Statement of Research Interests

Branton DeMoss

I study the nature of emergence using tools from algorithmic information theory, machine learning, statistical mechanics, and (soon) fractal geometry.

When can we say emergence has occurred? My recent work re-frames the question around compression, using the Minimum Description Length (MDL) principle. If we have a model M that we use to explain some dataset D, a naïve model selection criterion is to choose the model which achieves the best predictive performance on D. How can we understand whether a "simpler" model is a better choice compared to a more complex one which performs better? To give a poetic example, consider a puddle of water: the underlying dynamics of water are certainly governed by quantum mechanics. However, if we have limited memory and computational ability, we prefer to have access to coarse-grained hydrodynamic information, rather than information about the quantum state of the puddle. Though the quantum description is more complete and can in principle be used to predict the behavior of the puddle with greater accuracy, it requires considerably more information and computational resources to achieve this. With only a few bits of information about macroscopic hydrodynamic variables, we can achieve good prediction of the puddle's dynamics, if we allow for some inaccuracy.

The MDL principle makes this trade-off sharp, and formalizes Occam's Razor by asserting that we should minimize the following sum:

$$\text{Total Description Length of } D \text{ Given } M = H(D \mid M) + C(M) \tag{1}$$

Where $H(D \mid M)$ is the entropy of the data under the model, and $C(M)$ is the model complexity. That is, the best model for the data is the one which minimizes the *sum* of the entropy of the

data and the model complexity. Hence, we can identify emergence by the transition of *our preference* for one model to another, as determined by the MDL. This also makes it clear that emergence is relative to our computational resources: Laplace's demon has no need for emergent descriptions because its computational resources are unbounded. I extended this idea to formalize the notion of "coarse-graining" a model, then used the resultant machinery to explain the grokking phenomenon in neural networks, where networks exhibit a sudden phase transition from memorization of their training data to perfect generalization. Intriguingly, we observed a characteristic rise and fall of complexity in the networks during the transition from memorization to generalization. This same rise and fall of complexity was reported in work by Aaronson et al. [1] which studied the complexity dynamics of a cup of coffee. My next work explains the double descent phenomenon in terms of network complexity dynamics.

Understanding the intrinsic complexity of neural networks lets us bound their generalization performance. In particular, Lotfi et al. [3] show how we can quantify the expected risk $R(h)$ for a given hypothesis h, with probability $1 - \delta$:

$$R(h) \leq \hat{R}(h) + \sqrt{\frac{K(h) + 2\log K(h) + \log(1/\delta)}{2n}} \tag{2}$$

Where $K(\cdot)$ is the Kolmogorov complexity, $\hat{R}(h)$ is the empirical risk, and n the number of samples used to calculate the empirical risk (i.e. the training dataset size). Intriguingly, one can see that this statistical generalization bound corresponds with the MDL principle, so that models which achieve the best total compression of the data correspond to those which are expected to generalize best. Using ideas from algorithmic rate–distortion theory, I was able to produce tight bounds on the Kolmogorov complexity via network compression. This let me track the complexity dynamics of neural networks transitioning from memorization to generalization, providing a sharp picture of the emergence of abstraction and *understanding* in networks.

Statistical generalization bounds like Equation 2 still rely on the iid assumption. This is sufficient for simple stationary problems, but most decision problems in the natural world are non-

stationary: the world changes. Furthermore, for multi-step decision problems the policy directly induces distribution shift, since the next state depends on the current action taken. Generalization to open-ended, real world environments requires methods which can handle distributional adaptation. The Value Equivalence Principle [2] was recently proposed to generate optimal and efficient world models for model-based reinforcement learning agents. It suggests that models ought to coarse-grain away details which do not affect the agent's plan: formally, the agent value function induces an equivalence class of world models which achieve equivalent performance *under the value function*. This can be understood using the lossy compression framework I developed, where the distortion function is taken to be the agent's value function. I will use this link to lift compression-based generalization bounds to the non-stationary, online setting.

We have good reason to believe that much of the apparent complexity we observe is in fact due to repeated iteration of simple rules. It may be possible to bridge the algorithmic complexity of computation with the apparent complexity of the natural world using ideas from fractal geometry. It is already known that the fractal (effective Hausdorff) dimension d_{eff} of a set X can be understood in terms of its Kolmogorov complexity K, via:

$$d_{\text{eff}} = \liminf_{n} \frac{K(X \mid n)}{n} \tag{3}$$

That is, the effective dimension of objects is deeply linked with their computational complexity. This is another sign of emergence: *effective* dimensions which arise from irreducibly complex description. Machine learning must evolve to account for regularities beyond statistical description. The renormalization group and effective field theory gave us our best yet analytical understanding of the nature of emergence. Neural networks, Kolmogorov complexity, and fractal geometry will do the same for analytically intractable complex systems.

References

[1] Scott Aaronson, Sean M. Carroll, and Lauren Ouellette. Quantifying the rise and fall of complexity in closed systems: The coffee automaton, 2014.

[2] Christopher Grimm, André Barreto, Satinder Singh, and David Silver. The value equivalence principle for model-based reinforcement learning. *Advances in neural information processing systems*, 33:5541–5552, 2020.

[3] Sanae Lotfi, Marc Finzi, Yilun Kuang, Tim G. J. Rudner, Micah Goldblum, and Andrew Gordon Wilson. Non-vacuous generalization bounds for large language models. *ICML 2024*, 2024.