# On the Geometry of Memorization: Interpolation and Second-Order Representation Irregularity

**Satwik Bathula** [1]

## Abstract

We study memorization through the second-order geometry of learned representations. Using the pullback metric, we show that interpolating rapidly varying labels forces large second-order variation, linking interpolation to higher-order complexity. Empirically, both structured and random labels exhibit large magnitude, but differ in their spatial organization. Structured targets yield smooth, globally distributed variation, while random targets produce sparse, high-magnitude spikes. This distinction is captured by a curvature localization measure, showing that memorization is not characterized by the magnitude of second-order variation, but by its organization in representation space.

## 1. Introduction

Deep neural networks exhibit strong memorization, achieving near-zero training error even on random data (Zhang et al., 2017; Arpit et al., 2017; Feldman, 2020; Feldman & Zhang, 2020). This is characteristic of the modern interpolation regime, where overparameterized models fit data exactly while often still generalizing (Belkin et al., 2019; Hastie et al., 2022; Bartlett et al., 2020; Nakkiran et al., 2020). However, memorization is typically studied via error, loss, or optimization properties, leaving its manifestation in learned representations less understood.

We adopt a geometric perspective, viewing representations as inducing a geometry on the input domain via the pullback metric of a mapping

$$F_\theta : \mathcal{X} \to \mathbb{R}^d.$$

Rather than treating geometry as purely descriptive, we argue that it provides a structural characterization of memorization. In particular, interpolation of complex targets does not merely increase second-order variation, but forces distinct spatial organizations of curvature in the induced representation geometry. This perspective shifts the focus from the magnitude of complexity to its geometric organization, which we show is essential for distinguishing memorization regimes. Building on representation and information geometry (Amari, 2016; Poole et al., 2016; Raghu et al., 2017; Ansuini et al., 2019), and the spectral bias of neural networks (Rahaman et al., 2019; Xu et al., 2019), we analyze how this geometry evolves during training.

Fitting non-smooth labels induces large second-order variation in $F_\theta$, yielding rapid pullback-metric variation and linking interpolation to higher-order complexity. Empirically, decreasing training error coincides with increased second-order variation and degraded local structure, indicating a shift in how complexity is distributed.

**Contributions.**

- We provide a geometric framework for analyzing memorization through the pullback metric induced by learned representations.
- We establish that interpolation of rapidly varying targets imposes a *lower bound* on second-order variation: large curvature is not incidental but *unavoidable* under smooth representations.
- We show that memorization is not characterized by the magnitude of second-order variation, but by its *spatial organization*. While both structured and random targets induce large curvature, they differ fundamentally in how this curvature is distributed, which we quantify via a curvature localization measure.

Together, these results suggest that memorization can be understood through the distribution of second-order variation in learned representations, complementing traditional loss-based analyses. We define memorization as interpolation of rapidly varying labels under a smooth representation map. While memorization in large foundation models is often studied via outputs or loss, prior work analyzes it through membership inference and data extraction attacks,

[1]Department of Electrical and Computer Engineering, University of Southern California, Los Angeles, CA, USA. Correspondence to: Satwik Bathula <sbathula@usc.edu>.

*Published at ICML 2026 Workshop on the Impact of Memorization on Trustworthy Foundation Models*, Seoul, South Korea. PMLR 306, 2026. Copyright 2026 by the author(s).

whereas our results suggest it is reflected in the geometry of intermediate representations, where localized variation may indicate memorized instances. Our analysis relies only on smoothness and is independent of specific architectures, making it broadly applicable.

## 2. Related Work

Deep networks can memorize arbitrary labels (Zhang et al., 2017), with further analysis in (Arpit et al., 2017; Feldman, 2020; Feldman & Zhang, 2020). This behavior is closely tied to the interpolation regime (Belkin et al., 2019; Hastie et al., 2022; Bartlett et al., 2020; Nakkiran et al., 2020), where overparameterized models achieve near-zero training error while still generalizing. Implicit bias perspectives (Soudry et al., 2018; Jacot et al., 2018; Neyshabur et al., 2017; Dinh et al., 2017) explain aspects of this phenomenon via optimization dynamics.

