Belief space model predictive control for approximately optimal system identification

Boris Belousov
Department of Computer Science
Technische Universität Darmstadt, Germany
belousov@ias.tu-darmstadt.de

Hany Abdulsamad
Department of Computer Science
Technische Universität Darmstadt, Germany
abdulsamad@ias.tu-darmstadt.de

Matthias Schultheis
Department of Computer Science
Technische Universität Darmstadt, Germany
matthias.schultheis@gmail.com

Jan Peters
Department of Computer Science
Technische Universität Darmstadt, Germany
Max Planck Institute for Intelligent Systems
peters@ias.tu-darmstadt.de

Abstract

The fundamental problem of reinforcement learning is to control a dynamical system whose properties are not fully known in advance. Many articles nowadays are addressing the issue of optimal exploration in this setting by investigating the ideas such as curiosity, intrinsic motivation, empowerment, and others. Interestingly, closely related questions of optimal input design with the goal of producing the most informative system excitation have been studied in adjacent fields grounded in statistical decision theory. In most general terms, the problem faced by a curious reinforcement learning agent can be stated as a sequential Bayesian optimal experimental design problem. It is well known that finding an optimal feedback policy for this type of setting is extremely hard and analytically intractable even for linear systems due to the non-linearity of the Bayesian filtering step. Therefore, approximations are needed. We consider one type of approximation based on replacing the feedback policy by repeated trajectory optimization in the belief space. By reasoning about the future uncertainty over the internal world model, the agent can decide what actions to take at every moment given its current belief and expected outcomes of future actions. Such approach became computationally feasible relatively recently, thanks to advances in automatic differentiation. Being straightforward to implement, it can serve as a strong baseline for exploration algorithms in continuous robotic control tasks. Preliminary evaluations on a physical pendulum with unknown system parameters indicate that the proposed approach can infer the correct parameter values quickly and reliably, outperforming random excitation and naive sinusoidal excitation signals, and matching the performance of the best manually designed system identification controller based on the knowledge of the system dynamics.

Keywords: Bayesian experimental design, active exploration, curiosity, belief space planning, trajectory optimization

Acknowledgements

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 640554.

1 Introduction and related work

Adaptation and learning arise as a by-product of optimization in the belief space within the framework of Bayesian decision theory [Stratonovich, 1968a, Stratonovich, 1968b]. In modern terminology, learning is planning in a partially observable Markov decision process [Asmuth and Littman, 2011]. We pursue this line of reasoning and frame the problem of pure exploration (i.e., without any extrinsic reward) as a problem of online belief space trajectory optimization.

Optimal system identification and experimental design [Mehra, 1974, Bombois et al., 2011, Ryan et al., 2016] pursue a similar objective. They seek an optimal exploration strategy in stochastic sequential decision making problems. Contrary to the generic solution based on approximate dynamic programming [Feldbaum, 1960, Huan and Marzouk, 2016], we do not aim to find an optimal parametric policy but instead let a belief space planner choose the most explorative actions.

Approaches to (approximately) optimal system identification based on model predictive control (MPC) have been studied before [Larsson et al., 2013]. Algorithmically, our method is most closely related to [Kahn et al., 2015], who also used direct transcription in the belief space for trajectory optimization. However, what is different in our case is the objective function and its particular decomposition into a sum of terms that facilitates computation. More concretely, since robot dynamics is linear in the physics parameters [Atkeson, 1989], we can perform Bayesian inference in closed form.

The paper is structured as follows: Section 2 introduces the approach, Section 3 provides evaluations, and Section 4 highlights future directions.

2 Belief space optimization for system identification

Consider a dynamical system of the following form

$$x' = Ax + B(x, u)\theta \tag{1}$$

where $x \in \mathbb{R}^n$ is the current state, $u \in \mathbb{R}^m$ is the current action, $x' \in \mathbb{R}^n$ is the next state, matrix $A \in \mathbb{R}^{n \times n}$ is constant and matrix $B(x, u) \in \mathbb{R}^{n \times m}$ is dependent on the state and action. Many classical continuous control environments can be written in this way.

2.1 Example: pendulum dynamics

As a concrete instantiation of (1), consider the dynamics of a pendulum

$$\ddot{q} = \phi\left(\begin{pmatrix} q \\ \dot{q} \end{pmatrix}, u\right)^T \theta = \begin{pmatrix} -\sin(q + \pi) & -\dot{q} & u \end{pmatrix} \begin{pmatrix} \frac{3g}{2l} \\ \frac{3b}{ml^2} \\ \frac{3}{ml^2} \end{pmatrix} \tag{2}$$

with mass m, length l, and gravity g. The state of the pendulum $x = (q, \dot{q})$ is comprised of the angle q and the angular velocity $\dot{q}$. Crucially, the kinematic parameters $\phi(x, u)$ and the dynamic parameters θ separate. The system can be discretized using the implicit Euler integration scheme

$$x' = \begin{pmatrix} 1 & h \\ 0 & 1 \end{pmatrix} x + \begin{pmatrix} h^2 \\ h \end{pmatrix} \phi(x, u)^T \theta. \tag{3}$$

This representation directly corresponds to the generic form (1), with matrices A and $B(x, u)$ straightforward to identify.

