Statement of Purpose

Jacob Steinhardt

December 31, 2011

1 Career Goals

The advent of the computer, together with Turing's theory of universal computation, has revolutionized technology and science. Basic physical theories such as quantum mechanics have been formulated on a computational level, leading to inventions such as cryptographic protocols that are *guaranteed by the laws of physics* to be secure. At the same time, institutions such as Google and Wikipedia have consolidated and indexed the sum total of human knowledge, while bioinformaticians have automated the sequencing of the human genome — the very code of life, which will allow us to better understand, predict, and cure many common diseases.

However, for all this progress, one problem remains elusive to the computational paradigm — human intelligence. While computers are astoundingly good at solving formally specified problems, the kind of intuitive reasoning that humans perform each second are harder to reproduce. At the same time, this type of reasoning creates immense value for society — a large fraction of expenditures today go towards human labor, making the automation of "human" tasks an attractive target for economic growth. Furthermore, the conversion of intelligence to digital form would allow us to widely distribute and replicate brilliant digital thinkers, in much the same way that the chess-playing computer program Fritz has brought grandmaster-level chess to personal computers and even PDAs.

These considerations lead me to believe that artificial intelligence is the highest-impact technology that is currently being developed. Widespread deployment of intelligent systems will profoundly alter the pace of both economic and scientific progress, by allowing an increasingly large variety of presently labor-intensive tasks to be automated. My career goals are therefore to develop intelligent systems that will benefit society. Building intelligent systems is already a difficult task, and building them in a way that respects the complex desiderata of human values is even harder (for instance, even a simple task such as building a house has many implicit constraints such as not harming fellow construction workers, not unduly impeding traffic while transporting materials, etc.). These are both interdisciplinary efforts, and I draw my inspiration from such diverse fields as machine learning, computational neuroscience, cognitive science, philosophy, and the theory of computation. My current focus is on the computational aspects of statistical inference, both from a practical perspective and from the philosophical perspective of how to integrate computational constraints into the Bayesian paradigm. My ideas on this are described below.

2 Proposed Research Program

Most of the intelligent behavior that we see in humans can be seen as making good predictions in the face of uncertainty — inferential leaps that are not necessarily logically supported by the data, but are a strong (and usually accurate) guess. These range from picking up an object without a mathematically precise specification of its location, to determining whether or not our food is done cooking, to hearing a sequence of three thuds and concluding that someone is knocking on the door. A leading formalism for talking about inferences under uncertainty is *Bayesian statistics*, where we model the uncertainties as probabilities and *inference* acquires a technical term as the computation of conditional probabilities given the observed data.

Much work in machine learning has recognized the importance of Bayesian inference as a computational problem; this work has culminated in efforts to write probabilistic programming languages that perform general-purpose inference, including the Church programming language [Goodman et al., 2008]. However, despite the large body of work on the computational aspects of Bayesian inference, the Bayesian paradigm itself ignores the issue of computation (in other words, Bayes' theorem only tells us how to reason under uncertainty if we are computationally unbounded).

I believe that computational constraints are fundamental to the nature of intelligence. An intelligent agent is not a Bayes-rational agent with a sufficiently general prior, but rather an agent that can actually perform inference under that prior. Indeed, if we completely decouple modeling from inference then the modeling problem has already been solved: Solomonoff and Hutter have exhibited universal priors that can learn any computable function [Solomonoff, 1978; Hutter, 2001]. However, inference in these models is computationally infeasible. In fact, the infeasibility of inference is a quite general phenomenon. Bayesian updating is not a computationally simple operation, and posterior inference can be NP-hard even for simple models such as Latent Dirichlet Allocation [Sontag and Roy, 2012]. For more complex models, inference can even become uncomputable [Ackerman et al., 2011].

Due to the above considerations, I plan to focus on the computational aspects of inference and their relation to statistical modeling. My first approach will be to construct statistical models where Bayesian updates can be done efficiently (say, in logarithmic or polylogarithmic time). The problem is that most such models will only be able to express very simple concepts. In order to maintain a reasonable degree of expressive power, I will treat the results of yet-to-be-done computations as unknown, and incorporate them into the rest of the Bayesian model. In other words, the model will make predictions about the results of future computations, possibly conditioned on various latent parameters. This will enable efficiency gains — if the model is very confident about what the result of a long computation will be, there is no need to perform it. If the results of separate computations are not treated as independent of one another, it can also allow for the ability to recursively chain together many primitive computations in a way that acquires interesting and complex evidence about the environment. The eventual goal would be to build up a general statistical programming language, although with a somewhat different flavor than Church; instead of allowing arbitrary computable generative models, the syntax of the language should automatically restrict to generative models with efficient posterior inference, while ideally maintaining a deterministic subset that is still Turing-complete.

The approach outlined above solves the inference problem by committing to models where inference is easy, and then trying to understand how to still build rich statistical models out of simple, computationally tractable components. Another approach is to treat inference as a learning problem and to learn sophisticated inference techniques from past experience. Here the inference

problem would remain NP-hard, but the hope is that good learning techniques would allow for the development of sophisticated heuristics that could solve the inference problem for the cases that matter. This approach would focus more on problems such as structural abstraction and re-use, since the main goal is to give the inference engine a rich enough concept space to develop useful heuristics; I would build on work such as the hierarchical Bayesian approach to program learning using adaptor grammars [Liang et al., 2010].

Finally, there are the philosophical issues posed by the fact that all agents are necessarily computationally bounded. Bayesian theory is particularly beautiful because it is provably optimal in the case of a computationally unbounded agent; but the computationally-bounded versions of Bayesianism described above come with no such optimality guarantees, and both are unsatisfying. For instance, only having beliefs that can be updated efficiently means that our true beliefs might lie outside the range of permissible models (a complementary issue is what the notion of "true beliefs" even means in the computationally bounded setting, since it is impossible to even have explicit beliefs about all events on a finite computer, though it may be possible to compute a belief about all events of description length L in $O(L)$ time or some similar guarantee). On the other hand, using approximate algorithms to perform inference means that we never even have access to the underlying beliefs of our model, which is also troublesome. I believe that resolving these issues, in addition to being philosophically satisfying, will indicate promising directions of new research in artificial intelligence.

References

Nate Ackerman, Cameron Freer, and Daniel Roy. Noncomputable conditional distributions. *Proc. Logic in Computer Science*, 2011.

Noah Goodman, Vikash Mansingkha, Daniel Roy, Keith Bonawitz, and Josh Tenenbaum. Church: A language for generative models. *Uncertainty in Artificial Intelligence*, 2008.

Marcus Hutter. Towards a universal theory of artificial intelligence based on algorithmic probability and sequential decisions. *Lecture Notes in Computer Science*, pages 226–238, 2001.

Percy Liang, Michael I. Jordan, and Dan Klein. Learning programs: a hierarchical Bayesian approach. *ICML*, 2010.

Ray Solomonoff. Complexity-based induction systems: comparisons and convergence theorems. *IEEE Trans. Inform. Theory*, IT-24:422–432, 1978.

David Sontag and Daniel Roy. Complexity of inference in latent dirichlet allocation. *NIPS*, 2012.