The Limits of large learning rates: A Case Study in Single Index Models

Bhavesh Kumar BKUMAR2@UW.EDU
Libin Zhu LIBINZHU@UW.EDU
University of Washington, Seattle

Abstract

Gradient descent methods with large learning rates have recently been shown to improve generalization in deep networks by enhancing feature learning and acting as an implicit regularizer. In this work, we present a contrasting case study in structured nonlinear models, focusing on the single-index and multi-index settings. Using the central flow framework, we analyze training dynamics in the Edge of Stability (EoS) regime, where iterates oscillate around sharpness thresholds. Our analysis reveals that in the single-index model, the loss and sharpness gradients are colinear; therefore, the central flow projects away the only valid descent direction, leading to stalled optimization. Numerical experiments confirm that SGD with large learning rates halts learning in this setting. These results highlight fundamental limitations of large learning rates in structured models, refining our understanding of EoS dynamics and feature learning.

1. Introduction

Stochastic Gradient Descent (SGD) [19] serves as a foundational algorithm for training neural networks, underpinning their success across diverse applications [10, 12, 13, 15–17]. It is widely valued for its scalability and effectiveness in optimizing high-dimensional, non-convex objectives. Recent work has shown that surprisingly large learning rates can often improve generalization, by enhancing feature learning [4, 9, 20, 22].

In this paper, we present a contrasting case study in the single-index model, where large learning rates fail to improve learning but instead impede it. Our analysis focuses on the edge of stability regime, which universally occurs in the training of neural networks with large learning rates [4], and models the training dynamics with the central flow framework [7]. Analysis from Section 4 shows that in the single index model, the top eigendirection of the Hessian aligns with the ground-truth direction. Therefore, central flow projects the parameters into an orthogonal subspace, effectively blocking progress toward the target. Experiments in Section 5 confirm this theoretical prediction for the single-index model.

Extending beyond the single-index setting, we show that in multi-index models, large learning rates do not completely stall training but instead slow it down, since the relevant subspace only partially aligns with the top Hessian direction. This effect is likewise verified experimentally in Section 5.

2. Related Work

Recent works in (stochastic) gradient descent have started to investigate the effects of large learning rates since empirical findings show that these larger learning rates lead to better generalization and induce feature learning. Among these phenomena is the so-called Edge of Stability (EoS) regime, characterized by training dynamics of GD where the maximum Hessian eigenvalue (sharpness) stabilizes near the critical stability threshold $2/\eta$ (for learning rate η). In this regime, the iterates oscillate persistently rather than converge smoothly, yet the loss continues to decrease gradually despite these oscillations [4]. This phenomenon challenges classical analyses of gradient-based methods [16, 17] and has inspired new theoretical work that deepens our understanding of optimization dynamics, loss landscape geometry, and implicit regularization effects [1, 2, 4, 14, 21]. Central flows provide a feasible approach to analyzing EoS by deriving time-averaged differential equations that capture the effective trajectory under persistent sharpness constraints, revealing how optimizers balance loss minimization with stability [7].

Multi-index models provide a flexible yet structured framework for representing complex nonlinear functions on high-dimensional input spaces. Specifically, a multi-index model expresses the target function as a nonlinear mapping applied to several linear projections of the input. This low-dimensional structure facilitates interpretability while allowing rich function classes to be modeled and analyzed [6, 8, 11, 18]. A simplified special case of this is the single-index model, where the function depends on only one latent direction. Single-index models serve as fundamental mathematical abstractions to study key challenges in nonlinear optimization, including non-convex loss landscapes and initialization sensitivity in high-dimensional learning settings [3, 5].

3. Preliminaries and Notation

In this paper, we consider the single-index model defined by the ground-truth function

$$f^*(x) = \sigma(\langle w^*, x\rangle),$$

where $x \sim \mathcal{N}(0, I_d)$ is a Gaussian input, $w^* \in \mathbb{S}^{d-1}$ is the ground-truth parameter vector on the unit sphere, and $\sigma : \mathbb{R} \to \mathbb{R}$ is a nonlinear activation function. This model can be seen as a single neuron in the first hidden layer of a neural network. Letting the label $y = f^*(x)$, we consider the population loss

$$L(w) = \mathbb{E}_{x,y}[1 - \sigma(\langle w, x\rangle)y].$$

The *information exponent* k^* of a function σ is defined via its Hermite expansion as the smallest integer $k \geq 1$ such that the k-th Hermite coefficient of σ is nonzero:

$$k^* := \min\{k \geq 1 : c_k \neq 0\}, \quad \text{where} \quad c_k := \mathbb{E}_{z\sim\mathcal{N}(0,1)}[\sigma(z)He_k(z)].$$

