From Data to Complex Decisions

How to Use Data to Solve the Right Problem?

In the 21st century, we live in a world where data is abundant. We would like to take advantage of this opportunity to make more accurate and data-driven decisions in many areas of life such as industry, healthcare, business, and government. This opportunity has encouraged many machine learning and data mining researchers to develop tools to benefit from big data. Nonetheless, the developed methods so far have mostly been about the task of prediction and many complex decision-making problems, in particular the sequential ones, remain almost untouched.

An example of a complex sequential decision-making problem is designing adaptive treatment strategies for a patient with a chronic disease (ranging from depression to HIV/AIDS). Here the course of treatment is not static, but should dynamically change according to the current state of the patient. Moreover, in the modern practice of medicine much data is also collected about the patient, which could ideally be used to help in designing a customized course of treatment. There are many other problems of utmost importance to our economy and well-being that have this sequential nature. A non-exhaustive list includes problems in robotics and control engineering, advertisement, finance, smart video games, sustainable management of natural resources, and smart cities. Most current techniques in machine learning, data mining and statistics, however, cannot address these problems at all.

This is where my research helps: I develop algorithms to tackle sequential decision-making problems in the presence of uncertainty. In such problems, which can be formulated in the *reinforcement learning (RL)* framework, the goal is to find a policy (i.e., action-selection mechanism) based on interaction with the environment such that a long-term performance is maximized. I design and analyze these algorithms using rigorous mathematical tools from machine learning, statistics, optimization, and control theory. My particular focus is on problems with *high-dimensional* data, which are the most interesting and practically relevant, and yet the most challenging ones. My algorithms have already been applied to problems in robotics, healthcare, and computer vision; and I am eager to contribute to a wider range of application domains. Thanks to the interdisciplinary flavour of sequential decision-making problems, my research contributes to several theoretical disciplines and impacts many application domains.

I have also contributed to other areas of machine learning, such as manifold learning, nonparametric regression, and fundamental question of how to design regularizers. I will briefly describe them as well.

How to Solve High-Dimensional Reinforcement Learning Problems while Avoiding the Curse of Dimensionality?

An important challenge in solving high-dimensional RL problems is that the sample and computational complexity of finding a solution might increase exponentially with the dimension of the state space. This is known as the *curse of dimensionality*. The solution to avoid this curse is to benefit from intrinsic regularities (or structure) of the problem at hand. Examples of such regularities for RL problems include the smoothness of the long-term performance measure (also known as the value function), its sparsity in certain basis functions, or the input data lying close to a low-dimensional manifold. Note that it is not enough for the problem to have certain regularities;

the algorithm itself should have the capability to exploit the regularities that are present without knowing them a priori.

Even though much progress has been made in solving high-dimensional supervised learning problems, prior to my work fewer results had been available for RL problems due to their more challenging nature. In contrast to supervised learning problems like classification or regression, in RL there is no clear target function and we may only have access to an occasional delayed cost (or reinforcement) signal. Also, the input data typically violate the common independent and identically distributed (i.i.d.) assumption. Moreover, as the policy changes, the sampling distribution of input data changes as well, i.e., the data is non-stationary.

During my PhD research I took some key steps toward solving high-dimensional RL problems by introducing several **regularized nonparametric algorithms** [Farahmand et al., 2009b,a; Farahmand, 2011b]. These algorithms can benefit from the intrinsic regularities of a problem by estimating the value function in a large and rich function space (e.g., a reproducing kernel Hilbert space), while explicitly controlling the complexity of the estimate. The estimated value function can then be used to obtain a good policy. I have shown that for an important class of RL problems, these algorithms have strong statistical convergence guarantees, similar to the optimality guarantees for supervised learning problems (regression, in particular). These data-efficient algorithms make the task of solving a high-dimensional RL problem much easier. As an example, we have applied some of these algorithms to the problem of uncalibrated visual servoing for a robotic arm [Farahmand et al., 2009c]. Moreover, for the first time in the literature I rigorously addressed the important problem of model selection in this context [Farahmand and Szepesvári, 2011]. My PhD work has influenced other researchers in the community, and now several researchers work on various regularized algorithms to solve RL problems. According to Google Scholar, there are currently more than 170 citations to my work on this subject.

More recently, I have focused on the relation between the value function and the policy. It has been observed that often the policy is near optimal, even though the estimated value function, the main intermediate estimate in many algorithms, is still far from optimal. I theoretically analyzed this phenomenon and introduced a new type of regularity for RL problems, called the **action-gap regularity** [Farahmand, 2011a]. I showed that RL problems might sometimes be easier than what had been thought before.

This work led me to pay more attention to much less-studied regularities of policy. I designed and analyzed a family of algorithms called **Classification-based Approximate Policy Iteration (CAPI)** during my postdoctoral fellowship at McGill University [Farahmand et al., 2014, 2013]. CAPI is based on the intuition that in some problems the optimal policy has a simple structure even though the value function is quite complex, and vice versa. Using CAPI allows us to simultaneously benefit from both types of regularities, thus to solve more complex problems. CAPI has been successfully applied to the difficult problem of drug scheduling for HIV management.

