Research Statement

Nan Jiang

My research is on the theory of reinforcement learning (RL) under realistic assumptions. RL is a subfield of machine learning that studies how an agent can learn to make sequential decisions in environments with unknown dynamics. It provides a general and unified framework that captures many important AI applications, including dialog systems, self-driving cars, robots for daily life, and adaptive medical treatments.

Despite the generality of the framework, most theory and practice in RL are yet to directly address these real-life scenarios, due to two major reasons:

(**a**) We have a mature theory that deals with *small state spaces*. Real-world applications have large state spaces, which are considered to be difficult.

Example: for a robot equipped with a camera, the camera image is a part of its (information-theoretic) state, and the set of all possible images already forms a very large space.

(**b**) We have empirical successes in *simulators* (e.g., video/board games). Real-world applications seldom come with high-quality simulators, and how to do RL in this case is an open question.

Example: in dialog applications, a high-quality simulator would need to accurately reproduce human conversations in any hypothetical situation, which does not exist so far.

Recent empirical successes have demonstrated that the difficulty stated in (**a**) can be overcome by deploying powerful function approximation techniques, such as deep neural networks [1, 2]. Still, the state-of-the-art approaches heavily rely on the simulator nature of the environments described in (**b**): in a simulator, data can be generated indefinitely up to computational limits (e.g., AlphaGo generated 30 million self-play games), and taking arbitrarily bad actions has no real-world effects (e.g., crashing a car in a driving game is nothing compared to crashing a real car).

To bring RL to real life, a crucial step is to understand how to do RL when the state space is large, yet the amount of data is limited and the agent's actions incur real consequences. Topics such as overfitting, finite-sample performance, off-policy evaluation, and model selection become important in such a setting, yet many of them are surprisingly under-studied in the RL literature. I believe that good theory has fundamental importance and real impact, and this is what inspires my research.

1 Completed Research

Overfitting in RL: the role of discount factors In RL, a discount factor specifies how far an agent should look ahead into the future, and is closely related to the notion of planning horizon. Despite its importance, existing literature provided limited understanding of its role in RL algorithms, especially in the realistic setting of insufficient data. In [3], my coauthors and I showed a perhaps surprising result that with a limited amount of data, an agent can compute a *better* policy by using a discount factor in the algorithm that is smaller than the groundtruth specified in the problem definition. An explanation for this phenomenon is provided based on principles of learning theory: that a large discount factor causes *overfitting*. The statement is established theoretically by making an analogy between supervised learning (where we search over hypotheses) and reinforcement learning (where we search over policies), and showing that a small discount factor can control the effective size of the policy space and hence avoid overfitting. This work won the best paper award at AAMAS 2015.

Data-driven selection of state abstractions Generalization is key to scaling RL to problems with large state spaces. A particular type of generalization schemes that is convenient to deploy is state abstraction, which can be viewed as aggregation of equivalent or similar states, or specification of irrelevant or less important state features. However, coarse-resolution abstractions are often lossy, while fine-resolution abstractions provide little generalization, hence it can be hard to specify an abstraction that is coarse and accurate simultaneously [4].

A promising approach is to construct abstractions automatically for the problem at hand. Most previous efforts constructed abstractions for faster planning based on the MDP parameters, but such knowledge is not available in the learning setting. Fully data-driven abstraction selection was commonly believed to be hard since constructing abstractions needs model information, but we need abstractions for model learning to begin with. In [5], I took a bold attempt at the problem, and provided the first data-driven abstraction selection algorithm that enjoys statistical adaptivity: the algorithm automatically balances the coarseness-accuracy trade-off and selects a near-optimal abstraction as if the MDP parameters were given.

This work also showcases the idea that meta-learning problems in RL, such as abstraction selection, can often be formulated as an instance of *model selection* as in statistical learning. I expect this general idea to be useful in other aspects of RL as well, such as in option discovery, which I plan to explore in the future.

Policy validation based on historical data Validation, the assessment of the quality of a machine learning solution based on data, has been crucial to the success of supervised learning. In RL, validation of a new policy is relatively easy when the state space is small or we have an accurate simulator. In most real-world applications, neither of the two conditions holds and the validation problem becomes dramatically harder.