Neural networks induce structured geometric transformations of data (Amari, 2016; Poole et al., 2016; Raghu et al., 2017; Ansuini et al., 2019), while generalization has also been linked to curvature and sharpness in parameter space (Keskar et al., 2017; Dinh et al., 2017). In contrast, we study second-order variation in representation space and provide a geometric characterization of memorization, linking interpolation to second-order complexity and showing that its spatial distribution distinguishes memorization regimes.

## 3. Geometric Setup

We model learned representations as a mapping

$$F_\theta : \mathcal{X} \to \mathbb{R}^d,$$

where $\mathcal{X}$ denotes the input space and $F_\theta(x)$ is the representation of input $x$. This mapping induces a data-dependent geometric structure on $\mathcal{X}$ via the pullback of the Euclidean metric.

**Induced Geometry.** The representation map $F_\theta$ defines a pullback metric

$$g(x) = J_F(x)^\top J_F(x),$$

where $J_F(x)$ is the Jacobian of $F_\theta$ at $x$. This metric captures how distances and local structure in the input space are transformed by the network.

**Second-Order Complexity.** To quantify local geometric variation, we consider the second-order behavior of $F_\theta$:

$$\mathcal{C}(\theta) = \mathbb{E}_{x\sim\mathcal{D}}\left[\|\nabla^2 F_\theta(x)\|_F\right],$$

where $\nabla^2 F_\theta(x)$ denotes the Hessian and $\|\cdot\|_F$ is the Frobenius norm. This serves as a proxy for curvature and captures local nonlinear distortion in the representation.

**Interpretation.** The quantity $\mathcal{C}(\theta)$ captures the magnitude of second-order variation but does not characterize how this variation is distributed across the input domain. To quantify spatial organization, define $v(x) := |g'(x)|$, $\tilde{v}(x) := v(x)/\mathbb{E}[v(x)]$, and

$$L(\theta) := \mathbb{E}[\tilde{v}(x)^2].$$

Values near 1 indicate uniformly distributed variation, while larger values indicate localization in sparse regions.

**Proposition 3.1** (Localization)**.** *Let $v(x) \geq 0$ and define*

$$L := \mathbb{E}\left[\left(\frac{v(x)}{\mathbb{E}[v(x)]}\right)^2\right].$$

*If $v$ is supported on a set of measure $\epsilon$, then*

$$L \geq \frac{1}{\epsilon},$$

*with equality $L = 1$ when $v$ is uniform.*

Thus, $L(\theta)$ quantifies the concentration of geometric variation, distinguishing uniform from localized regimes.

## 4. Interpolation and Geometric Irregularity

We now establish a theoretical connection between interpolation and geometric irregularity in learned representations.

**Theorem 4.1** (Rough-label interpolation forces second-order complexity)**.** *Let $F : \mathcal{X} \subset \mathbb{R}^m \to \mathbb{R}^d$ be $C^2$, and let $\gamma : [0,1] \to \mathcal{X}$ be a smooth path. Let $w \in \mathbb{R}^d \setminus \{0\}$ and define*

$$f(t) := \langle w, F(\gamma(t))\rangle.$$

*Let $x_1 < \cdots < x_n$ be points in $[0,1]$ with alternating labels $y_i = (-1)^i$. Suppose there exists $\gamma > 0$ such that*

$$y_i f(x_i) \geq \gamma, \quad \forall i.$$

*Let*

$$h := \max_i (x_{i+1} - x_i).$$

*Then there exists $t^* \in (0,1)$ such that*

$$\|F''(t^*)\| \geq \frac{2\gamma}{\|w\| h^2}.$$

*Moreover, if $h \leq C/n$ for some constant $C > 0$, then*

$$\sup_{t\in[0,1]} \|F''(t)\| \geq \frac{2\gamma}{C^2\|w\|} n^2.$$

**Proof.** See Appendix A.

*Intuition.* Rapidly varying labels force repeated sign changes in $\langle w, F(x)\rangle$, implying large first derivatives over

small intervals, and hence large second derivatives by smoothness. Thus, Theorem 4.1 shows that interpolating "rough" label patterns necessarily induces large second-order variation in the representation, linking interpolation to geometric irregularity.