We make the following following assumption on the activation function:

Assumption 1 *The activation function σ has information exponent $k^* \geq 3$.*

For the unit sphere $\mathbb{S}^{d-1} := \{w \in \mathbb{R}^d : \|w\|_2 = 1\}$, the tangent space at w is defined as $T_w\mathbb{S}^{d-1} := \{v \in \mathbb{R}^d : w^\top v = 0\}$. For a C^1-smooth function $g : \mathbb{S}^{d-1} \to \mathbb{R}$, its Euclidean gradient $\nabla g(w)$ can be projected onto the tangent space to define the Riemannian gradient:

$$\operatorname{grad}^R g(w) = P_w^\perp \nabla g(w) = (I - ww^\top)\nabla g(w).$$

If g is twice differentiable, its Riemannian Hessian at w is the linear operator

$$\operatorname{Hess}^R g(w) := P_w^\perp \nabla^2 g(w) P_w^\perp - \langle w, \nabla g(w)\rangle P_w^\perp,$$

which acts on the tangent space $T_w\mathbb{S}^{d-1}$.

We denote the Riemannian gradient and Hessian of the population loss by $\operatorname{grad}^R L(w)$ and $\operatorname{Hess}^R L(w)$, respectively.

4. Main results

In this section, we analyze the Edge of Stability (EoS) dynamics on the unit sphere to reveal how EoS impedes learning in the single-index model. We begin by deriving the central flow dynamics restricted to the unit sphere, following the approach of [7] for the Euclidean setting.

4.1. Central Flow on the Sphere

We analyze the behavior of Riemannian gradient descent given by the discrete updates

$$w_{t+1} = \frac{w_t - \eta \operatorname{grad}^R L(w_t)}{\|w_t - \operatorname{grad}^R L(w_t)\|_2},$$

where the normalization ensures w_{t+1} remains on the sphere. Note that in the EoS regime, there exists an iteration t_0 where the Riemannian sharpness

$$S_R(w(t)) := \lambda_{\max}\big(\operatorname{Hess}^R L(w(t))\big)$$

stabilizes at the critical value $2/\eta$. For all $t \geq t_0$, the iterates cease to follow a smooth trajectory and instead undergo rapid oscillations due to the large learning rate.

Following [7], to characterize the effective optimization path, we consider the time-averaged trajectory $\bar{w}_t$ around which the iterates oscillate. We model each iterate as a small discrete perturbation from this average,

$$w_t = \bar{w}_t + (x_t u_t) + O(x_t^3),$$

where u_t is the top eigenvector of the Hessian $\operatorname{Hess}^R L(\bar{w}_t)$ and x_t is a scalar random variable with zero mean, $\mathbb{E}[x_t] = 0$, representing the oscillation magnitude in that direction. The variance of the oscillation is denoted $\sigma^2(t) := \mathbb{E}[x_t^2]$.

Expanding the Riemannian gradient around $\bar{w}_t$ with Taylor series then taking the expectation similarly to [7], we have

$$\mathbb{E}[\operatorname{grad}^R L(w_t)] = \operatorname{grad}^R L(\bar{w}_t) + \frac{\sigma^2(t)}{2} \operatorname{grad}^R S_R(\bar{w}_t) + O(\sigma^3(t)).$$

Hence, the continuous-time evolution of the time-averaged trajectory $\bar{w}(t)$ satisfies

$$\frac{d\bar{w}}{dt} = -\eta\left(\operatorname{grad}^R L(\bar{w}) + \frac{\sigma^2(t)}{2}\operatorname{grad}^R S_R(\bar{w})\right).$$

Similar to the derivations in [7], we can evaluate the oscillation variance by setting $\frac{dS_R(\bar{w})}{dt} = 0$, which stabilizes S_R to $\frac{2}{\eta}$. Evaluating for $\sigma^2(t)$ then yields

$$\sigma^2(t) = \frac{2\left\langle \operatorname{grad}^R S_R(\bar{w}), \operatorname{grad}^R L(\bar{w})\right\rangle}{\left\|\operatorname{grad}^R S_R(\bar{w})\right\|_2^2}.$$

Plugging this back yields the projected form of the central flow:

$$\frac{d\bar{w}}{dt} = -\eta \left(\operatorname{grad}^R L(\bar{w}) - \frac{\langle \operatorname{grad}^R S_R(\bar{w}), \operatorname{grad}^R L(\bar{w}) \rangle}{\left\| \operatorname{grad}^R S_R(\bar{w}) \right\|_2^2} \operatorname{grad}^R S_R(\bar{w}) \right),$$

which can be rewritten as

$$\frac{d\bar{w}}{dt} = -\eta \Pi^{\perp}_{\operatorname{grad}_R S_R(\bar{w})} \operatorname{grad}_R L(\bar{w}), \tag{1}$$

where $\Pi_v^{\perp}(g) := g - \frac{\langle v,g \rangle}{\|v\|^2} v$ denotes the orthogonal projection in the tangent space onto the complement of v.