However, this is also the situation where a reliable validation paradigm is indispensable: to safely deploy a new policy, we need to verify that its performance is non-degenerate, possibly by estimating the policy's value based on historical data. Along this line of research, my coauthor and I recently developed a new estimator that advances the state of the art, which blends two existing estimators in an organic way to reduce variance without introducing bias [6]. We also proved the first fine-grained lower bound for the problem, revealing its fundamental difficulty and showing that our new estimator matches the lower bound in certain cases. Due to its appealing properties, the estimator quickly got attention from other RL experts who built on our work [7].

Towards practical predictive state representations In realistic environments, an agent often perceives the state of the world via sensors of limited capability, and may need to reason about the sensorimotor history to infer its current situation. (For example, we human beings need memory because what we see and hear at the moment does not provide enough information for decision making.) Predictive State Representations (PSRs) [8] are models of such *partially observable* systems, and can be learned via a spectral algorithm which enjoys statistical consistency and computational efficiency [9].

Despite the elegant theory, spectrally-learned PSRs often fail to deliver practical performance that is comparable to latent-variable models (such as HMMs/POMDPs) learned via Expectation-Maximization. Recently, a surprising defect of the spectral algorithm was discovered, which had been hidden by a widely adopted but unrealistic assumption, and can cause the algorithm to break down in the agnostic setting [10]. In response to such a challenge, my coauthors and I proved that in a particular limiting case, the spectral algorithm retains robustness guarantees in the agnostic setting [11]. Inspired by the theoretical analysis, we also developed practical algorithms that achieve substantially improved performance on synthetic problems and a real-world natural language dataset [12]. To further strengthen the practicality of PSRs, we proposed an optimization-based procedure as a post-processing step for spectral learning [13]. This provides a complete picture for training PSRs that parallels the Spectral-then-EM framework for latent-variable models [14].

A new theory for exploration under rich sensory inputs In RL, exploration refers to taking randomized or diversified actions to collect datasets that provide a well-rounded characterization of the environment. It is also arguably the topic where practice and theory have the largest gap: on one hand, state-of-the-art practice deploys function approximation to handle rich sensory inputs, but often explores in a data-inefficient manner with naïve heuristics. On the other hand, a mature theory has been developed for systematic exploration in small state-space problems, which cannot handle practical scenarios.

In my recent internship at Microsoft Research, my coauthors and I developed a new theory for exploration in large state/observation spaces [15]. Via a lower bound, we showed that a function approximator with low statistical complexity, which guarantees generalization in supervised learning, does *not* necessarily generalize for the purpose of exploration in RL. This result calls for a new measure that characterizes the difficulty of exploration. Our main contributions include proposing the notion of *Bellman rank* as such a measure, and providing a sample-efficient algorithm that provably explores whenever Bellman rank is low. Furthermore, a broad range of RL settings are shown to yield low Bellman rank, from classical control problems to currently popular experimental settings. Overall, this work substantially bridges the gap between theory and practice in RL, and establishes a formal framework for studying exploration in complex environments.

2 Future Directions

In the work mentioned above, I have made contributions to several important directions of RL, and in each direction we are just getting started and there is far more to be done. Just to give an example, I showed that validating new policies using historical data is a fundamentally hard problem; an important next step is to think through how to leverage domain knowledge (e.g. in healthcare) to resolve the difficulty.

Besides these directions, which I will keep pursuing in the future, below I describe two new directions I am eager to explore. They address topics of wide interest, fit well with my expertise, and inspire multi-disciplinary collaboration.

Partial observability meets rich sensory inputs Consider a robot learning about a partially observable environment via a stream of camera images. A typical unsupervised learning approach would involve building a predictive model over *raw* images. However, the sensory inputs from a realistic environment can contain lots of useless yet complex details (e.g., a TV broadcasting an irrelevant show), and predicting every bit of them is unnecessary and demanding.

I would argue that this issue is best addressed under the RL framework. The agent's planning goal (maximizing long-term rewards) provides an absolute criterion, according to which the agent could figure out in a principled manner which aspects of its raw observations should *not* be modeled. Studying this research question may also lead to a new theory on memory formation, a topic that attracts interest from the wider AI community.

AI safety and inverse RL With the rise of AI, there is an increasing amount of public attention on the safety of AI solutions. One of the main concerns is on the negative side-effects that an agent can cause even when it faithfully optimizes for the given reward function [16]. How can we prevent a cooking robot from harvesting the decorative plants for dinner veggies? How can we prevent a cleaning robot from throwing away a bed for tidiness?