**Lemma 4.2** (Second-order variation induces metric irregularity)**.** *Let $F : U \subset \mathbb{R}^m \to \mathbb{R}^d$ be $C^2$, and let $g = DF^\top DF$ be the induced pullback metric. Then*

$$\|\nabla g(x)\|_F \leq 2m^{3/2}\|DF(x)\|_{\mathrm{op}}\|D^2F(x)\|_{\mathrm{op}}.$$

**Proof.** See Appendix B.

The lemma shows that large second-order variation in $F$ induces rapid variation in the induced geometry. Combining this with Theorem 4.1 yields a geometric consequence of interpolation.

**Corollary 4.3** (Interpolation induces geometric irregularity)**.** *Under the assumptions of Theorem 4.1, let $t^*$ be a point satisfying*

$$\|F''(t^*)\| \geq \frac{2\gamma}{\|w\|h^2}.$$

*Assume further that there exist constants $\sigma > 0$ and $\eta > 0$ such that*

$$\|F'(t^*)\| \geq \sigma,$$
$$\left|\langle F''(t^*), F'(t^*)\rangle\right| \geq \eta\,\|F''(t^*)\|\,\|F'(t^*)\|.$$

*Let*

$$g(t) := \|F'(t)\|^2.$$

*Then*

$$|g'(t^*)| \geq \frac{4\eta\sigma\gamma}{\|w\|h^2}.$$

**Proof.** See Appendix C.

**Interpretation** Theorem 4.1 implies that interpolating rapidly varying labels forces *unavoidable* second-order variation, so any smooth representation must exhibit high second-order complexity, which by Lemma 4.2 and Corollary 4.3 induces rapid variation in the pullback geometry. Thus, curvature is necessary for interpolation but does not determine its spatial organization, which we study empirically and find depends on the structure of the target function.

## 5. Experiments

We empirically investigate how interpolation is realized geometrically in learned representations, and how this realization depends on the structure of the target function.

### 5.1. Setup

We consider a one-dimensional domain $x \in [0, 1]$ with $n = 80$ uniformly spaced points and two label regimes:

- **Structured:** $y_i = (-1)^i$
- **Random:** $y_i \sim \mathrm{Unif}\{-1, +1\}$

We train models

$$f(x) = w^\top F_\theta(x),$$

where $F_\theta : \mathbb{R} \to \mathbb{R}^d$ is a deep Tanh network with Fourier features, using mean squared error to exact interpolation.

We analyze:

$$\|F''(x)\|, \quad g(x) = \|F'(x)\|^2, \quad g'(x) = 2\langle F''(x), F'(x)\rangle$$

computed via automatic differentiation.

### 5.2. Interpolation Regime

In both label settings, the model achieves zero training error and perfect classification accuracy. The margin satisfies

$$\min_i y_i f(x_i) > 0,$$

ensuring that the conditions of Theorem 4.1 hold.

### 5.3. Second-Order Complexity

Across both regimes, $\|F''(x)\|$ increases during training and stabilizes at large values, consistent with the results of Section 3.

### 5.4. Geometric Realization of Complexity

While both regimes exhibit large second-order complexity, their spatial organization differs markedly.

**Structured labels.** Alternating labels induce smooth but highly nonlinear representations with globally distributed metric variation $|g'(x)|$ (Figure 1).

**Random labels.** In contrast, random labels yield near-linear representations with localized perturbations, where $|g'(x)|$ exhibits sharp spikes in narrow regions (Figure 2).