4.2. Edge of Stability impedes learning in Single-index Models

With central flow (1) established, we now apply it to the single-index model. Using Hermite polynomial expansions of σ (Lemma 1), the Riemannian gradients take explicit forms:

$$\begin{aligned} \operatorname{grad}_R L(w) &= -\Pi_{T_w} w^* \sum_{k=1}^{\infty} \frac{c_k^2}{(k-1)!} \alpha^{k-1}, \\ \operatorname{grad}_R S_R(w) &= s'(\alpha) \Pi_{T_w} w^*, \end{aligned}$$

where $\alpha = \langle w, w^* \rangle$ and $s'(\alpha)$ is a scalar function determined by the Hermite coefficients. Notably, both gradients point in the same direction: parallel to $\Pi_{T_w} w^*$.

This collinearity has a significant impact on the central flow. When we project the loss gradient orthogonally to the sharpness gradient, the entire effective update is eliminated, leading to a stationary trajectory.

Theorem 1 (Riemannian EoS Prevents Convergence in Single-Index Models) *Suppose Assumption 1 holds. Under the Riemannian central flow dynamics equation* (1)

$$\frac{d\bar{w}}{dt} = -\eta \, \Pi^{\perp}_{\operatorname{grad}_R S_R(\bar{w})} \operatorname{grad}_R L(\bar{w})$$

where $S_R(\bar{w}) = \lambda_{\max}(\operatorname{Hess}_R L(\bar{w}))$ *is the Riemannian sharpness, the central flow dynamics become stationary:*

$$\frac{d\bar{w}}{dt} = 0$$

for all $t \geq 0$*, implying* $\bar{w}_t = \bar{w}_0$ *for all* $t \geq 0$.

Proof sketch. We highlight the key ideas of the proof of Theorem 1. The full proof is deferred to Appendix B. To begin, we analyze $\operatorname{grad}_R L(\bar{w})$ and $\operatorname{grad}_R S_R(\bar{w})$ from Lemma 1. We express $\operatorname{grad}_R L(\bar{w}) = \beta_L(\bar{w}) \cdot \Pi_{T_w} w^*$ and $\operatorname{grad}_R S_R(\bar{w}) = s'(\alpha) \cdot \Pi_{T_{\bar{w}}} w^*$, where β_L is a scalar coefficient and $s'(\alpha)$ is a scalar function determined by the Hermite coefficients. We then plug these values into the projection in the central flow equation (1):

$$\Pi^{\perp}_{\operatorname{grad}_R S_R(\bar{w})} \operatorname{grad}_R L(\bar{w}) = \beta_L(\bar{w}) \Pi_{T_{\bar{w}}} w^* - \frac{\beta_L(\bar{w})}{s'(\alpha)} \cdot s'(\alpha) \Pi_{T_{\bar{w}}} w^* = 0.$$

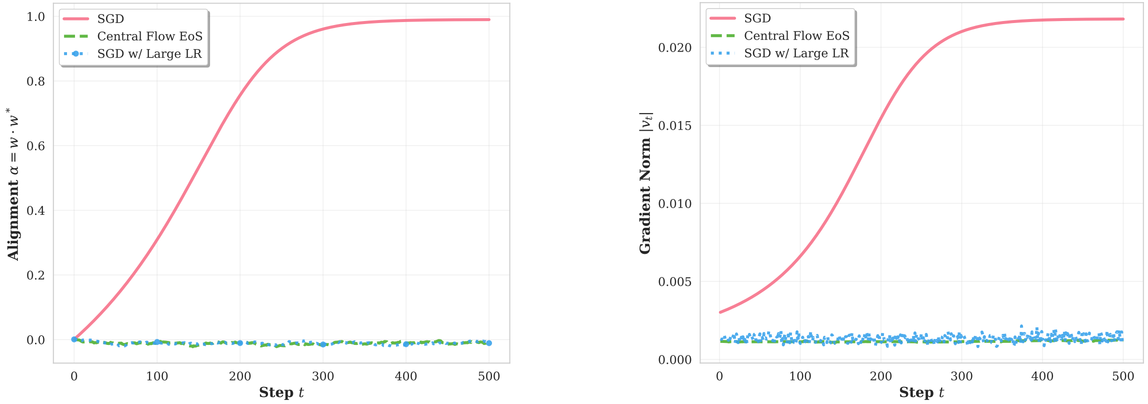
Thus, we have $\frac{d\bar{w}}{dt} = 0$, stalling the optimization. □

