines the valley picture: the basin of low-loss solutions is structured by curvature variations that give rise to entropic barriers.

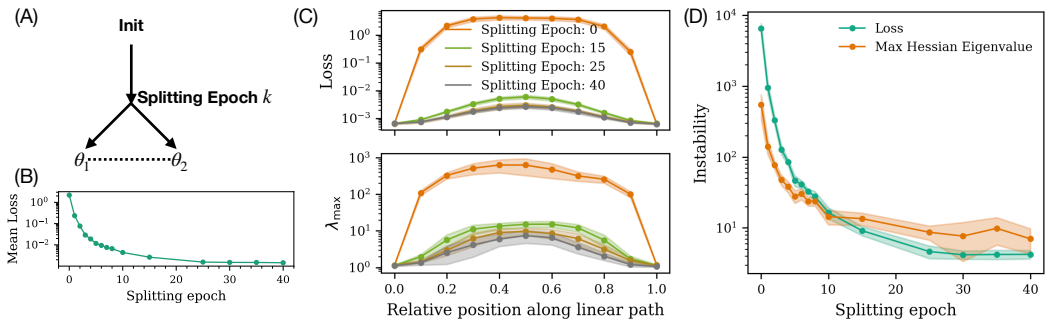

Figure 4: Entropic barriers are relevant later in training. **(A)** Linear mode connectivity schematic (Frankle et al., 2020). We train a network to epoch k, then produce two new networks via different data ordering, and measure the loss along a linear path. **(B)** The average loss along such a path goes down as k increases, decreasing rapidly with k. **(C)** *Top:* The loss profile along linear paths for various k. *Bottom:* The curvature profile, measured by the maximum Hessian eigenvalue, for various k. **(D)** We plot the *instability* (The fractional change along the path) of the loss and the curvature. For small k, the loss exhibits larger instability, while for larger k, the curvature exhibits larger instability.

We show that the forces produced by curvature variations along connecting paths consistently drive optimization dynamics back toward flatter regions near the minima (Figure 3). In particular, models initialized away from a minimum but constrained to remain on the path show persistent drift back toward the endpoint, even though the loss profile is nearly flat. We also observe that smaller batches accelerate relaxation, showing that the strength of the entropic force depends on the noise level of the dynamics. Entropic forces are not necessarily negligible – we show empirically that they can drive models *up* a loss gradient.

Entropic Linear Mode Connectivity. Our analysis of linear mode connectivity further shows that entropic forces play an important role late in training. As the splitting epoch increases, energetic barriers along linear paths decrease, but curvature barriers persist for longer into training. This suggests a two-phase picture of training: early dynamics are dominated by energetic forces that drive the model into a low-loss basin, while later dynamics are governed by entropic confinement that selects a specific region within that basin. Our experiments have important implications for late-time dynamics of deep network training and parameter-space ensembling techniques.

Confinement and Generalization. Our findings may also provide insight into generalization properties of overparameterized models. Empirically, models trained with SGD tend to find a generalizing solutions and not overfit the data, even after many epochs of training. This occurs *despite* the fact that the loss landscape is energetically flat, raising the question of why optimization dynamics do not diffuse into regions of parameter space that overfit the training data.

We posit that generalizing minima may be dynamically disconnected from overfit minima. Entropic barriers could make paths to such regions dynamically inaccessible: even when overfitting solutions are connected to flatter ones by low-loss paths, entropic forces could shield the generalizing solutions by repelling SGD away from regions of parameter space that do not generalize. Our results suggest that this is a promising avenue for future work.

Weight-space averaging. Our work also provides a new lens through which to view weight-space ensembling techniques. The study of global loss landscape features, such as mode connectivity (Draxler et al., 2018; Frankle et al., 2020), has been crucial in developing methods like Stochastic Weight Averaging (SWA) (Izmailov et al., 2018; Wortsman et al., 2021). Our findings suggest a more nuanced picture of the global landscape: techniques like SWA may be averaging minima that, while energetically connected within a single low-loss valley, may be dynamically disconnected by the entropic barriers we observe. This would imply that the SWA solution cannot be easily found by diffusive optimization dynamics at the bottom of a valley in the loss landscape. A valuable avenue for future work would be to analyze the connectivity properties of these averaged minima to better

understand how weight-space averaging is able to construct solutions with favorable generalization properties.

