My research goal is to understand the computational and statistical mechanisms required to design efficient AI agents that interact with their environment and adaptively improve their long-term performance. I approach this goal through the lens of *reinforcement learning* (RL) [Bertsekas and Tsitsiklis, 1996; Sutton and Barto, 2018; Farahmand, 2021].[1]

Studying RL is not merely an intellectual goal. RL is also a flexible framework for formulating and solving many complex and impactful real-world sequential decision-making problems in games, healthcare, robotics, advertisement, recommender systems, finance, etc. Refer to Li [2019] for a survey. RL has the potential of having a huge impact on our economy and society, more so than any other area of machine learning (ML).

Despite several success stories and the flexibility of its framework, RL as a technology is not ready for many real-world applications. The reason is partly because existing algorithms are sample inefficient, computationally expensive, do not properly deal with the risk of their decisions, etc. Let me focus on sample inefficiency, which informally means that the RL agent requires too many interactions with its environment before performing well. This is not an issue for problems that are defined on simulators, such as Go or Atari games, but is a great hindrance for applications in which obtaining samples is costly and time consuming.

My research career so far and my goal for the next 5–10 years have been and will be to surpass the limitations of existing RL agents that prevent their applicability to a broad range of real-world problems. My approach, at the high-level, is to *rethink the fundamental algorithms in RL*. We at the Adaptive Agents Lab (Adage) strive towards this goal using a diverse set of research methodologies, ranging from theoretical/mathematical analysis to empirical studies to solving novel and challenging applications. With a robust background in RL research gained at prominent academic institutes such as the University of Alberta, McGill, and CMU, and industrial experience at MERL, coupled with leadership roles as a PI and professor at the Vector Institute and the University of Toronto, I am well-prepared to establish a dynamic RL research lab at your university.

1 Prior and Current Research Directions

I explain some of the major research directions of mine and my lab's. I have a more detailed version of this statement on my website.

1.1 Rethinking the Curse of Dimensionality: Regularization

Challenge: The *curse of dimensionality*, the exponential increase of sample complexity as a function of the input dimension, can be a significant barrier in solving real-world high-dimensional RL problems. Roughly speaking, this curse indicates that the error in estimation decreases with the rate of $O(n^{-1/d})$ for a problem with n data points and input dimension d. This is too slow for high-dimensional problems. This curse can be mitigated by benefitting from the *intrinsic regularities* (or structure) of the problem. Although the ML community had studied inductive biases and regularizers to mitigate this curse for supervised learning problems, this had been very much under-studied in the RL community before my work.

Solution: By rethinking how function approximation is used in RL, I developed a flexible family of regularized RL algorithms that can benefit from the regularity of the value function (long-term performance measure) or the policy (action selection mechanism of the agent) [Farahmand et al., 2009b,a, 2010; Farahmand, 2011b; Farahmand and Szepesvári, 2011; Farahmand, 2011a; Farahmand and Precup, 2012; Farahmand et al., 2015, 2016a]. I was also the first who proved strong theoretical guarantees for such algorithms. As a concrete example, I proved $O(n^{\frac{-2k}{2k+d}})$ error rate for estimating a k-times differentiable value function. This shows that if the value function is smooth enough, we get a much faster rate than

[1] This document covers my research experience until early 2024 and my plans as of that date.

$O(n^{-1/d})$, hence mitigating the curse of dimensionality.

Impact: The insights and techniques of this research program, which I started during my PhD studies, have significantly influenced how the research community at large conduct theoretical RL research these days. Many modern RL theory papers refer to this body of work, and in fact, many researchers have just started working on these topics more than a decade after my initial publications. This research program has received more than 800+ citations so far. I have been invited to give presentations at CMU, McGill, etc.

1.2 Rethinking World Models: Decision-Aware Model Learning

Model-based RL (MBRL) is a promising approach to design sample-efficient agents. An MBRL agent learns a model of the environment (i.e., the physical world) and uses the learned model in an internal simulator along with a Planner to find a good policy. The advantage of MBRL lies in its potential for reducing the need for real-world interactions, thus enhancing sample efficiency of an RL agent.

Challenge: The conventional approach to model learning attempts to learn a good *predictive* model of the environment. As no model can be completely accurate due to the approximation and estimation errors, there are always some errors between the real-world and the learned model. The conventional model learning approach cannot distinguish between decision- and planning-relevant and irrelevant aspects of the environment, hence wastes the capacity of the model and samples on unnecessary detail that are not relevant to the decision making. This leads to less efficient MBRL agents than potentially possible.

