0.1 Hypothesis: Trained Models as Approximate Riemannian Submersions

Let $\mathcal{X}$ denote the input space (e.g., token embeddings), and let $F = \pi \circ f : \mathcal{X} \to \Delta^{k-1}$ be the full mapping from inputs to output probability distributions via a trained neural model, where π denotes the softmax map and Δ^{k-1} is the $(k-1)$-dimensional probability simplex. Let g_{FR} denote the Fisher-Rao metric on Δ^{k-1}.

[Pullback Fisher-Rao Metric] The pullback metric on $\mathcal{X}$ is defined as

$$g_{\mathcal{X}}(x) := F^* g_{\mathrm{FR}} = J_F(x)^\top \cdot G_{\mathrm{FR}}(F(x)) \cdot J_F(x),$$

where $J_F(x) \in R^{k \times n}$ is the Jacobian of F at x, and G_{FR} is the Fisher-Rao metric tensor at $F(x) \in \Delta^{k-1}$.

[Trained Models Approximate Riemannian Submersions] Let $F = \pi \circ f : \mathcal{X} \to \Delta^{k-1}$ be a model trained to minimize a divergence loss (e.g., cross-entropy). Then in regions of semantic regularity, F approximates a Riemannian submersion in the following sense:

- The Jacobian $J_F(x)$ is approximately of constant rank $r \ll n$ in local neighborhoods.
- The pullback metric $g_{\mathcal{X}}(x)$ is non-degenerate on a horizontal distribution $\mathcal{H}_x \subset T_x\mathcal{X}$ of dimension r.
- Directions in the kernel of $J_F(x)$ (the vertical space $\mathcal{V}_x$) correspond to semantically invariant perturbations (i.e., the model's output distribution is approximately constant along those directions).

Interpretation. This conjecture proposes that trained models learn to compress the input space into a lower-dimensional semantic manifold embedded in output distribution space, preserving Fisher-Rao distances along directions relevant to prediction, and flattening irrelevant variations.

Empirical Testing Protocol

To test the conjecture, the following steps may be applied:

1. **Compute the model output:** Evaluate $F(x)$ for various input points $x \in \mathcal{X}$.
2. **Estimate the Jacobian:** Numerically compute $J_F(x) \in R^{k \times n}$ using automatic differentiation.
3. **Estimate the rank:** Perform singular value decomposition (SVD) on $J_F(x)$ to estimate the effective rank and determine the horizontal space dimension r.

4. **Compute pullback metric:** Construct $g_{\mathcal{X}}(x)$ via

$$g_{\mathcal{X}}(x) = J_F(x)^\top \cdot G_{\mathrm{FR}}(F(x)) \cdot J_F(x),$$

where G_{FR} is diagonal with entries $1/p_i$ for $p = F(x)$.

5. **Analyze geometry:** Study curvature, geodesics, and degeneracy patterns in $g_{\mathcal{X}}$ to determine whether the model behaves like a submersion locally.

Conclusion. If verified, this hypothesis supports the idea that LLMs construct a semantic manifold via compression and projection, where the pullback of the Fisher-Rao geometry provides a natural Riemannian structure on the input space.