2.2 Propagation of uncertainty

If parameter values θ are uncertain, they should be characterized by a probability distribution $p(\theta)$. The full state of the system should then include it and we have to describe its dynamics. Assuming the initial belief $p(\theta) = N(\theta|\mu, \Sigma)$ and the system dynamics $p(x'|x, u; \theta) = N(x'|Ax + B(x, u)\theta, Q)$ are Gaussian, the posterior after observing a transition (x, u, x') is also Gaussian with parameters given by the standard Kalman filter update equations [Bishop, 2006]

$$K(x, u, \Sigma) = \Sigma B(x, u)^T \left(Q + B(x, u)\Sigma B(x, u)^T\right)^{-1}, \tag{4}$$

$$L(x, u, \Sigma) = I - K(x, u, \Sigma)B(x, u), \tag{5}$$

$$\mu' = \mu + K(x, u, \Sigma)\left(x' - Ax - B(x, u)\mu\right), \tag{6}$$

$$\Sigma' = L(x, u, \Sigma)\Sigma. \tag{7}$$

Kalman gain $K(x, u, \Sigma)$ and matrix $L(x, u, \Sigma)$ are introduced for convenience to simplify Equations (6) and (7) that describe the dynamics of the sufficient statistics of the belief state.

To plan using the model (6)-(7), future observations x' need to be integrated out. This results in the maximum likelihood transition dynamics $x' = Ax + B\mu$ and the constant mean update $\mu' = \mu$. Intuitively, such constancy is a manifestation of the fact that the mean of the parameter estimate μ cannot be improved before observing any data. Nevertheless, its variance Σ can be controlled.

Equation (7) gives the update rule for the covariance matrix and serves as the key to our formulation of the objective function. Namely, we exploit the fact that the covariance matrix at the next time step is given by a product of matrices. For example, after two time steps, $\Sigma'' = L(x', u', \Sigma')L(x, u, \Sigma)\Sigma$.

2.3 Entropy minimization objective

What should the objective function be? A conceptually straightforward approach is to minimize the entropy of the posterior distribution over the parameters at the end of the planning horizon. This objective essentially asks for the most informative actions and can be identified with the information gain criterion [Lindley et al., 1956]. It also fits nicely with the multiplicative form of the covariance matrix, turning the product into a sum. For example, for a two-stage problem,

$$J = \frac{1}{2} \log\det\left(2\pi e \Sigma''\right) \propto \log\det \Sigma'' = \log\det L(x', u', \Sigma') + \log\det L(x, u, \Sigma) + \log\det \Sigma. \tag{8}$$

Similarly, for an N-step trajectory,

$$J \propto \sum_{k=0}^{N-1} \log\det L(x_k, u_k, \Sigma_k). \tag{9}$$

Thus, the summand $L(x_k, u_k, \Sigma_k)$ can be viewed as a running cost. Adding a regularization term $u^T Ru$ for smoothness, we arrive at the following optimization problem

$$\underset{u_{0:N-1}}{\text{minimize}} \quad \sum_{k=0}^{N-1} \log\det L(x_k, u_k, \Sigma_k) + u_k^T R u_k \tag{10}$$

$$\text{subject to} \quad x_{k+1} = Ax_k + B(x_k, u_k)\mu, \quad k = 0, 1, \ldots, N-1, \tag{11}$$

$$\Sigma_{k+1} = L(x_k, u_k, \Sigma_k)\Sigma_k, \quad k = 0, 1, \ldots, N-1, \tag{12}$$

where $L(x, u, \Sigma) = I - \Sigma B(x, u)^T \left(Q + B(x, u)\Sigma B(x, u)^T\right)^{-1} B(x, u)$. This problem can be directly plugged into a trajectory optimizer, e.g., CasADi [Andersson et al., 2012]; state and control constraints can be added if needed.