To quantify this distinction, we compute the localization score $L(\theta)$. While both regimes have comparable second-order magnitude, structured labels yield small values of $L(\theta)$, whereas random labels produce approximately an order-of-magnitude larger values, indicating concentration of curvature in sparse regions.

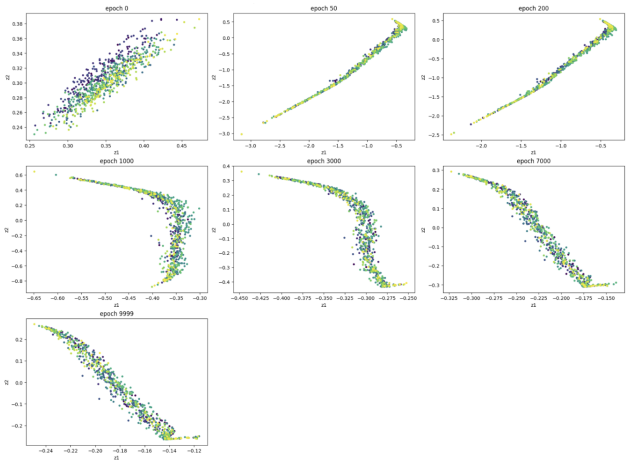


*Figure 1.* Structured labels induce smooth, globally distributed second-order variation (low $L(\theta)$).

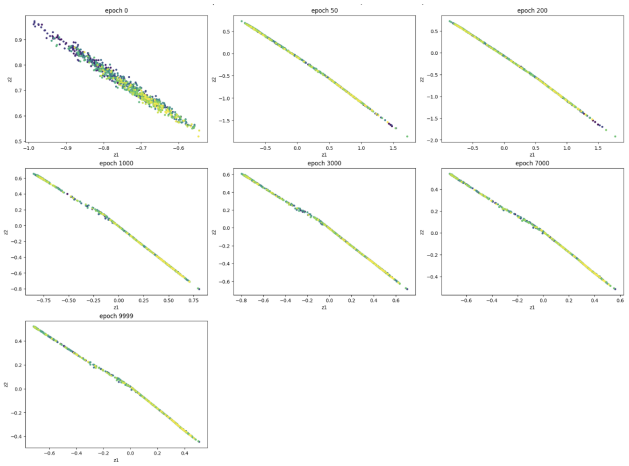


*Figure 2.* Random labels yield localized second-order variation with sharp spikes in $|g'(x)|$ (high $L(\theta)$).

**Real data with label corruption.** We evaluate curvature localization on MNIST under varying levels of label corruption. While all models achieve near-zero training error, the organization of second-order variation differs across regimes. As shown in Figure 3 and Table 1, the localization score decreases with increasing corruption ($L(\theta) = 1.83$ for clean data vs. $1.51$ and $1.40$ under 20% and 50% corruption), indicating a shift from more concentrated to more uniformly distributed geometric variation.

| Noise Level | $L(\theta)$ |
|---|---|
| 0% | 1.83 |
| 20% | 1.51 |
| 50% | 1.40 |

*Table 1.* Curvature localization $L(\theta)$ on MNIST under varying label corruption.

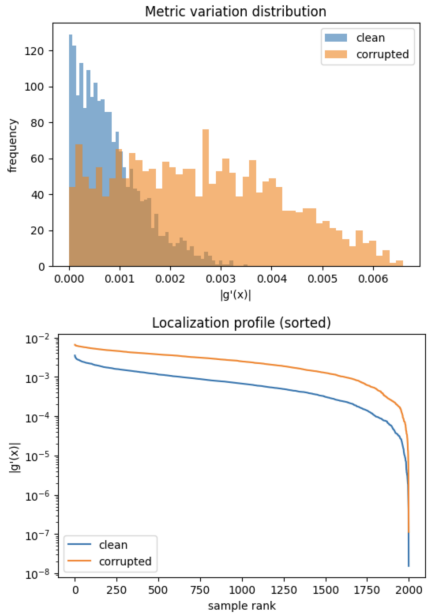


*Figure 3.* Metric variation on MNIST under label corruption. Corruption yields broader $|g'(x)|$ but lower $L(\theta)$, indicating more uniform geometric variation.