Solution: Rethinking what kind of a model an RL agent actually needs for its decision making led me to develop a new general framework/philosophy that I call *Decision-Aware Model Learning (DAML)*. DAML postulates that instead of learning a good predictor of the environment, one should only learn about those aspects of the environment that are relevant to (1) the decision problem itself, and (2) how the Planner uses the model.

Understanding and designing DAML-based methods have been one of the main themes at the Adage lab in the past few years. We have laid theoretical foundations [Farahmand et al., 2017a; Farahmand, 2018; Abachi et al., 2020; Kastner et al., 2023] and studied ways to make them practical for deep neural network (DNN)-based architectures [Voelcker et al., 2022, 2023].

Impact: The decision-aware approach to MBRL has significantly impacted the research community, and I expect it to be one of the main approaches to MBRL in this decade. This research program has attracted more than 200+ citations so far. As a notable example, it has inspired the empirically successful MuZero algorithm by Google DeepMind. I have been invited to give talks on this topic at various places, including Google DeepMind, Microsoft Research, and the book launch event of a new textbook on Distributional RL [Bellemare et al., 2023]. Furthermore, I co-organized the successful Decision Awareness in Reinforcement Learning workshop at the prestigious International Conference on Machine Learning (ICML) in 2022, which drew a large audience.

HQP Training and Grants: PhD students Claas Voelcker, Tyler Kastner, and Romina Abachi are working on various aspects of DAML. My CIFAR AI Chair Award (2018–2024) is partly and my NSERC Discovery Grant (2021–2026) is centrally focused on DAML.

1.3 Application Pull: Controlling Partial Differential Equations

In addition to my theoretical research, I have worked on several applications of RL, including energy management systems for hybrid cars [Farahmand et al., 2016c], smart air-conditioning systems [Farahmand et al., 2017b], and a Question Answering (QA) system for diagnosing health conditions [Akrout et al., 2019]. Here I highlight an applied research direction where I pushed the limits of existing RL agents.

Challenge: RL has regularly been applied to control dynamical systems that are described by Ordinary Differential Equations (ODE), such as rigid robots. The underlying dynamics of many physical systems such as any fluid mechanical or electromagnetic ones, however, are better described by Partial Differential Equations (PDE). Designing a smarter air conditioning system that makes people comfortable was indeed my motivating application when I was at Mitsubishi Electric Research Laboratories (MERL) and initiated this research direction. This is a difficult task, partly because of the complex airflow and temperature fields within a room, which are described by PDEs, and also because the input to an RL agent controlling a PDE is *infinite* dimensional, causing extreme concerns regarding the curse of dimensionality.

Solution: I have been one of the first in the world, if not the first, who considered the possibility of directly controlling PDEs using a learning-based approach. By applying methods developed to mitigate the curse of dimensionality (Section 1.1), I have shown that an RL-based approach can effectively control a time-varying convection-diffusion PDE [Farahmand et al., 2016b, 2017b; Pan et al., 2018; Pirmorad et al., 2021].

Impact: The published papers have attracted 67 citations until now, along with another 22 citations on a patent application. I have been invited to give talks at the Vector Institute, Google Brain, Autodesk Research, Manulife, AI for Engineering Summer School, and McMaster and Carleton Universities. MERL has been actively continuing this research direction since I left, and is pursuing commercialization.

1.4 Other Projects

I briefly outline several additional research projects conducted at the Adage Lab.

- **Search Control Mechanism:** The success of an MBRL agent crucially depends on where in the state space it queries the model. This is controlled by the *search-control* mechanism. Together with my graduated PhD student Yangchen Pan, we showed that the conventional approaches are inefficient, and suggested a new mechanism based on hill climbing a utility function [Pan et al., 2019, 2020, 2022]. This research shows that one of the main modules of an MBRL agent, previously uncontested since 1990s, can be significantly improved. Dr. Pan notably secured a Lecturer position at the University of Oxford.
- **Accelerated Planning:** An RL agent must plan far ahead into the future in order to consider the long-term consequences of its decisions. This long *planning horizon*, however, increases the computational and sample complexities of the agent. The root cause of the increase in complexities is the slowness of the underlying planning algorithms, which have essentially been unchanged since their advent in 1950s by Richard Bellman. By rethinking the planners and making novel connections between planning and other areas of applied mathematics, we developed multiple algorithms with accelerated convergence rates [Farahmand and Ghavamzadeh, 2021; Rakhsha et al., 2022, 2024; Lee et al., 2023].
- **Adversarial Robustness:** DNNs are at the core of modern ML, and understanding their limitations is crucial for designing better RL agents. My PhD student Avery Ma and I have been studying the adversarial risk and robustness of DNNs. We developed a regularizer that improves the adversarial robustness of DNNs [Ma et al., 2021]. We also showed that the choice of optimizer has a dramatic effect on the adversarial robustness and explained it through a frequency domain lens [Ma et al., 2023]. This paper got a *Featured Certificate* at TMLR, putting it at the top 3% of the accepted papers.
- **RL in Frequency Domain:** Going beyond the *expected* sum of rewards is needed for risk-aware RL agents. I developed and theoretically investigated a novel frequency-based representation of the uncertainty of the agent's rewards, called *Characteristic Value Function* [Farahmand, 2019]. This sole-authored NeurIPS paper is mentioned in the textbook on Distributional RL [Bellemare et al., 2023].