The stalling phenomenon in single-index models stems from a distinctive geometric structure: on the unit sphere, both the loss gradient and the sharpness gradient are colinear, each aligned with the tangential projection of the ground-truth w^*. As a result, when the EoS mechanism projects away from the sharpness direction to preserve stability, it simultaneously eliminates the only available descent direction for minimizing the loss. The optimizer is therefore left without a valid update direction, causing training to stall. This outcome is not an artifact of approximation but a direct consequence of the statistical geometry of single-index models.

5. Numerical results

In this section, we validate our theoretical findings through numerical experiments on single-index and multi-index models. These simulations compare three optimization approaches: (1) SGD with a small learning rate, (2) SGD with a large learning rate which naturally enters the EoS regime, and (3) the central flow dynamics. This comparison demonstrates that (1) the central flow accurately models the SGD with large learning rates, i.e., EoS regime, and (2) EoS induces stalling in the single-index setting, while in the multi-index setting, it slows but does not stall the optimization.

We track two metrics across training steps: (1) the alignment between the parameter w_t and the ground-truth w^* given by $\alpha = \langle w_t, w^* \rangle$ and (2) gradient norm $\|v_t\|_2$ where updates follow $w_{t+1} = -\eta v_t$. For the central flow (1), the gradient takes the form $v_t^{\text{CF}} = \Pi^{\perp}_{\text{grad}_R S_R(w)} \text{grad}_R L(w)$, whereas for SGD it is simply $v_t^{\text{SGD}} = \text{grad}_R L(w)$. Inputs are sampled as $X \sim \mathcal{N}(0, I_d)$ with sample size $n = 10^4$ and dimension $d = 500$. We use the smooth activation function $\sigma(z) = z^4 - 6z^2 + 3$. We set the learning rate to $\eta = 10^{-3}$ in the small–learning rate regime and $\eta = 1.0$ in the large–learning rate regime.



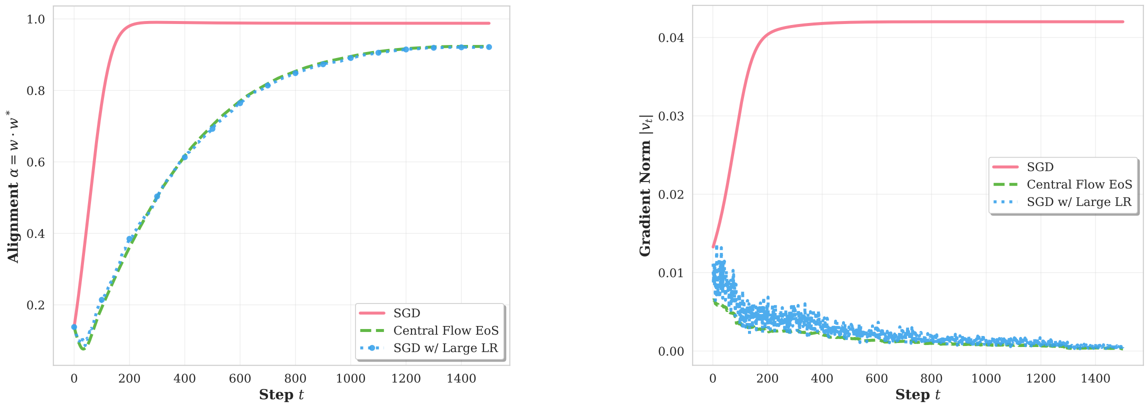
Alignment

Gradient Norm

Figure 1: Training dynamics in the single-index model.

In Figure 1, we observe that SGD with a small learning rate η (red) steadily increases the alignment α toward 1 while maintaining non-negligible gradient norms, thereby enabling effective learning. In contrast, both the central flow (green) and SGD with a large learning rate η (blue) lead to oscillations of α around zero, accompanied by vanishing gradient norms. This behavior is consistent with the stationary dynamics characterized in Theorem 1.

Next, we extend our experiments to multi-index models. Here, the label is $y = \sum_{i=1}^{m} \sigma(\langle w_i^*, x \rangle)$, with $m = 5$ orthonormal vectors $\{w_i^*\}_{i=1}^{m}$. We track subspace alignment $\alpha = \sqrt{\sum_{i=1}^{m} \cos^2(\theta_i)}$ (where θ_i is the principal angles between w and the subspace spanned by $\{w_i^*\}_{i=1}^{m}$) and the gradient norm.