3 Evaluation

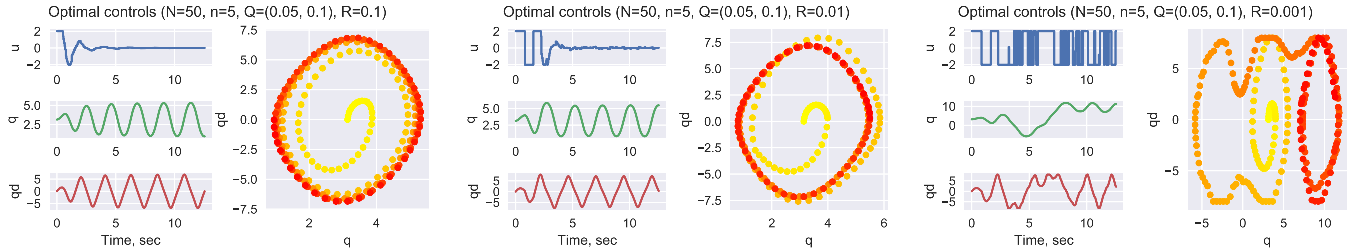
Having solved the problem above, we obtain a sequence of actions $u_{0:N-1}$ that should reveal the most about the system. Note that this sequence of actions depends on our prior belief $p(\theta|\mu, \Sigma)$ because μ enters the state dynamics and Σ figures in the covariance cost. Thus, the optimal sequence of actions is a function of the prior together with the initial state x_0, i.e., $u_{0:N-1} = \psi(x_0, \mu, \Sigma)$. We can think of ψ as a call to the trajectory optimizer.

The main question is whether this sequence of actions is better than any other one given that the true value $\mu^\star$ is different from μ. One way to evaluate this hypothesis is to execute $u_{0:N-1}$ on the real system with parameters $\mu^\star$ and then find the posterior $p(\theta|x_{0:N}, u_{0:N-1})$ given the observed trajectory. An even better solution is to replan after every time step. Such closed loop control should intuitively speed up convergence to the true parameter value. We call this approach belief space model predictive control for approximately optimal system identification.

We compare the belief space MPC approach (Figure 1) against random and sinusoidal excitations (Figure 2) on the pendulum environment from OpenAI Gym [Brockman et al., 2016]. Optimal exploration performs well and beats random actions and a naively chosen excitation signal by a large margin (Figure 3). However, a wisely chosen excitation signal can be as good as the optimal one (Figure 4). The optimization approach was found quite insensitive to the choice of the action cost R in a reasonable range, although extremely small values were found to cause instability.

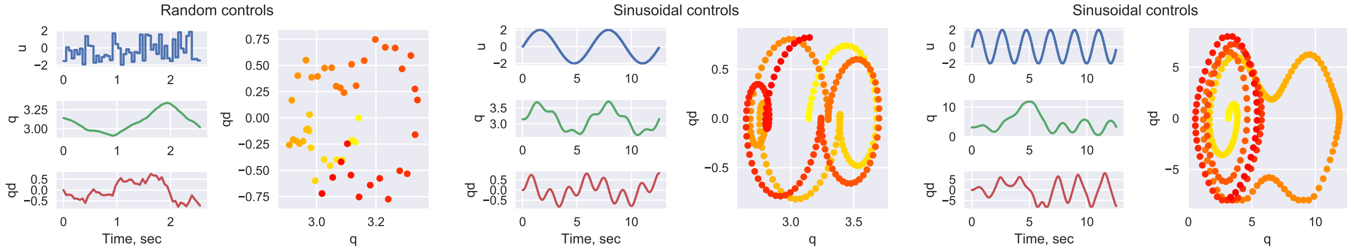
4 Conclusion

Although the preliminary results are encouraging, further investigation is required. First, evaluation on more complex systems must be performed to demonstrate the scalability of the approach. Second, comparison to other exploration strategies is needed to better understand the trade-offs between optimality and heuristics. Third, the assumption on the system dynamics (1) can be relaxed to allow for more flexible models; for example, the feature mapping ϕ can be learned by exploiting its invariance to dynamics parameters, or a non-parametric model, such as a Gaussian process, can be employed to represent the system dynamics.



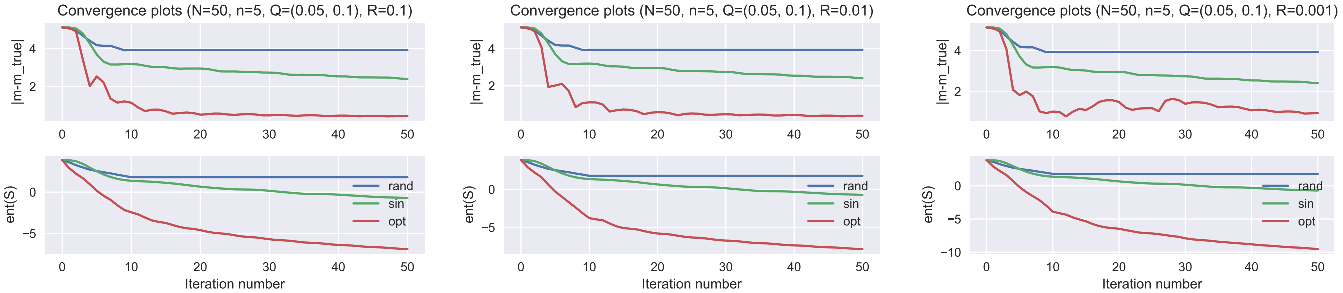
(a) high action cost, $R = 0.1$ (b) medium action cost, $R = 0.01$ (c) low action cost, $R = 0.001$