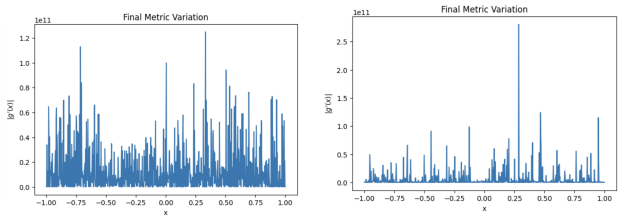


*Figure 4.* Metric variation $|g'(x)|$: structured labels show distributed curvature, while random labels exhibit localized spikes (low vs high $L(\theta)$).

### 5.5. Representation vs Output Complexity

The output $f(x)$ remains simple, while $F(x)$ carries the complexity required for interpolation, indicating that memorization is reflected not in the magnitude of complexity, but in how geometric complexity is organized within the representation. Figure 4 illustrates the contrast between globally distributed curvature and localized spikes.

## 6. Conclusion

We show that memorization is better understood through the geometric organization of representations rather than scalar complexity. While interpolation induces large second-order variation, its spatial distribution distinguishes regimes, with structured targets yielding smoothly distributed variation and random or corrupted labels producing more localized patterns. Thus, memorization is governed not by complexity magnitude but by its organization in representation space.

## Impact Statement

This work develops a geometric perspective on memorization in neural networks, showing how target structure shapes the organization of learned representations. While primarily theoretical, these insights may inform the design of models with improved control over memorization and generalization. Understanding how complexity is encoded at the representation level could also aid in identifying and mitigating unintended memorization of sensitive or proprietary data.

The work has no direct negative societal impact. However, as with many advances in understanding model behavior, it may indirectly influence systems deployed in high-stakes settings. Care should therefore be taken in applications involving personal data, where representation-level memorization could have privacy implications.

## References


Amari, S.-i. *Information Geometry and Its Applications*. Springer, 2016.

Ansuini, A., Laio, A., Macke, J. H., and Zoccolan, D. Intrinsic dimension of data representations in deep neural networks. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2019.

Arpit, D., Jastrzebski, S., Ballas, N., Krueger, D., Bengio, E., Kanwal, M. S., Maharaj, T., Fischer, A., Courville, A., Bengio, Y., and Lacoste-Julien, S. A closer look at memorization in deep networks. In Precup, D. and Teh, Y. W. (eds.), *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pp. 233–242. PMLR, 06–11 Aug 2017. URL https://proceedings.mlr.press/v70/arpit17a.html.

Bartlett, P. L., Long, P. M., Lugosi, G., and Tsigler, A. Benign overfitting in linear regression. *Proceedings of the National Academy of Sciences (PNAS)*, 2020.

Belkin, M., Hsu, D., Ma, S., and Mandal, S. Reconciling modern machine learning practice and the classical bias-variance trade-off. In *PNAS*, 2019.

Dinh, L., Pascanu, R., Bengio, S., and Bengio, Y. Sharp minima can generalize for deep nets. In *International Conference on Machine Learning (ICML)*, 2017.

Feldman, V. Does learning require memorization? a short tale about a long tail. In *Proceedings of the ACM Symposium on Theory of Computing (STOC)*, 2020.

Feldman, V. and Zhang, C. What neural networks memorize and why: Discovering the long tail via influence estimation. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2020.

Hastie, T., Montanari, A., Rosset, S., and Tibshirani, R. Surprises in high-dimensional ridgeless least squares interpolation. *Annals of Statistics*, 2022.

Jacot, A., Gabriel, F., and Hongler, C. Neural tangent kernel: Convergence and generalization in neural networks. In *NeurIPS*, 2018.