2 Future Research Directions

I outline some research directions that I currently believe are the right steps towards my goal of understanding and designing a general-purpose and efficient AI agent. These $\approx 3-5$-year projects are suitable as thesis topics and will be the basis of grant proposals.

2.1 MBRL Agents: The Next Generation

There are still many open questions and design choices that must be resolved before realizing the full potential of MBRL. The next generation of MBRL agents, I believe, should be *decision-aware*, *error-aware*, and *risk-sensitive*.

Decision-Awareness. I expect the decision-aware MBRL approach to be an active research area for the RL community in this decade. Some of my planned directions are:

- **How can we design a Bayesian DAML?** In contrast to existing non-Bayesian approaches, a Bayesian DAML provides a distribution over the agent's belief about the truth of each model. This is crucial for active exploration, safety, etc.
- **When do DAML-based approaches work best?** Fully characterizing when DAML is superior to Prediction-based approaches is an open theoretical question whose solution may lead to design of better model learning algorithms too.
- **How can we make DAML a practical approach?** Existing DAML-based methods are rather difficult to optimize. Designing algorithms that can easily be used by practitioners broadens the impact of the decision-aware line of work.

Grant: Some of these questions are the topic of my NSERC DG (2021–2026).

Error-Awareness. Regardless of how good the model learning algorithm of an MBRL agent performs, decision-aware or not, there is always some error between the learned model and the true dynamics. Designing a Planner that explicitly considers the errors in the model is a key underdeveloped step in successful applications of MBRL.

Risk-Awareness. Your level of risk aversion is different than mine, so should be our AI agents'. An RL agent should be able to work with a wide range of risk functionals based on the user's current level of risk-aversion. Despite the apparent simplicity of this desideratum, the literature on risk-awareness, especially in the context of MBRL, is rather limited. In the context of decision-awareness, the literature is essentially non-existent, except our own recent work [Kastner et al., 2023].

2.2 Accelerated RL Agents

Our recent results show that redesigning the core planning algorithms to create accelerate planners is indeed possible (Section 1.4). This vastly unexplored area promises RL agents that learn much faster, in both computational and sample complexity senses, than the existing ones. My future research addresses the followings:

- **Accelerated RL algorithms for high-dimensional problems:** To apply the proposed accelerated algorithms to high-dimensional RL problems, we need to use function approximators such as DNNs. The interplay between the underlying accelerated planner and DNNs requires careful investigations in order to design practical algorithms.
- **Sample complexity of accelerated RL algorithms:** Characterizing the sample complexity of accelerated RL algorithms requires an analysis of the interaction of the planning error and the statistical error due to finiteness of data, using tools developed in the work described in Section 1.1. This study may lead to better planners too.
- **Co-design of Planner and Model Learner:** Any MBRL algorithm can be interpreted as two coupled dynamical systems induced by (1) Planner and (2) the Model Learner. As Planner uses the

learned model, these two dynamical systems are coupled. The coupling, however, is often ignored when one designs them separately. By taking a dynamical system viewpoint, we can *co-design* Planner and Model Learner so that the coupled dynamical systems have desirable properties, such as accelerated convergence or robustness to errors.

Grant: This research is the topic of my recently awarded Ontario Early Researcher Award (2024–2029).

2.3 Other Future Research Directions

I mention some other research directions of three different flavours: algorithmic/empirical, interdisciplinary, and applied. These are complementary to my theoretical directions.