6 Conclusion

We identify a key geometric feature of neural network loss landscapes and its impact on optimization dynamics. Our central finding is that low-loss paths connecting distinct minima consistently exhibit a rise in curvature away from their endpoints. We show that this variation, when coupled with the inherent noise of stochastic gradient descent, gives rise to entropic barriers. We demonstrate empirically that these barriers generate effective forces that dynamically confine the optimizer to flatter regions near the minima, even when the path is energetically favorable.

Our experiments exploring the curvature along linearly mode-connected networks reveal that the mechanism of entropic confinement is particularly relevant during the later stages of training, shaping the final localization and stability of the learned solution. Our results establish these curvature-induced forces as a crucial element in understanding the behavior of stochastic optimizers. This geometric perspective offers new insights into how the landscape itself guides the discovery of stable and well-generalizing models, providing a promising direction for future research.

Ethics Statement. We do not foresee any direct ethical concerns arising from this work. Our study focuses on better understanding optimization in machine learning, without direct deployment in sensitive application domains. We note that we have used language models to polish the text of this manuscript in places.

Reproducibility Statement. We have provided descriptions of all algorithms, models, and experimental setups in the main text and appendix. Training procedures and dataset details are documented to facilitate replication. When the author list is unblinded, we will release our codebase to enable full reproducibility of our results.

References

Samuel Ainsworth, Jonathan Hayase, and Siddhartha Srinivasa. Git re-basin: Merging models modulo permutation symmetries. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=CQsmMYmlP5T.

Marco Baity-Jesi, Levent Sagun, Mario Geiger, Stefano Spigler, GÃ©rard Ben Arous, Chiara Cammarota, Yann LeCun, Matthieu Wyart, and Giulio Biroli. Comparing dynamics: deep neural networks versus glassy systems*. *Journal of Statistical Mechanics: Theory and Experiment*, 2019(12):124013, dec 2019. doi: 10.1088/1742-5468/ab3281. URL https://dx.doi.org/10.1088/1742-5468/ab3281.

Jeremy Cohen, Simran Kaur, Yuanzhi Li, J Zico Kolter, and Ameet Talwalkar. Gradient descent on neural networks typically occurs at the edge of stability. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=jh-rTtvkGeM.

Laurent Dinh, Razvan Pascanu, Samy Bengio, and Yoshua Bengio. Sharp minima can generalize for deep nets. In *International Conference on Machine Learning*, pp. 1019–1028. PMLR, 2017.

Felix Draxler, Kambis Veschgini, Manfred Salmhofer, and Fred Hamprecht. Essentially no barriers in neural network energy landscape. In Jennifer Dy and Andreas Krause (eds.), *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pp. 1309–1318. PMLR, 10–15 Jul 2018. URL https://proceedings.mlr.press/v80/draxler18a.html.

Jonathan Frankle, Gintare Karolina Dziugaite, Daniel M. Roy, and Michael Carbin. Linear mode connectivity and the lottery ticket hypothesis. In *Proceedings of the 37th International Conference on Machine Learning*, ICML'20. JMLR.org, 2020.

Jean-Christophe Gagnon-Audet, Ricardo Pio Monti, and David J. Schwab. AWE: Adaptive weight-space ensembling for few-shot fine-tuning. In *ICLR 2023 Workshop on Mathematical and Empirical Understanding of Foundation Models*, 2023. URL https://openreview.net/forum?id=rrMIPIboZL.

Timur Garipov, Pavel Izmailov, Dmitrii Podoprikhin, Dmitry P Vetrov, and Andrew G Wilson. Loss surfaces, mode connectivity, and fast ensembling of dnns. In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett (eds.), *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc., 2018. URL https://proceedings.neurips.cc/paper_files/paper/2018/file/be3087e74e9100d4bc4c6268cdbe8456-Paper.pdf.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition, 2015. URL https://arxiv.org/abs/1512.03385.