Alignment

Gradient Norm

Figure 2: Training dynamics in the multi-index model.

Figure 2 shows that in multi-index models all methods improve the subspace alignment α, but with different patterns: SGD with a small η (red) converges rapidly, while central flow (green) and large-η SGD (blue) advance more slowly and plateau below the maximum, though without stalling. Their gradient norms remain non-zero, indicating sustained progress.

Unlike the single-index case, where EoS cancels the entire descent direction, the presence of multiple directions prevents perfect alignment between loss and sharpness gradients. As a result, EoS slows but does not stall the optimization.

6. Conclusion

This paper demonstrates that while large learning rates often benefit neural network training, their effects depend crucially on model structure. For the single-index model, the Edge of Stability regime eliminates all effective descent directions, causing optimization to stall completely. In multi-index models, where gradients only partially align, training continues but at a diminished rate. Together, these findings clarify when large learning rates act as a boon versus a hindrance, bridging recent empirical successes in deep learning with theoretical insights into structured models. More broadly, our results emphasize that the geometry of the loss and sharpness landscapes dictates the efficacy of large learning rates, suggesting future directions for designing adaptive optimization strategies that exploit EoS without succumbing to its limitations.

Acknowledgments

The research L.Z. was supported by NSF TRIPODS II, DMS-2023166.

References

[1] Kwangjun Ahn and Se-Young Yun. Understanding progressive sharpening in gradient descent. *arXiv preprint arXiv:2206.00012*, 2022.

[2] Sanjeev Arora, Wei Li, Georg Stadler, and Ruosong Zhang. Understanding gradient descent on the edge of stability in deep learning. In *International Conference on Learning Representations (ICLR)*, 2022. URL https://proceedings.mlr.press/v162/arora22a/arora22a.pdf.

[3] Gerard Ben Arous, Reza Gheissari, Jiaoyang Huang, and Aukosh Jagannath. Online stochastic gradient descent on non-convex losses from high-dimensional inference. *arXiv preprint arXiv:2002.10311*, 2020.

[4] Jeremy M. Cohen, Simran Kaur, Yuanzhi Li, J. Zico Kolter, and Ameet Talwalkar. Gradient descent on neural networks typically occurs at the edge of stability. *arXiv preprint arXiv:2103.00065*, 2021.

[5] Alex Damian, Eshaan Nichani, Rong Ge, and Jason D. Lee. Smoothing the landscape boosts the signal for sgd: Optimal sample complexity for learning single index models. In *Advances in Neural Information Processing Systems*, volume 36, page 72726. Curran Associates, Inc., 2023. URL https://papers.nips.cc/paper_files/paper/2023/hash/9f89de4c5ccfd11e1dd2eb5ec0bbd5e3-Paper-Conference.pdf. Oral presentation.

[6] Alex Damian, Jason D Lee, and Joan Bruna. The generative leap: Sharp sample complexity for efficiently learning gaussian multi-index models. *arXiv preprint arXiv:2506.05500*, 2024.

[7] Alex Damian, Eshaan Nichani, Yuheng Bu, Max Tegmark, and Gregory Berkolaiko. Understanding optimization in deep learning with central flows, 2024. URL https://arxiv.org/abs/2410.24206. arXiv preprint arXiv:2410.24206.

[8] Rishabh Dudeja and Daniel Hsu. Learning single-index models in gaussian space. *Conference on Learning Theory*, 2018.

[9] Jonathan Frankle, David J Schwab, and Ari S Morcos. The early phase of neural network training. *arXiv preprint arXiv:2002.10365*, 2020.

[10] Elad Hoffer, Itay Hubara, and Daniel Soudry. Train longer, generalize better: closing the generalization gap in large batch training of neural networks. *Advances in neural information processing systems*, 30, 2017.

[11] Zeljko Kereta, Timo Klock, and Valeriya Naumova. Nonlinear generalization of the monotone single index model. *arXiv preprint arXiv:1902.09024*, 2019.

[12] 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. *arXiv preprint arXiv:1609.04836*, 2016.

[13] Yann A LeCun, Leon Bottou, Genevieve B Orr, and Klaus-Robert Müller. Efficient backprop. *Neural networks: Tricks of the trade*, pages 9–48, 2012.

[14] Chao Ma and Lei Wu. Linear stability analysis of the edge of stability. *arXiv preprint arXiv:2109.04312*, 2021.