A common approach to solving complex sequential decision-making problems in robotics is to rely on a human expert to provide example solutions. The robot then tries to imitate the expert, and thus approximate the expert solution. Even though this approach, which is called Learning from Demonstration, can lead to reasonable policies for the robot, it is ultimately limited by the quality of the policy provided by the expert. To overcome this limitation, I designed a new algorithm, Approximate Policy Iteration with Demonstration (APID), that benefits from both the expert's data, even if few or inaccurate, and the cost (or reinforcement) signal. The algorithm is formulated as a constrained optimization problem in which the expert's suggestions define a set of linear constraints for the value function. This framework borrows ideas from the large-margin classification. This allows us to benefit from the regularities of both the value function and the extra information provided by an expert. Under my supervision, a graduate student at McGill

University implemented and applied APID to a robot navigation problem with great success: the quality of solution was much better than both pure RL-based and the conventional Learning from Demonstration approaches. This work has recently been presented as a spotlight talk at the NIPS conference (4% acceptance rate) [Kim et al., 2013].

Other Machine Learning Problems

While studying the low-dimensional manifold regularity of data, the subject of manifold learning, and its consequences for RL problems, I developed and analyzed a simple method to estimate the dimension of a manifold based on sampled data. The analysis showed a remarkable result: The difficulty of manifold dimension estimation depends mainly on the intrinsic dimension of the manifold and not the extrinsic dimension of the input space [Farahmand et al., 2007]. This work was one of the first that rigorously proved the possibility of having a *manifold-adaptive* method.

I also studied the regularized nonparametric regression when the data violates the usual i.i.d. assumption, which often happens in RL problems. I showed that if the data is exponentially β-mixing, the convergence rate of the estimation error is the same as the optimal minimax rate of the i.i.d. case (up to a logarithmic factor) [Farahmand and Szepesvári, 2012]. This work shows that having some mild dependence, as characterized by exponential β-mixing, has almost no effect on learning.

Most current regularized algorithm, in either supervised learning or RL contexts, use generic and analytically tractable regularizers. But for some problems it is more natural to choose the regularizer problem-dependently, for example by regularizing the smoothness according to the underlying distribution of data. The downside of problem-dependent regularizers is that they may not be analytically tractable anymore. To address this issue, I introduced and analyzed Sample-based Approximate Regularization (SAR), which uses Finite Difference approximation and Monte Carlo estimation to approximate the intractable regularizer [Bachman et al., 2014]. I showed that SAR is indeed a sound general procedure. SAR opens up new opportunities to define more complex and task-dependent regularizers.

Future Directions

My research has opened up several exciting future research directions with both theoretical significance and practical impact. I briefly describe a non-exhaustive list of research projects, which might be used as the basis for grant proposals at the beginning of my faculty career.

Fast Algorithms for Big Data Reinforcement Learning Problems: For many real-world applications of RL problems (e.g., recommendation systems on web, financial problems, lifelong robotic problems, epilepsy control based on multi-channel EEG data, etc.), we must process a massive amount of data in order to get any reasonable policy (big data scenario). For these problems, any algorithm that is super-linear will not be practical. The current regularized algorithms, even though sample-efficient, are not computationally cheap (the computational cost is $O(n^3)$ for n being the number of data points). A practically fruitful research topic is to develop regularized *online* RL algorithms that are both sample and computational-efficient and apply them to some of the aforementioned problems. Tools such as stochastic gradient-like methods for optimization, sparsification, and Fast Multipole Methods can be used to develop fast online regularized RL algorithms.

Regularized RL algorithms for Other Models of Sequential Decision-Making Problems:

The current regularized RL algorithms are formulated for an important class of discounted reward Markov Decision Processes (MDP), but some problems are more naturally described as average reward MDPs or Semi-MDPs. Extending regularized RL algorithms to other models of sequential decision making expands the applicability of RL.

Continuous-action CAPI: Most current RL algorithms are not suitable for problems with continuous action spaces, yet in many application domains such as robotics and control engineering we require to use continuous actions. I plan to extend CAPI to continuous action spaces, which seems quite possible, and apply it to control engineering and/or robotic problems.

Deep Architectures for Reinforcement Learning Problems: Deep learning has shown great promise in learning rich representation of data, and is a topic of active research in the machine learning community. The idea of using deep architectures, however, has almost not been explored for RL problems. One exception is the *Value Pursuit Iteration* framework that uses the most recent approximation of the optimal value function as a part of the representation itself and gradually increases the richness of the representation in a data-dependent fashion [Farahmand and Precup, 2012]. Studying theoretical and empirical properties of Value Pursuit Iteration and other deep learning methods for RL is a fruitful and unexplored research area.

Adaptive Treatment Strategies for Chronic Diseases: Designing adaptive treatment strategy for patients with mental or physical chronic diseases is of significant social and economical importance. A few examples of chronic diseases are depression, alcoholism, cardiovascular disorders, inflammatory bowel diseases, and HIV/AIDS. One difficulty to treat chronic diseases, such as depression, is that one type of treatment may not be suitable for all. Moreover, the drugs might take a long time before showing their effectiveness, or lack thereof. Reinforcement learning methods, which find a close to optimal policy based on patient's data, are the kind of analytical tools that can tremendously help us fight against these debilitating diseases.