Pavel Izmailov, Dmitrii Podoprikhin, Timur Garipov, Dmitry P. Vetrov, and Andrew Gordon Wilson. Averaging weights leads to wider optima and better generalization. *CoRR*, abs/1803.05407, 2018. URL http://arxiv.org/abs/1803.05407.

Stanislaw Jastrzebski, Zachary Kenton, Devansh Arpit, Nicolas Ballas, Asja Fischer, Yoshua Bengio, and Amos Storkey. Finding flatter minima with sgd, 2018. URL https://openreview.net/forum?id=r1VF9dCUG.

Simran Kaur, Jeremy Cohen, and Zachary C. Lipton. On the maximum hessian eigenvalue and generalization, 2023. URL https://arxiv.org/abs/2206.10654.

Nitish Shirish Keskar, Dheevatsa Mudigere, Jorge Nocedal, Mikhail Smelyanskiy, and Ping Tak Peter Tang. On large-batch training for deep learning: Generalization gap and sharp minima. In *International Conference on Learning Representations*, 2017. URL https://openreview.net/forum?id=H1oyRlYgg.

Alex Krizhevsky. Learning multiple layers of features from tiny images. Technical report, 2009.

Kangqiao Liu, Liu Ziyin, and Masahito Ueda. Noise and fluctuation of finite learning rate stochastic gradient descent. In *International Conference on Machine Learning*, pp. 7045–7056. PMLR, 2021.

Stephan Mandt, Matthew D Hoffman, and David M Blei. Stochastic gradient descent as approximate bayesian inference. *Journal of Machine Learning Research*, 18(134):1–35, 2017.

Sidak Pal Singh and Martin Jaggi. Model fusion via optimal transport. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin (eds.), *Advances in Neural Information Processing Systems*, volume 33, pp. 22045–22055. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper_files/paper/2020/file/fb2697869f56484404c8ceee2985b01d-Paper.pdf.

Sidak Pal Singh, Linara Adilova, Michael Kamp, Asja Fischer, Bernhard Schölkopf, and Thomas Hofmann. Landscaping linear mode connectivity. In *High-dimensional Learning Dynamics 2024: The Emergence of Structure and Reasoning*, 2024. URL https://openreview.net/forum?id=OSNMqvPii6.

Samuel Smith, Erich Elsen, and Soham De. On the generalization benefit of noise in stochastic gradient descent. In *International Conference on Machine Learning*, pp. 9058–9067. PMLR, 2020.

Samuel L Smith and Quoc V Le. A bayesian perspective on generalization and stochastic gradient descent. In *International Conference on Learning Representations*, 2018.

Mingwei Wei and David J Schwab. How noise affects the hessian spectrum in overparameterized neural networks, 2019. URL https://arxiv.org/abs/1910.00195.

Mitchell Wortsman, Gabriel Ilharco, Jong Wook Kim, Mike Li, Simon Kornblith, Rebecca Roelofs, Raphael Gontijo-Lopes, Hannaneh Hajishirzi, Ali Farhadi, Hongseok Namkoong, and Ludwig Schmidt. Robust fine-tuning of zero-shot models. *arXiv preprint arXiv:2109.01903*, 2021. https://arxiv.org/abs/2109.01903.

Mitchell Wortsman, Gabriel Ilharco, Samir Ya Gadre, Rebecca Roelofs, Raphael Gontijo-Lopes, Ari S Morcos, Hongseok Namkoong, Ali Farhadi, Yair Carmon, Simon Kornblith, and Ludwig Schmidt. Model soups: averaging weights of multiple fine-tuned models improves accuracy without increasing inference time. 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. 23965–23998. PMLR, 17–23 Jul 2022. URL https://proceedings.mlr.press/v162/wortsman22a.html.

Zeke Xie, Issei Sato, and Masashi Sugiyama. A diffusion theory for deep learning dynamics: Stochastic gradient descent exponentially favors flat minima. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=wXgk_iCiYGo.

Zhewei Yao, Amir Gholami, Kurt Keutzer, and Michael W Mahoney. Pyhessian: Neural networks through the lens of the hessian. In *2020 IEEE international conference on big data (Big data)*, pp. 581–590. IEEE, 2020.