Based on a classical RL viewpoint we could argue that this is a type of reward misspecification. However, it is simply infeasible in practice to specify everything that the agent should *not* do. A promising approach is to infer these undesired situations from human behavior, which fits into the research of Inverse RL [17]. However, traditional Inverse RL focuses on simple and individual environments, and the framework needs to be substantially generalized to handle the safety concerns. I am interested in contributing to the theoretical foundations in this important direction.

References

[1] Volodymyr Mnih, Koray Kavukcuoglu, David Silver, Andrei A Rusu, Joel Veness, Marc G Bellemare, Alex Graves, Martin Riedmiller, Andreas K Fidjeland, Georg Ostrovski, et al. Human-level control through deep reinforcement learning. *Nature*, 518(7540):529–533, 2015.

[2] David Silver, Aja Huang, Chris J Maddison, Arthur Guez, Laurent Sifre, George van den Driessche, Julian Schrittwieser, Ioannis Antonoglou, Veda Panneershelvam, Marc Lanctot, et al. Mastering the game of Go with deep neural networks and tree search. *Nature*, 529(7587):484–489, 2016.

[3] **Nan Jiang**, Alex Kulesza, Satinder Singh, and Richard Lewis. The dependence of effective planning horizon on model accuracy. In *Proceedings of the 2015 International Conference on Autonomous Agents and Multiagent Systems*, pages 1181–1189, 2015.

[4] **Nan Jiang**, Satinder Singh, and Richard Lewis. Improving UCT planning via approximate homomorphisms. In *Proceedings of the 2014 International Conference on Autonomous Agents and Multiagent Systems*, pages 1289–1296, 2014.

[5] **Nan Jiang**, Alex Kulesza, and Satinder Singh. Abstraction selection in model-based reinforcement learning. In *Proceedings of the 32nd International Conference on Machine Learning*, pages 179–188, 2015.

[6] **Nan Jiang** and Lihong Li. Doubly robust off-policy value evaluation for reinforcement learning. In *Proceedings of the 33rd International Conference on Machine Learning*, pages 652–661, 2016.

[7] Philip S Thomas and Emma Brunskill. Data-efficient off-policy policy evaluation for reinforcement learning. In *Proceedings of the 33rd International Conference on Machine Learning*, pages 2139–2148, 2016.

[8] Satinder Singh, Michael R James, and Matthew R Rudary. Predictive state representations: A new theory for modeling dynamical systems. In *Proceedings of the 20th Conference on Uncertainty in Artificial Intelligence*, pages 512–519. AUAI Press, 2004.

[9] Byron Boots, Sajid M Siddiqi, and Geoffrey J Gordon. Closing the learning-planning loop with predictive state representations. In *Proceedings of the 9th International Conference on Autonomous Agents and Multiagent Systems*, pages 1369–1370, 2010.

[10] Alex Kulesza, N Raj Rao, and Satinder Singh. Low-rank spectral learning. In *Proceedings of the 17th International Conference on Artificial Intelligence and Statistics*, pages 522–530, 2014.

[11] Alex Kulesza, **Nan Jiang**, and Satinder Singh. Low-rank spectral learning with weighted loss functions. In *Proceedings of the 18th International Conference on Artificial Intelligence and Statistics*, pages 517–525, 2015.

[12] Alex Kulesza, **Nan Jiang**, and Satinder Singh. Spectral learning of predictive state representations with insufficient statistics. In *Proceedings of the 29th AAAI Conference on Artificial Intelligence*, 2015.

[13] **Nan Jiang**, Alex Kulesza, and Satinder Singh. Improving predictive state representations via gradient descent. In *Proceedings of the 30th AAAI Conference on Artificial Intelligence*, 2016.

[14] Yuchen Zhang, Xi Chen, Denny Zhou, and Michael I Jordan. Spectral methods meet EM: A provably optimal algorithm for crowdsourcing. In *Advances in Neural Information Processing Systems*, pages 1260–1268, 2014.

[15] **Nan Jiang**, Akshay Krishnamurthy, Alekh Agarwal, John Langford, and Robert E Schapire. Contextual Decision Processes with low Bellman rank are PAC-learnable. *arXiv preprint arXiv:1610.09512*, 2016.

[16] Dario Amodei, Chris Olah, Jacob Steinhardt, Paul Christiano, John Schulman, and Dan Mané. Concrete problems in AI safety. *arXiv preprint arXiv:1606.06565*, 2016.

[17] Andrew Y Ng and Stuart J Russell. Algorithms for inverse reinforcement learning. In *Proceedings of the 17th International Conference on Machine Learning*, pages 663–670, 2000.