Literature Review

Abraham Frei-Pearson
Department of Mathematics

Sheroze Sheriffdeen
Oden Institute

November 15, 2019

We are interested in applying reinforcement learning techniques to noisy optimal control problems. The general case will be in the following form. Given an unknown $n \times n$ matrix valued function A, our state will be a vector $x_t \in \mathbb{R}^n$, and our dynamics will be

$$x_{t+1} = (I + A(x_t))x_t + u(x_t) + g,$$

where g is Gaussian noise. Here u will be a control function, which we are trying to learn, which will be specified in different problem domains. In general, we'll take u to be bounded. We could consider the avoid problem: we would like x to avoid a certain region. We could also consider the reach problem: we would like x to reach a specified neighborhood of the origin (say) in time as small as possible.

Cartpole, or inverted pendulum, is a basic example of this sort of problem in the noise free setting, and it will be our starting point. To start with we'll consider *bang-bang* controls only: the cart can be accelerated either left or right with full force at each time. With this in mind, the problem admits the following description:

> *State space*: This is the position of the cart, and the angle and angular velocity of the pendulum, so it is $\mathbb{R} \times (0, \pi) \times \mathbb{R}$.
>
> *Action space*: This is just whether we are accelerating the box left or right, so $\{\pm 1\}$.
>
> *Reward function*: We will take this to be -1 if the pendulum falls past a given threshold, and 0 otherwise.

In the noise-free setting, the inverted pendulum is stable with an appropriate control. On the other hand, if the noise term is large enough compared to the size of the control, the pendulum will tip over quickly no matter what you do. It would be interesting to compute this relationship empirically and see how common reinforcement learning algorithms compare. We will also explore the same problem in the noise-free setting with noisy observations. In this case, the state should include previous observations, since these can be used to more appropriately approximate the true state of the system. If this goes well, we

would like to explore some of the other optimal control problems in the OpenAI gym.

There are a number of reach goals which would be of interest. What if damping is added to the inverted pendulum, i.e. the acceleration caused by the control depends on the position of the cart. This problem is interesting in the one-shot case: suppose the dynamics (i.e. the matrix A above) is chosen randomly. What is the best strategy for computing u? In this case, we will need to move away from bang-bang controls: it makes sense to hold off on firing the control, or fire it at a small velocity, early on to learn the system dynamics. Could the reach or avoid problems be solved safely in this case?

There are several abstractions we are interested in considering: noise in the system itself, noise in the observations, and noise in the reward. In what follows, we will refer to a problem as "traditional optimal control" if the noise-free dynamics are known, and "reinforcement learning" if they are not.

Noisy dynamics introduce maximization bias, as discussed in Sutton and Barto [3], into the standard reinforcement learning algorithms. The point is that, early in training, noise causes the best-looking action to look better than it actually is. This is dealt with effectively by double-Q learning. Another algorithm for dealing with maximization bias is discussed in [4], so called "G-learning". This algorithm begins with a stochastic policy ρ. A pseudo-reward is calculated for each state action pair called $G_\pi(s,a)$, where $G_\pi(s,a)$ is $Q(s,a)$ less an information cost which is a discounted version of the Kullback-Liebler divergence $KL(\pi(;s)||\rho(;s))$.

In the case of traditional optimal control, at least when the state dynamics are locally linear, an optimal estimate for the state is given by extended Kalman filtration [10]. This is optimal in the sense that, given knowledge about the noise, the Kalman estimate for the state minimizes the expected squared error. The Kalman estimate is as follows: suppose the matrix A above is known, and our estimate for x_t is $\hat{x}_t$. We obtain an estimate for x_{t+1} by applying $H := I + A$ to $\hat{x}_t$. If z is the observed value at time $t+1$, then we set $\hat{x}_{t+1} = x_{t+1} + K(z - x_{t+1})$, where K is the *Kalman gain*, obtained by minimizing the a posteriori variance of the estimate. We would like to see how much applying the Kalman filter to our data in a noisy environment improves the usual reinforcement learning methods. Although we need a model to apply the Kalman filter, it would still be interesting to study improvements to reinforcement learning from filtered data. Kalman filtration is also applied to reinforcement learning problems in [9] in a somewhat different way: in this case Kalman filtration is applied to the update of the Q-function directly, and then Q-learning is applied.