Keskar, N. S. et al. On large-batch training for deep learning: Generalization gap and sharp minima. In *ICLR*, 2017.

Nakkiran, P., Kaplun, G., Bansal, Y., Yang, T., Barak, B., and Sutskever, I. Deep double descent: Where bigger models and more data hurt. In *International Conference on Learning Representations (ICLR)*, 2020.

Neyshabur, B., Bhojanapalli, S., McAllester, D., and Srebro, N. Exploring generalization in deep learning. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2017.

Poole, B., Lahiri, S., Raghu, M., Sohl-Dickstein, J., and Ganguli, S. Exponential expressivity in deep neural networks through transient chaos. *NeurIPS*, 2016.

Raghu, M., Poole, B., Kleinberg, J., Ganguli, S., and Sohl-Dickstein, J. On the expressive power of deep neural networks. In *International Conference on Machine Learning (ICML)*, 2017.

Rahaman, N., Baratin, A., Arpit, D., Draxler, F., Lin, M., Hamprecht, F., Bengio, Y., and Courville, A. On the spectral bias of neural networks, 2019. URL https://openreview.net/forum?id=r1gR2sC9FX.

Soudry, D., Hoffer, E., Nacson, M. S., Gunasekar, S., and Srebro, N. The implicit bias of gradient descent on separable data. In *ICLR*, 2018.

Xu, Z.-Q. J., Zhang, Y., Luo, T., Xiao, Y., and Ma, Z. Frequency principle: Fourier analysis sheds light on deep neural networks. *arXiv preprint arXiv:1901.06523*, 2019.

Zhang, C., Bengio, S., Hardt, M., Recht, B., and Vinyals, O. Understanding deep learning requires rethinking generalization. In *ICLR*, 2017.

# Appendix

## Proofs

### A. Proof of Theorem 4.1

*Proof.* Define $\tilde{F}(t) := F(\gamma(t))$. Then $\tilde{F} \in C^2([0,1], \mathbb{R}^d)$, and the result follows by applying the one-dimensional argument to $\tilde{F}$.

For each $i = 1, \ldots, n-1$, define

$$\Delta_i := x_{i+1} - x_i.$$

Then $\Delta_i > 0$ and $\Delta_i \leq h$.

Since $y_i = (-1)^i$ and $y_i f(x_i) \geq \gamma$, we have

$$(-1)^i f(x_i) \geq \gamma, \quad (-1)^{i+1} f(x_{i+1}) \geq \gamma.$$

Multiplying the second inequality by $-1$ yields

$$(-1)^i f(x_{i+1}) \leq -\gamma.$$

Subtracting,

$$(-1)^i \big(f(x_i) - f(x_{i+1})\big) \geq 2\gamma,$$

so

$$|f(x_{i+1}) - f(x_i)| \geq 2\gamma.$$

By the mean value theorem, for each $i$ there exists $c_i \in (x_i, x_{i+1})$ such that

$$f'(c_i) = \frac{f(x_{i+1}) - f(x_i)}{\Delta_i}.$$

Thus,

$$|f'(c_i)| \geq \frac{2\gamma}{\Delta_i} \geq \frac{2\gamma}{h}.$$

Because the signs of $f(x_{i+1}) - f(x_i)$ alternate, the signs of $f'(c_i)$ also alternate. Hence for each $i = 1, \ldots, n-2$,

$$|f'(c_{i+1}) - f'(c_i)| = |f'(c_{i+1})| + |f'(c_i)| \geq \frac{4\gamma}{h}.$$

Now $c_i \in (x_i, x_{i+1})$ and $c_{i+1} \in (x_{i+1}, x_{i+2})$, so

$$0 < c_{i+1} - c_i < x_{i+2} - x_i \leq 2h.$$

Applying the mean value theorem to $f'$ on $[c_i, c_{i+1}]$, there exists $t^* \in (c_i, c_{i+1})$ such that