Sergey Zagoruyko and Nikos Komodakis. Wide residual networks. *CoRR*, abs/1605.07146, 2016. URL http://arxiv.org/abs/1605.07146.

Zhanpeng Zhou, Yongyi Yang, Xiaojiang Yang, Junchi Yan, and Wei Hu. Going beyond linear mode connectivity: The layerwise linear feature connectivity, 2023. URL https://arxiv.org/abs/2307.08286.

Liu Ziyin, Kangqiao Liu, Takashi Mori, and Masahito Ueda. Strength of minibatch noise in sgd. *arXiv preprint arXiv:2102.05375*, 2021.

A Appendix

A.1 k-step projected SGD

In order to directly measure the effect of entropic forces in a controlled setting, we use a modified version of SGD. Our algorithm deals with two conflicting considerations: First, we would like to limit the scope of our observation to models that lie on a linear path, or more generally models that lie on a MEP. However, we would also like to run optimization in such a way that entropic forces arising from curvature are still relevant to the optimization dynamics. The key observation is that if we were to run SGD on a line in parameter space, projecting back to the line after each optimization step, we would remove the effect of entropic forces, which arise from noisy multi-step optimization dynamics Wei & Schwab (2019). Motivated by this, we propose a natural algorithm that trades off between these two considerations by taking multiple SGD steps between before projecting the parameters back to the liner path (or MEP).

Algorithm 1 k-step projected SGD

Input: A model $f_\theta(x)$, a loss function $L(\theta, x, y)$, an integer k, path pivots $\theta_0, \theta_1, \ldots \theta_N$ SGD learning rate η, SGD batch size B.
while not converged **do**
  for $i = 1$ to k **do**
    Draw a batch $b \leftarrow \{x_j, y_j\}_{j=1}^{B} \sim D_{\text{train}}$
    $\theta \leftarrow \theta - \eta \nabla_\theta L(\theta, b)$
  end for
  Project θ on to the closest segment $\theta_n - \theta_{n+1}$.
end while

In this way, k trades off the effect of entropic forces (large k) vs how close to the linear, low-loss path optimization stays (small k). In Figure 2, we use $k = 4$, $\alpha = 1.6 \times 10^{-4}$, and run the algorithm on along the MEP 1-2.

A.2 The Fisher Trick for Estimating the Hessian

Computing the full Hessian of the training loss is intractable for modern neural networks due to both memory and runtime constraints. The Hessian matrix has $\mathcal{O}\left(N_p^2\right)$ parameters, where N_p is the number of parameters of the network, hence even just computing and storing the Hessian matrix is prohibitive. A common workaround is to exploit the equivalence between the Hessian of the loss and the Fisher Information Matrix (FIM) at a minimum.

If, as in the case of image classification, the loss is the negative log-likelihood,

$$\mathcal{L}(\theta) = -\sum_{(x,y)\in D} \log p_\theta(y \mid x).$$

At any parameter vector θ that minimizes the loss function $\mathcal{L}(\theta)$, we have that $\mathbb{E}_{p_\theta(y|x)}\left[\log(p_\theta(y|x))\right] = 0$, taking the derivative of this equation with respect to θ and using the log-derivative trick we have:

$$\mathbb{E}_{p_\theta(y|x)}\left[\nabla_\theta \log(p_\theta(y|x))\right] = \mathbb{E}_{p_\theta(y|x)}\left[\nabla_\theta^2 p_\theta(y|x)\right] - \mathbb{E}_{p_\theta(y|x)}\left[\nabla_\theta \log(p_\theta(y|x)) \nabla_\theta \log(p_\theta(y|x))\right] \tag{6}$$

Therefore at any minimum $\theta^\star$ of the loss, the Hessian of the loss coincides with the Fisher information matrix $\mathcal{F}(\theta^\star)$,

$$\mathcal{F}(\theta^\star) \equiv \mathbb{E}_{(x,y)\sim D}\left[\nabla_\theta \log p_\theta(y \mid x)\, \nabla_\theta \log p_\theta(y \mid x)^\top\right]\Big|_{\theta^\star}. \tag{7}$$

This identity allows us to approximate Hessian eigenvalues using stochastic estimates of the FIM. In practice