We are also interested, time permitting, in studying noisy problems in the one-shot case. Consider a discrete, linear dynamical system with noise

$$x_{t+1} = (I + A)x_t + \eta + u_t x_t,$$

where η is Gaussian and u_t is a linear control. Our reward will be 1 for all t where x_t is within a ball around the origin. What if the matrix A is unknown a priori?

This problem is studied in [6] and [1], using the theory of differential inclusions. This is also addressed in [2] in the min-max setting. Given a set of transitions $\mathcal{P}$, an algorithm for producing a policy maximizing $\mathbb{E}_{p\in\mathcal{P},\Pi}\sum_t \gamma^t R(X_t)$, where X_t is the trajectory under π, is presented. It would be interesting to see if this algorithm could be adjusted to include a prior on the transitions, and optimize the expected reward according to that prior rather than the min-max reward. We would be interested in studying some simple examples of this in the context of reinforcement learning, in simple tabular or one dimensional cases. This problem is also studied in [5] in the context of optimal control.

A reformulation of the stochastic optimal control problem in terms of KL divergence minimization is examined by Rawlik et. al [7]. This work considers reinforcement learning as an instance of stochastic optimal control which does not assume knowledge of the dynamics. A cost is associated with each policy relating to a measure of the control, and the stochastic optimal policy is posed as finding an optimal policy that minimizes the expected cost of controls following trajectories under the given policy. An experiment performed with the cart-pole problem with zero mean Gaussian noise added to the state and an extended Kalman smoother [8] is use to estimate a Gaussian approximation to the full posterior of the state variables, leading to a Gaussian posterior policy.

References

[1] Mohamadreza Ahmadi, Arie Israel, and Ufuk Topcu. Controller Synthesis for Safety of Physically-Viable Data-Driven Models. *arXiv e-prints*, page arXiv:1801.04072, Jan 2018.

[2] J. Andrew (Drew) Bagnell, Andrew Y. Ng, and Jeff Schneider. Solving uncertain markov decision problems. Technical Report CMU-RI-TR-01-25, Carnegie Mellon University, Pittsburgh, PA, August 2001.

[3] Andrew Barto and Richard Sutton. *Rinforcement Learning: An Introduction*. 2nd edition, 2018.

[4] Roy Fox, Ari Pakman, and Naftali Tishby. Taming the Noise in Reinforcement Learning via Soft Updates. *arXiv e-prints*, page arXiv:1512.08562, Dec 2015.

[5] Steven I Marcus, Emmanual Fernández-Gaucherand, Daniel Hernández-Hernandez, Stefano Coraluppi, and Pedram Fard. Risk sensitive markov decision processes. In *Systems and control in the twenty-first century*, pages 263–279. Springer, 1997.

[6] Melkior Ornik, Arie Israel, and Ufuk Topcu. Control-Oriented Learning on the Fly. *arXiv e-prints*, page arXiv:1709.04889, Sep 2017.

[7] Konrad Rawlik, Marc Toussaint, and Sethu Vijayakumar. On stochastic optimal control and reinforcement learning by approximate inference.

In *Twenty-Third International Joint Conference on Artificial Intelligence*, 2013.

[8] Robert F Stengel. Stochastic optimal control: theory and application. *New York*, 1986.

[9] Charles Tripp and Ross D. Shachter. Approximate Kalman Filter Q-Learning for Continuous State-Space MDPs. *arXiv e-prints*, page arXiv:1309.6868, Sep 2013.

[10] Greg Welch and Gary Bishop. An introduction to the kalman filter. Technical report, Chapel Hill, NC, USA, 1995.