- **LLM-augmented RL Agents:** The rich inductive biases encoded by LLMs can be used to design more sample-efficient AI agents (Section 1.1). I see two particular projects overlapping my current expertise: (1) *Deep Reasoning*, in which we create agents capable of complex multi-step reasoning tasks, beyond the current emergent reasoning capabilities of LLMs, by formulating the reasoning task as a sequential decision-making problem in a large prompt space and solving them using modern RL tools; (2) *Low-Cost Reasoning*, in which we use the ideas from RL Acceleration, specifically the OS-VI framework [Rakhsha et al., 2022], to use two LLMs together, a cheap one and an expensive one, to perform deep reasoning tasks.
- **Social Learning and Theory of Mind:** Rarely our agents are alone in the world. How can we design them to deal with humans or other AI agents? We are exploring this question from the perspective of Theory of Mind, with Prof. William Cunningham, a cognitive social psychologist at the University of Toronto. This research benefits both AI and Psychology: it allows developing more efficient and human-friendly AI agents, and may suggest novel hypotheses on how humans behave socially.
- **RL Applications:** I am interested in collaborating with domain experts in challenging applications of RL in healthcare, sustainability, engineering, etc. Two recently initiated collaborations are: (1) **RL for Smart Carbon Storage**, in which we formulate the question of where to inject CO2 in the ground and by how much as an RL problem with a PDE dynamics (Section 1.3). This is in collaboration with Prof. Swidinsky, a geophysicist at the University of Toronto. (2) **RL for Operations Research Problems**, in which we investigate designing a unified RL framework to solve a variety of online stochastic optimization problems appearing in Operations Research, including Electric ride-hail problem and Scheduling of home charging operations for electric vehicles. This is in collaboration with Profs. Cousineau and Mendoza from HEC Montréal. For both collaborations, we have submitted grant applications.

References

Romina Abachi, Mohammad Ghavamzadeh, and Amir-massoud Farahmand. Policy-aware model learning for policy gradient methods. *arXiv:2003.00030v2*, 2020. 2

Mohamed Akrout, Amir-massoud Farahmand, Tory Jarmain, and Latif Abid. Improving skin condition classification with a visual symptom checker trained using reinforcement learning. In *International Conference on Medical Image Computing and Computer Assisted Intervention (MICCAI)*, 2019. 2

Marc G. Bellemare, Will Dabney, and Mark Rowland. *Distributional Reinforcement Learning*. MIT Press, 2023. http://www.distributional-rl.org. 2, 3

Dimitri P. Bertsekas and John N. Tsitsiklis. *Neuro-Dynamic Programming*. Athena Scientific, 1996. 1

Amir-massoud Farahmand. Action-gap phenomenon in reinforcement learning. In J. Shawe-Taylor, R.S. Zemel, P.L. Bartlett, F. Pereira, and K.Q. Weinberger, editors, *Advances in Neural Information Processing Systems (NIPS - 24)*, pages 172–180. Curran Associates, Inc., 2011a. 1

Amir-massoud Farahmand. *Regularization in Reinforcement Learning*. PhD thesis, University of Alberta, 2011b. 1

Amir-massoud Farahmand. Iterative value-aware model learning. In *Advances in Neural Information Processing Systems (NeurIPS - 31)*, pages 9072–9083, 2018. 2

Amir-massoud Farahmand. Value function in frequency domain and the characteristic value iteration algorithm. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2019. 3

Amir-massoud Farahmand. *Lecture Notes on Reinforcement Learning*. 2021. 1

Amir-massoud Farahmand and Mohammad Ghavamzadeh. PID accelerated value iteration algorithm. In *International Conference on Machine Learning (ICML)*, 2021. 3

Amir-massoud Farahmand and Doina Precup. Value pursuit iteration. In F. Pereira, C.J.C. Burges, L. Bottou, and K.Q. Weinberger, editors, *Advances in Neural Information Processing Systems (NIPS - 25)*, pages 1349–1357. Curran Associates, Inc., 2012. 1

Amir-massoud Farahmand and Csaba Szepesvári. Model selection in reinforcement learning. *Machine Learning*, 85(3):299–332, 2011. 1

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. 1

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 21 (NIPS 2008)*, pages 441–448. MIT Press, 2009b. 1

Amir-massoud Farahmand, Rémi Munos, and Csaba Szepesvári. Error propagation for approximate policy and value iteration. In J. Lafferty, C. K. I. Williams, J. Shawe-Taylor, R. S. Zemel, and A. Culotta, editors, *Advances in Neural Information Processing Systems (NeurIPS - 23)*, pages 568–576. 2010. 1

Amir-massoud Farahmand, Doina Precup, Mohammad Ghavamzadeh, and André M.S. Barreto. Classification-based approximate policy iteration. *IEEE Transactions on Automatic Control*, 60(11): 2989–2993, November 2015. 1