Intention-aware Robot Learning: Having robots that can easily learn to collaborate with people would be a revolution in our society and industry. There is, however, a major barrier to achieve this goal, and that is to automatically understand the intentions and goals of a human. If the task objectives were known to a robot (i.e., they were programmed), the previously discussed RL algorithms could be used to plan for the task. But programming those objectives is difficult, especially in unstructured and flexible environments, wherein the goals and intentions of a human collaborator might change often. The goal of this project is to develop a mathematical framework and statistically sound algorithms that can provide a predictive model of a human's purpose in his or her actions based on perceptual cues, reliably extract the goals and intentions, and use them to compute a collaborative plan. To achieve this goal we should use ideas and techniques from a wide range of fields such as supervised machine learning, reinforcement learning, and game theory. I started working on this research project since March 2014 as a postdoctoral fellow with J. Andrew Bagnell at the Robotics Institute, Carnegie Mellon University.

Conclusion

The human race just started a new era: Computers are so powerful and data are so abundant that data-driven decisions have the potential to be competitive to the decisions of human experts. We now have reasonable speech recognition systems, are close to having visual search engines, and can personalize ads or movie recommendations on a scale that nobody could dream of 10 years ago. It is an exciting time, but we must not forget that this is just the beginning of an era.

The range of decision making problems that humans solve is much more diverse than the problems of recognizing a pattern or recommending a movie. The science and technology of solving more complex problems such as sequential decision-making under uncertainty, which is arguably the most fundamental problem any human solves, is still far from being fully developed. Many open problems should be solved and technical issues must be addressed. Now one can focus on simpler problems that have immediate commercial impact, or work on fundamental problems that will have huge impact in the future. I am offering a research program that focuses on the latter while being aware of the need to produce technology with immediate impact as it progresses.

References

Philip Bachman, Amir-massoud Farahmand, and Doina Precup. Sample-based approximate regularization. In *Proceedings of the 31st International Conference on Machine Learning (ICML)*, volume 32 of *JMLR: W & CP*, 2014.

Amir-massoud Farahmand. Action-gap phenomenon in reinforcement learning. In *Advances in Neural Information Processing Systems (NIPS - 24)*, 2011a.

Amir-massoud Farahmand. *Regularization in Reinforcement Learning*. PhD thesis, University of Alberta, 2011b.

Amir-massoud Farahmand and Doina Precup. Value pursuit iteration. In *Advances in Neural Information Processing Systems (NIPS - 25)*, 2012.

Amir-massoud Farahmand and Csaba Szepesvári. Model selection in reinforcement learning. *Machine Learning Journal*, 85(3):299–332, 2011.

Amir-massoud Farahmand and Csaba Szepesvári. Regularized least-squares regression: Learning from a β-mixing sequence. *Journal of Statistical Planning and Inference*, 142(2):493 – 505, 2012. URL http://dx.doi.org/10.1016/j.jspi.2011.08.007.

Amir-massoud Farahmand, Csaba Szepesvári, and Jean-Yves Audibert. Manifold-adaptive dimension estimation. In *ICML '07: Proceedings of the 24th international conference on Machine learning*, pages 265–272, New York, NY, USA, 2007. ACM.

Amir-massoud Farahmand, Mohammad Ghavamzadeh, Csaba Szepesvári, and Shie Mannor. Regularized fitted Q-iteration for planning in continuous-space Markovian Decision Problems. In *Proceedings of American Control Conference (ACC)*, pages 725–730, June 2009a.

Amir-massoud Farahmand, Mohammad Ghavamzadeh, Csaba Szepesvári, and Shie Mannor. Regularized policy iteration. In D. Koller, D. Schuurmans, Y. Bengio, and L. Bottou, editors, *Advances in Neural Information Processing Systems (NIPS - 21)*, pages 441–448. MIT Press, 2009b.

Amir-massoud Farahmand, Azad Shademan, Martin Jägersand, and Csaba Szepesvári. Model-based and model-free reinforcement learning for visual servoing. In *Proceedings of IEEE International Conference on Robotics and Automation (ICRA)*, pages 2917–2924, May 2009c.

Amir-massoud Farahmand, Doina Precup, Mohammad Ghavamzadeh, and André M.S. Barreto. Classification-based approximate policy iteration. *IEEE Transactions on Automatic Control (Submitted)*, 2013.

Amir-massoud Farahmand, Doina Precup, André M.S. Barreto, and Mohammad Ghavamzadeh. Classification-based approximate policy iteration: Experiments and extended discussions. *arXiv e-print: 1407.0449*, 2014. URL http://arxiv.org/abs/1407.0449.

Beomjoon Kim, Amir-massoud Farahmand, Joelle Pineau, and Doina Precup. Learning from limited demonstrations. In *Advances in Neural Information Processing Systems (NIPS - 26)*, pages 2859–2867, 2013.