[15] Dominic Masters and Carlo Luschi. Revisiting small batch training for deep neural networks. *arXiv preprint arXiv:1804.07612*, 2018.

[16] Yurii Nesterov. *Introductory Lectures on Convex Optimization: A Basic Course*. Springer Science & Business Media, 2003.

[17] Boris T. Polyak. *Introduction to Optimization*. Optimization Software, 1987.

[18] Yunwei Ren and Jason D Lee. Learning orthogonal multi-index models: A fine-grained information exponent analysis. *arXiv preprint arXiv:2410.09678*, 2024.

[19] Herbert Robbins and Sutton Monro. A stochastic approximation method. *The annals of mathematical statistics*, pages 400–407, 1951.

[20] Leslie N Smith and Nicholay Topin. Super-convergence: Very fast training of neural networks using large learning rates. In *Artificial intelligence and machine learning for multi-domain operations applications*, volume 11006, pages 369–386. SPIE, 2019.

[21] L Wen, C Ma, W Li, and Y Li. Phase diagram of initial condensation for two-layer neural networks. *arXiv preprint arXiv:2305.03803*, 2023.

[22] Libin Zhu, Chaoyue Liu, Adityanarayanan Radhakrishnan, and Mikhail Belkin. Catapults in SGD: spikes in the training loss and their impact on generalization through feature learning. In *Forty-first International Conference on Machine Learning*, 2024. URL https://openreview.net/forum?id=jJmGl01S4l.

Appendix A. Supporting Lemmas

Lemma 1 (Riemannian Hermite Gradient Expansions) *Consider a single index model with activation function $\sigma : \mathbb{R} \to \mathbb{R}$ having information exponent $k^* \geq 3$, and population loss $L : \mathbb{S}^{d-1} \to \mathbb{R}$ given by*

$$L(w) = \mathbb{E}_{x,y}[1 - \sigma(\langle w, x\rangle)y]$$

where $x \sim \mathcal{N}(0, I_d)$, $y = \sigma(\langle w^*, x\rangle)$, and $w^* \in \mathbb{S}^{d-1}$. The Riemannian gradients admit the representations:

$$\operatorname{grad}^R L(w) = -\Pi_{T_w} w^* \sum_{k=1}^{\infty} \frac{c_k^2}{(k-1)!}\alpha^{k-1}, \tag{2}$$

$$\operatorname{grad}^R S_R(w) = s'(\alpha)\Pi_{T_w} w^*, \tag{3}$$

where $\alpha = \langle w, w^*\rangle$, c_k are the Hermite coefficients of σ, $S_R(w) = \lambda_{\max}(\operatorname{Hess}^R L(w))$ is the Riemannian sharpness, $\Pi_{T_w} = I - ww^T$ is the orthogonal projection onto the tangent space $T_w\mathbb{S}^{d-1}$, and $s'(\alpha)$ is a scalar function determined by the Hermite coefficients.

Proof The population loss on the sphere is

$$L(w) = \mathbb{E}_{x,y}[1 - \sigma(\langle w, x\rangle)y] = 1 - \mathbb{E}_x[\sigma(\langle w, x\rangle)\sigma(\langle w^*, x\rangle)].$$

Since σ has information exponent $k^* \geq 3$, it admits a Hermite polynomial expansion:

$$\sigma(z) = \sum_{k=0}^{\infty} \frac{c_k}{k!}\operatorname{He}_k(z),$$

where $c_k = \mathbb{E}_{z\sim\mathcal{N}(0,1)}[\sigma(z)\operatorname{He}_k(z)]$. By the orthogonality of Hermite polynomials, for $w, w^* \in \mathbb{S}^{d-1}$ with correlation $\alpha = \langle w, w^*\rangle$,

$$\mathbb{E}_x[\operatorname{He}_j(\langle w, x\rangle)\operatorname{He}_k(\langle w^*, x\rangle)] = \delta_{jk}k!\alpha^k.$$

Expanding the cross-correlation yields

$$\mathbb{E}_x[\sigma(\langle w, x\rangle)\sigma(\langle w^*, x\rangle)] = \sum_{k=0}^{\infty} \frac{c_k^2}{k!}\alpha^k,$$

so that

$$L(w) = 1 - \sum_{k=0}^{\infty} \frac{c_k^2}{k!}\alpha^k = \sum_{k=0}^{\infty} \frac{c_k^2}{k!}(1 - \alpha^k).$$

