Open Problems / Questions

Open Problems / Questions

- One of the things that makes SLT interesting is the potential for feedback between **experimental progress** and **theoretical progress**. A principal aim of this meeting is to kick that flywheel into motion.
- Good researchers operate in alternating phases of narrow and broad attention. There are many areas related to SLT that we care about (e.g. goal non-identifiability, degeneracy broadly interpreted, other areas of deep learning theory, …).
- **Phases and phase transitions** is an area in which we are keen to see near-term theoretical work.
- The emergence of **internal structure** in learning machines, and the connection to phase transitions, is an area in which we are keen to see near-term experimental work.

Open Problems / Questions

Mechanics

- Discussions and working groups are a key organising frame for this meeting.
- Today we write up the **beginning** of a list of open problems / questions.
- Volunteer to be an **advocate** (read: water carrier, not Owner / Boss) for an open problem, to (i) save it from being ignored (ii) be a point of call for questions and updating curious people about progress (iii) coordinate summary / presentation on Friday.
- **Intended outcomes:** ongoing research collaborations, papers, blog posts, code repositories.
- **Possible follow-up workshop** in November, in Europe.

Open Problems / Questions

Mechanics

- In a group, ways to get started: **read** some papers, **code** a minimal thing, **calculate** a minimal thing.
- Engage with details as soon as possible, go from there.
- It's easy to see reasons why something might not work. But nitpickers make for poor collaborators (so do credulous people). Try to be neither.
- Be respectful.
- Teach.

Dev Interp

The Construction Is The Program

- **Key Questions**: when do structures form? where do they form? how do existing structures constrain new ones?
- Let's fail fast, if we can: here are the easy ways to fail
    - **Too few:** Only a small fraction of structure forms in phase transitions
    - **Too many:** Phase transitions are too common, no way to triage them
    - **Too big**: transitions are irreducibly complex, not much better than interpreting a whole network.
    - **Too diffuse**: transitions are too far apart (in time, or weight space, or …) so that they cannot be effectively pieced together to form a bigger picture
- Help contribute to a research agenda, to be posted at the end of this week.

Theory

- **Subleading terms**: experiments that exhibit them, theoretical clarification of the geometry involved.
- **Singular fluctuation**: what is the significance of this in geometry? How does it affect phase transitions?
- **Phase banding**: under what conditions do phases form collective structures such as “bands”?
- **Bayesian posterior vs SGD**: what is the precise mathematical and experimental relationship between the Bayesian posterior and SGD “invariant measures”.
- **Other limits**: we understand only the $n \to \infty$ limit. What about other limits (e.g. d large relative to n, or going to infinity in a fixed ratio).

Theory

- **Recursive phases**: in-context learning and SLT of mesa-optimisers
- **MDL**: how do classical treatments of MDL and Kolmogorov complexity etc need to be adapted to situations where the “receiver” and “sender” are singular models.
- **Geometry change & phase transitions**: beyond changes in RLCT, what geometry is visible at the level of statistics?
- **Irreducible components in Bayesian statistics**: what is the meaning of the irreducible components of the exceptional divisor of the resolution in Bayesian statistics?
- **How does data determine phase structure?**

What is?

Finding the Right Definitions

- **What is a circuit?** What does it mean, at every level from experimental through to highfalutin theory, for one circuit to be an input to another, etc?
- **What is a feature?** Beyond experimental ideas, what does it mean to “represent” something.
- **What is a Natural Abstraction?** In terms of RG flow & SLT

Experiments

- **Toy Model of Superposition:** classifying local minima, outside of high sparsity, classification of phase transitions.
- **Toy Model of Circuit Formation:** are there phase transitions in the formation of of circuits in vision models?
- **Toy Model of Deception:** what are the high level targets for interpretability?
- **Toy Model Zoo**
- **Automated phase classification**
- **Inference of phase structure**
- **Phase transitions from distribution shift:** we know phase transitions can be induced by changes in the true distribution. This seems relevant to fine-tuning.

Physics

- **Renormalisation Group (RG) flow**: coarse-graining for SLT
- **Statistical physics**: SLT & the replica trick
- **Where are the fermions**: supersymmetry and statistics (Parisi etc)

Dummy's Theories

- By a Dummy's Theory, I mean an easy argument based on the Free Energy Formula and internal model selection.
- **Dummy's Theory of Scaling Laws:** with care, think about scaling laws in D, C
- **Dummy's Theory of Double Descent**
- **Dummy's Theory of Grokking**
- **Dummy's Theory of Modularity**

Bridges

- **Science and governance**: how does science of DL & SLT in particular assist governance?
- **Neuroscience**: what can we learn from critical phenomena in the brain?
- **Stochastic DEs**: resolving diffusion / noise
- **Biology**: developmental biology, gene regulatory networks (degeneracy)

Random Questions

- Do features move?
- Where is a feature before it is “put into” superposition?
- What is the governing singularity of the Induction Head phase transition?