Amir-massoud Farahmand, Mohammad Ghavamzadeh, Csaba Szepesvári, and Shie Mannor. Regularized policy iteration with nonparametric function spaces. *Journal of Machine Learning Research (JMLR)*, 17(139):1–66, 2016a. 1

Amir-massoud Farahmand, Saleh Nabi, Piyush Grover, and Daniel N. Nikovski. Learning to control partial differential equations: Regularized fitted Q-iteration approach. In *IEEE Conference on Decision and Control (CDC)*, pages 4578–4585, December 2016b. 3

Amir-massoud Farahmand, Daniel N. Nikovski, Yuji Igarashi, and Hiroki Konaka. Truncated approximate dynamic programming with task-dependent terminal value. In *AAAI Conference on Artificial Intelligence*, February 2016c. 2

Amir-massoud Farahmand, André M.S. Barreto, and Daniel N. Nikovski. Value-aware loss function for model-based reinforcement learning. In *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics (AISTATS)*, pages 1486–1494, April 2017a. 2

Amir-massoud Farahmand, Saleh Nabi, and Daniel N. Nikovski. Deep reinforcement learning for partial differential equation control. In *American Control Conference (ACC)*, 2017b. 2, 3

Tyler Kastner, Murat A. Erdogdu, and Amir-massoud Farahmand. Distributional model equivalence for risk-sensitive reinforcement learning. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023. 2, 4

Jongmin Lee, Ernest Ryu, Amin Rakhsha, and Amir-massoud Farahmand. Deflated dynamics value iteration. Under review, 2023. 3

Yuxi Li. Reinforcement learning applications. *arXiv, 1908.06973*, 2019. 1

Avery Ma, Fartash Faghri, Nicolas Papernot, and Amir-massoud Farahmand. SOAR: Second-order adversarial regularization. *arXiv preprint 2004.01832(v2)*, February 2021. 3

Avery Ma, Yangchen Pan, and Amir-massoud Farahmand. Understanding the robustness difference between stochastic gradient descent and adaptive gradient methods. *Transactions on Machine Learning Research (TMLR)*, 2023. 3

Yangchen Pan, Amir-massoud Farahmand, Martha White, Saleh Nabi, Piyush Grover, and Daniel Nikovski. Reinforcement learning with function-valued action spaces for partial differential equation control. In *Proceedings of the 35th International Conference on Machine Learning (ICML)*, volume 80, pages 3986–3995, Jul 2018. 3

Yangchen Pan, Hengshuai Yao, Amir-massoud Farahmand, and Martha White. Hill climbing on value estimates for search-control in Dyna. In *Proceedings of International Joint Conference on Artificial Intelligence (IJCAI)*, 2019. 3

Yangchen Pan, Jincheng Mei, and Amir-massoud Farahmand. Frequency-based search-control in Dyna. In *International Conference on Learning Representations (ICLR)*, 2020. 3

Yangchen Pan, Jincheng Mei, Amir-massoud Farahmand, Martha White, Hengshuai Yao, Mohsen Rohani, and Jun Luo. Understanding and mitigating the limitations of prioritized experience replay. In *Conference on Uncertainty in Artificial Intelligence (UAI)*, 2022. 3

Erfan Pirmorad, Faraz Khoshbakhtian, Farnam Mansouri, and Amir-massoud Farahmand. Deep reinforcement learning for online control of stochastic partial differential equations. In *NeurIPS Workshop on the Symbiosis of Deep Learning and Differential Equations*, 2021. 3

Amin Rakhsha, Andrew Wang, Mohammad Ghavamzadeh, and Amir-massoud Farahmand. Operator splitting value iteration. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2022. 3, 5

Amin Rakhsha, Mete Kemertas, Mohammad Ghavamzadeh, and Amir-massoud Farahmand. Maximum entropy model correction in reinforcement learning. In *International Conference on Learning Representations (ICLR)*, 2024. 3

Richard S. Sutton and Andrew G. Barto. *Reinforcement Learning: An Introduction.* The MIT Press, second edition, 2018. 1

Claas A. Voelcker, Victor Liao, Animesh Garg, and Amir-massoud Farahmand. Value gradient weighted model-based reinforcement learning. In *International Conference on Learning Representations (ICLR)*, 2022. 2

Claas A. Voelcker, Arash Ahmadian, Romina Abachi, Igor Gilitschenski, and Amir-massoud Farahmand. λ-AC: Effective decision-aware reinforcement learning with latent models. 2023. 2