Since L depends only on $\alpha = \langle w, w^*\rangle$, we use the chain rule to obtain

$$\frac{\partial L}{\partial \alpha} = \sum_{k=1}^{\infty} \frac{c_k^2}{k!}(-k)\alpha^{k-1} = -\sum_{k=1}^{\infty} \frac{c_k^2}{(k-1)!}\alpha^{k-1} =: L'(\alpha).$$

The Riemannian gradient is then $\operatorname{grad}^R L(w) = \frac{\partial L}{\partial \alpha} \operatorname{grad}^R \alpha(w)$. Since $\alpha(w) = \langle w, w^* \rangle$ has Euclidean gradient $\nabla \alpha(w) = w^*$, the Riemannian gradient of α is $\operatorname{grad}^R \alpha(w) = \Pi_{T_w} \nabla \alpha(w) = \Pi_{T_w} w^*$. Therefore,

$$\operatorname{grad}^R L(w) = -\Pi_{T_w} w^* \sum_{k=1}^{\infty} \frac{c_k^2}{(k-1)!} \alpha^{k-1},$$

which establishes (2).

We next compute the Riemannian Hessian using the embedding formula for functions on the sphere:

$$\operatorname{Hess}^R L(w) = \Pi_{T_w} \nabla^2 L(w) \Pi_{T_w} - \langle w, \nabla L(w) \rangle \Pi_{T_w}.$$

In the ambient space $\mathbb{R}^d$, we have $\nabla L(w) = L'(\alpha) w^*$ and $\nabla^2 L(w) = L''(\alpha) w^* w^{*\top}$, where

$$L''(\alpha) = -\sum_{k=2}^{\infty} \frac{c_k^2}{(k-2)!} \alpha^{k-2}.$$

Let $v := \Pi_{T_w} w^*$ denote the tangential component of w^*. For any $u \in T_w \mathbb{S}^{d-1}$, since $u \perp w$ and $w^* = v + \alpha w$, we have $w^{*\top} u = v^\top u$. Thus $\nabla^2 L(w) u = L''(\alpha) w^* (v^\top u)$, and hence

$$\Pi_{T_w} \nabla^2 L(w) \Pi_{T_w} u = L''(\alpha) v (v^\top u).$$

The curvature term evaluates to $\langle w, \nabla L(w) \rangle = \langle w, L'(\alpha) w^* \rangle = L'(\alpha)\alpha$. Combining these, the Riemannian Hessian acts as

$$\operatorname{Hess}^R L(w)[u] = L''(\alpha) \langle v, u \rangle v - L'(\alpha)\alpha\, u.$$

The tangent space decomposes orthogonally as $T_w \mathbb{S}^{d-1} = \operatorname{span}\{v\} \oplus \{v\}^\perp$. On directions $u \in \{v\}^\perp$, the first term vanishes, yielding eigenvalue $\lambda_\perp(\alpha) = -L'(\alpha)\alpha$ with multiplicity $d-2$. On the direction $u \parallel v$, both terms contribute, yielding eigenvalue $\lambda_\parallel(\alpha) = L''(\alpha)\|v\|_2^2 - L'(\alpha)\alpha$. Therefore, the Riemannian sharpness is

$$S_R(w) = \lambda_{\max}(\operatorname{Hess}^R L(w)) = s(\alpha) := \max\{\lambda_\parallel(\alpha), \lambda_\perp(\alpha)\},$$

which depends only on α.

Since $S_R(w) = s(\alpha)$ for a scalar function $s : [-1, 1] \to \mathbb{R}$, wherever s is differentiable we apply the chain rule on the sphere to obtain $\operatorname{grad}^R S_R(w) = s'(\alpha) \operatorname{grad}^R \alpha(w) = s'(\alpha) \Pi_{T_w} w^*$, which establishes (3). The derivative $s'(\alpha)$ is determined by differentiating whichever eigenvalue is active in the maximum, both expressed via the Hermite coefficients through $L'(\alpha)$ and $L''(\alpha)$. ■

Appendix B. Main Theoretical Result

Theorem 1 (Riemannian EoS Prevents Convergence in Single Index Models) *Consider the single index model $f^*(x) = \sigma(\langle w^*, x \rangle)$ where $x \sim \mathcal{N}(0, I_d)$, $w^* \in \mathbb{S}^{d-1}$, and $\sigma : \mathbb{R} \to \mathbb{R}$ is a nonlinear activation function with information exponent $k^* \geq 3$. Let the population loss be*

$$L(w) = \mathbb{E}_{x,y}[1 - \sigma(\langle w, x \rangle) y],$$

where $y = \sigma(\langle w^*, x\rangle)$. Under the Riemannian central flow dynamics from (1)

$$\frac{d\bar{w}}{dt} = -\eta\,\Pi^{\perp}_{\operatorname{grad}^R S_R(\bar{w})} \operatorname{grad}^R L(\bar{w}),$$

where $S_R(\bar{w}) = \lambda_{\max}(\operatorname{Hess}^R L(\bar{w}))$ is the Riemannian sharpness, the central flow dynamics become stationary:

$$\frac{d\bar{w}}{dt} = 0,$$

for all $t \geq 0$, implying $\bar{w}_t = \bar{w}_0$ for all $t \geq 0$.

