On the *Mechanistic Stability* of Language Models

Alan Sun, Ethan Sun, Warren Shepard

awsun@cmu.edu, {ethan.k.sun.26, warren.a.shepard}@dartmouth.edu

January 1, 2025

Herein, we introduce a notion which we call *mechanistic stability* which characterizes the sensitivity of a models' prediction criteria with respect to perturbations of the input data.

1 What are we doing?

For any task T, we care about whether a model is mechanistically stable on T. That is, is it using the same causal graph to perform inference? The main result that we want to prove in our paper is that mechanistic stability is good for the model in terms of both generalization and robustness. Throughout our paper, we assume the supervised learning setting.

We denote $\mathcal{X}, \mathcal{Y}$ to the universe of all possible input and outputs. Then, a task is simply a distribution over the product of $\mathcal{X}, \mathcal{Y}$. Concretely,

Definition 1.1 (Task) *A task T is a distribution over $\mathcal{X} \times \mathcal{Y}$. We denote this distribution as $\mathcal{D}_{\mathcal{X}\times\mathcal{Y}}$. In the literature, this is also called the data distribution.*

Stability is a measure of change in "output" versus changes in "input." Here, the terms input/output are used loosely. Mechanistic stability is when the output is the mechanism of our model and where the input is across partitions of $\mathcal{X} \times \mathcal{Y}$ according to

Definition 1.2 (Permissible Partition) *For a given task T with distribution $\mathcal{D}_{\mathcal{X}\times\mathcal{Y}}$, a T-partition, denoted by $\mathcal{S}$, is a countable collection of measurable subsets of $\mathcal{X} \times \mathcal{Y}$ such that $\coprod_{s\in\mathcal{S}} s = \mathcal{X} \times \mathcal{Y}$ and $\mathbb{P}_{\mathcal{D}_{\mathcal{X}\times\mathcal{Y}}}[s] > 0$ for all $s \in \mathcal{S}$.*

Now that we have a very loose notion of what the "input" looks like, we need to measure what big changes in our output corresponds to. I'm not too sure exactly what this looks like yet and we probability need theory from the causality and category theory papers.

After we perform mechanism extraction, we need to assume that by virtue of this being the mechanism, this is essentially the function that we care about. So, if two mechanism are exactly the same, then their functions must be the same. Is this the case in the category-theoretic sense? In other words, if two causal models are equivalence in the category- theoretic sense, how different can they be in terms of their output?

Trivially, mechanistic equivalence should give us robustness. The main issue that we care about here is generalization. In the sense that

Definition 1.3 *Let $\rho(\cdot, \cdot)$ denote a metric on the set of all causal graphs with respect to a distribution $\mathcal{D}_{\mathcal{X}\times\mathcal{Y}}$.*

At the core of our paper, we are making the following assumption. This is common for many papers in mechanistic interpretability:

Assumption 1.4 *The extracted mechanisms of a neural network is its causal graph.*

Now, we are ready to define mechanistic stability. Let M be a model that represents a stochastic mapping from $\mathcal{X} \to \mathcal{Y}$. From now on, we shall slightly abuse notation and refer to both the model and its mechanism as M. It should be clear from context what exactly we mean.

Definition 1.5 *We say that M is ε-mechanistically stable on a task T if*

$$\sup_{\mathcal{S}:T\text{-partition}} \mathbb{E}_{p_1,p_2\in\mathcal{S}}\rho(M_{p_1}, M_{p_2}) < \varepsilon,$$

where M_{p_i} is the mechanism of M when restricted to the truncated distribution over p_i.

At this point, there are still some questions that need to be resolved. Firstly, how do we define $\rho(\cdot, \cdot)$ and does it need to be an actual metric, for example maybe we can only determine equivalence up to some isomorphism? Secondly, does this definition of all possible permissible partitions make sense? Is it too weak? That is, are there pathological partitions across whose stability we do not care about?

2 Main results

Herein, is a wish list of the main results that we wish to achieve.

Theorem 2.1 *Mechanistic stability $\Rightarrow$ in-distribution robustness, assuming that both $\mathcal{X}, \mathcal{Y}$ are metric spaces. We also need that the task distribution $\mathcal{D}_{\mathcal{X}\times\mathcal{Y}}$ is smooth.*

Theorem 2.2 *Mechanistic stability $\Rightarrow$ in-distribution generalization.*

Maybe for proving this latter theorem, we can look at some proofs relating regularization. What is a notion of in-distribution generalization that makes sense to look at here. Can we say something stronger maybe? What about out-of-distribution? like distribution shift?

3 Casual Equivalence: A Category-Theoretic Perspective

Herein, we describe briefly how to understand the equivalence between two mechanisms. Our fundamental assumption is that the extracted mechanism of a model is a causal graph. A causal graph is a functor from $F : \mathsf{Syn} \to \mathsf{Stoch}$

4 Related Literature

What is the relationship between our notion of mechanistic stability and the more general notion of algorithmic stability? Is it potentially the same as algorithmic stability applied to in-context learning?
Algorithmic stability. What is the relationship between algorithmic stability and the notion of mechanistic stability that we are defining, both in terms of the generalization bounds that are guaranteed with algorithmic stability and in terms of the concept itself (how are we measuring the distance between two learned hypotheses?)
Mechanistic interpretability. Recently, there has a been a push to interpret neural networks by uncovering their *mechanisms*. A mechanism of a neural network with respect to some task is a minimal subgraph of its computational graph that wholly characterizes the network's behavior on this task (Wang et al., 2022). After exposing the driving mechanisms
A brief introduction to the field and how our work provides rigorous guarantees for a lot of the work being done in MI.
Graphical models. Probably need to rethink the title here, we want to explain the relationship between directed acyclic graphs, causal graphs, and the mechanisms that we are extracting from the neural network. Our underlying assumption is that all of these things are the same.

5 Experimental Methods

Can we show mechanistic stability and instability on a host of tasks?

1. IOI
2. Colored Objects
3. Arithmetic
4. General algorithmic tasks
5. General reasoning tasks (this and the previous task would benefit from increased test time compute and enhanced supervision; so we can we somehow increase stability by increasing one of these factors?)
6. General knowledge tasks (would not benefit from chain-of-thought or more test time compute)

What is the effect of scale on mechanistic stability? Or conversely, the benefits that we see in terms of performance as a result of increased scale, does this come from increased mechanistic stability?

References

Kevin Ro Wang, Alexandre Variengien, Arthur Conmy, Buck Shlegeris, and Jacob Steinhardt. Interpretability in the Wild: a Circuit for Indirect Object Identification in GPT-2 Small. In *The Eleventh International Conference on Learning Representations*, September 2022. URL https://openreview.net/forum?id=NpsVSN6o4ul.

6 Category-Theoretic Causality