$$f''(t^*) = \frac{f'(c_{i+1}) - f'(c_i)}{c_{i+1} - c_i}.$$

Therefore,

$$|f''(t^*)| \geq \frac{4\gamma/h}{2h} = \frac{2\gamma}{h^2}.$$

Since $f''(t) = \langle w, F''(t) \rangle$, we have

$$|f''(t)| \leq \|w\| \, \|F''(t)\|.$$

Thus,

$$\|F''(t^*)\| \geq \frac{2\gamma}{\|w\| h^2}.$$

If $h \leq C/n$, then $h^{-2} \geq n^2/C^2$, yielding

$$\sup_{t \in [0,1]} \|F''(t)\| \geq \frac{2\gamma}{C^2 \|w\|} n^2.$$

□

## B. Proof of Lemma 4.2

*Proof.* Let $F = (F^1, \ldots, F^d)$. Then

$$g_{ij}(x) = \sum_{\alpha=1}^{d} \partial_i F^\alpha(x)\, \partial_j F^\alpha(x).$$

Differentiating with respect to $x^k$ gives

$$\partial_k g_{ij}(x) = \sum_{\alpha=1}^{d} \partial_{ki} F^\alpha(x)\, \partial_j F^\alpha(x) + \sum_{\alpha=1}^{d} \partial_i F^\alpha(x)\, \partial_{kj} F^\alpha(x).$$

This can be written as

$$\partial_k g_{ij}(x) = \langle \partial_{ki} F(x), \partial_j F(x) \rangle + \langle \partial_i F(x), \partial_{kj} F(x) \rangle.$$

Applying Cauchy–Schwarz,

$$|\partial_k g_{ij}(x)| \le \|\partial_{ki} F(x)\|\, \|\partial_j F(x)\| + \|\partial_i F(x)\|\, \|\partial_{kj} F(x)\|.$$

Using operator norm bounds,

$$\|\partial_j F(x)\| \le \|DF(x)\|_{\mathrm{op}}, \quad \|\partial_{ki} F(x)\| \le \|D^2 F(x)\|_{\mathrm{op}},$$

and similarly for the other terms, we obtain

$$|\partial_k g_{ij}(x)| \le 2\, \|DF(x)\|_{\mathrm{op}}\, \|D^2 F(x)\|_{\mathrm{op}}.$$

Now summing over $i, j, k$,

$$\|\nabla g(x)\|_F^2 = \sum_{i,j,k=1}^{m} |\partial_k g_{ij}(x)|^2 \le m^3 \cdot \left(2\, \|DF(x)\|_{\mathrm{op}}\, \|D^2 F(x)\|_{\mathrm{op}}\right)^2.$$

Taking square roots yields

$$\|\nabla g(x)\|_F \le 2m^{3/2} \|DF(x)\|_{\mathrm{op}} \|D^2 F(x)\|_{\mathrm{op}}.$$

□

## C. Proof of Corollary 4.3

*Proof.* In one dimension,

$$g(t) = \|F'(t)\|^2 = \langle F'(t), F'(t) \rangle.$$

Differentiating,

$$g'(t) = 2\langle F''(t), F'(t) \rangle.$$

Thus,

$$|g'(t^*)| = 2\, |\langle F''(t^*), F'(t^*) \rangle|.$$

By the alignment assumption,

$$|\langle F''(t^*), F'(t^*) \rangle| \ge \eta \|F''(t^*)\| \|F'(t^*)\|.$$

Hence,

$$|g'(t^*)| \ge 2\eta \|F''(t^*)\| \|F'(t^*)\|.$$

Using $\|F'(t^*)\| \ge \sigma$ and the bound from Theorem 4.1,

$$\|F''(t^*)\| \ge \frac{2\gamma}{\|w\| h^2},$$

we obtain

$$|g'(t^*)| \geq 2\eta \left( \frac{2\gamma}{\|w\| h^2} \right) \sigma = \frac{4\eta\sigma\gamma}{\|w\| h^2}.$$

□