Proof From Lemma 1, the Riemannian gradients take the forms

$$\operatorname{grad}^R L(\bar{w}) = -\Pi_{T_{\bar{w}}} w^* \sum_{k=1}^{\infty} \frac{c_k^2}{(k-1)!}\alpha^{k-1},$$
$$\operatorname{grad}^R S_R(\bar{w}) = s'(\alpha)\Pi_{T_{\bar{w}}} w^*,$$

where $\alpha = \langle \bar{w}, w^*\rangle$ and $s'(\alpha)$ is a scalar function determined by the Hermite coefficients. The key observation is that both Riemannian gradients are parallel to $\Pi_{T_{\bar{w}}} w^*$, the tangential component of w^* at $\bar{w}$. We may therefore write

$$\operatorname{grad}^R L(\bar{w}) = \beta_L(\bar{w}) \cdot \Pi_{T_{\bar{w}}} w^*,$$
$$\operatorname{grad}^R S_R(\bar{w}) = s'(\alpha) \cdot \Pi_{T_{\bar{w}}} w^*,$$

where $\beta_L(\bar{w}) = -\sum_{k=1}^{\infty} \frac{c_k^2}{(k-1)!}\alpha^{k-1}$.

Applying the Riemannian projection from the EoS dynamics, we compute

$$\Pi^{\perp}_{\operatorname{grad}^R S_R(\bar{w})} \operatorname{grad}^R L(\bar{w}) = \operatorname{grad}^R L(\bar{w}) - \frac{\langle \operatorname{grad}^R S_R(\bar{w}), \operatorname{grad}^R L(\bar{w})\rangle_{\bar{w}}}{\|\operatorname{grad}^R S_R(\bar{w})\|_{\bar{w}}^2} \operatorname{grad}^R S_R(\bar{w}).$$

Since both gradients are parallel to $\Pi_{T_{\bar{w}}} w^*$, the inner product and norm evaluate to

$$\langle \operatorname{grad}^R S_R(\bar{w}), \operatorname{grad}^R L(\bar{w})\rangle_{\bar{w}} = s'(\alpha)\beta_L(\bar{w})\|\Pi_{T_{\bar{w}}} w^*\|_{\bar{w}}^2,$$

and

$$\|\operatorname{grad}^R S_R(\bar{w})\|_{\bar{w}}^2 = s'(\alpha)^2\|\Pi_{T_{\bar{w}}} w^*\|_{\bar{w}}^2.$$

Therefore,

$$\frac{\langle \operatorname{grad}^R S_R(\bar{w}), \operatorname{grad}^R L(\bar{w})\rangle_{\bar{w}}}{\|\operatorname{grad}^R S_R(\bar{w})\|_{\bar{w}}^2} = \frac{s'(\alpha)\beta_L(\bar{w})\|\Pi_{T_{\bar{w}}} w^*\|_{\bar{w}}^2}{s'(\alpha)^2\|\Pi_{T_{\bar{w}}} w^*\|_{\bar{w}}^2} = \frac{\beta_L(\bar{w})}{s'(\alpha)}.$$

Substituting back gives

$$\begin{aligned}\Pi^{\perp}_{\operatorname{grad}^R S_R(\bar{w})} \operatorname{grad}^R L(\bar{w}) &= \beta_L(\bar{w})\Pi_{T_{\bar{w}}} w^* - \frac{\beta_L(\bar{w})}{s'(\alpha)} \cdot s'(\alpha)\Pi_{T_{\bar{w}}} w^* \\ &= \beta_L(\bar{w})\Pi_{T_{\bar{w}}} w^* - \beta_L(\bar{w})\Pi_{T_{\bar{w}}} w^* \\ &= 0.\end{aligned}$$

Therefore,

$$\frac{d\bar{w}}{dt} = -\eta\,\Pi^{\perp}_{\operatorname{grad}^R S_R(\bar{w})} \operatorname{grad}^R L(\bar{w}) = -\eta \cdot 0 = 0,$$

showing that the Riemannian central flow is stationary with $\bar{w}_t = \bar{w}_0$ for all $t \geq 0$. ■