Figure 1: Effects of the action cost R on the closed-loop system performance. Trajectories are executed on Pendulum-v0 using belief-space MPC with horizon N and replanning every n steps. System noise Q is fixed and the action cost R is varying. Three scenarios are shown. In (a), the action cost is high, therefore the controller quickly pumps the energy into the system and fades away to observe the oscillations; this is possible because Pendulum-v0 is frictionless (although the controller has a non-zero prior on the friction coefficient). In (b), the cost of actions is lower, therefore the controller can enjoy taking larger actions a bit longer. In (c), the controller gets unstable, probably because the reward function is quite flat without action regularization and the action limits are too small to escape the flat region.



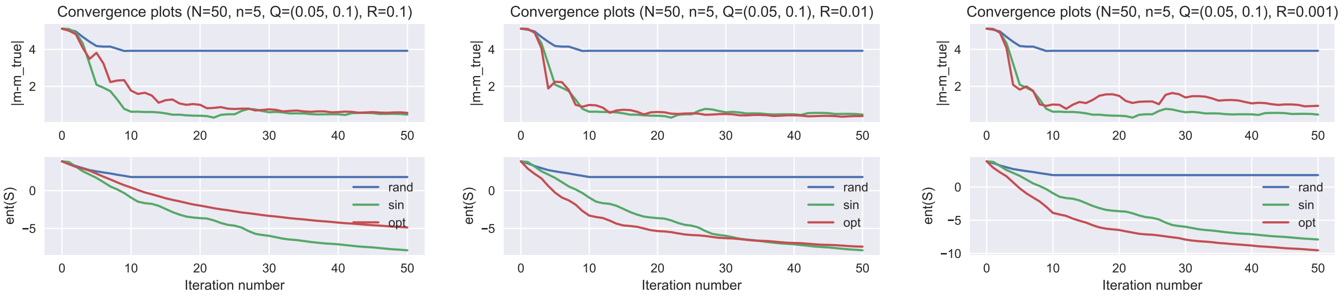
(a) random actions do not explore (b) slow sinusoid — sufficient excitation (c) fast sinusoid — best coverage

Figure 2: Compared to the optimal controls, random actions (a) perform very badly because they fail to explore the state space. On the other hand, a naive sinusoidal signal (b) works quite well on the pendulum, making it swing in all kinds of ways. However, the quality of system identification crucially depends on finding the right frequency of the sinusoid. A more oscillatory signal (c) turns out to be better for system identification (see convergence plots below).



(a) high action cost, $R = 0.1$ (b) medium action cost, $R = 0.01$ (c) low action cost, $R = 0.001$

Figure 3: Convergence plots show how quickly the posterior concentrates around the true parameter value; convergence in terms of distance from the mean and in terms of entropy of the posterior are shown. The posterior is updated after every n steps in the environment with the newly obtained data; one iteration on the x-axis corresponds to one posterior update. Three excitation signals are compared: random actions (blue), slow sinusoid (green), and optimal controls (red). Three scenarios are displayed from left to right that correspond to different action costs; only the red curve is different among the subplots, the other two curves are the same and kept for reference. All subplots demonstrate that the optimal excitation controls are significantly better than random or sinusoidal ones. Subplots (a) and (b) show similar red curves, which means that optimization is insensitive to the choice of the action cost in a reasonable range. Subplot (c) demonstrates that extremely low action costs may lead to oscillations; also observe that the final entropy in (c) is lower, meaning that the controller is more certain in the end.



(a) high action cost, $R = 0.1$

(b) medium action cost, $R = 0.01$

(c) low action cost, $R = 0.001$

Figure 4: A properly chosen excitation signal can yield very good results. These plots show that using a faster sinusoid (green), one can obtain as good parameter estimates as with an optimal signal. In (a), the fast sinusoid discovers the correct value faster and in the end it is even more certain than the optimal controller. In (b), both the optimal controls and the sinusoid perform on par. In (c), the posterior mean found with the optimal actions is further away from the true value and at the same time the controller is more confident about it; this shows the importance of the choice of costs.

References

[Andersson et al., 2012] Andersson, J., Åkesson, J., and Diehl, M. (2012). Casadi: A symbolic package for automatic differentiation and optimal control. In *Recent advances in algorithmic differentiation*, pages 297–307. Springer.

[Asmuth and Littman, 2011] Asmuth, J. and Littman, M. (2011). Learning is planning: near bayes-optimal reinforcement learning via monte-carlo tree search. In *Proceedings of the Twenty-Seventh Conference on Uncertainty in Artificial Intelligence*, pages 19–26. AUAI Press.

[Atkeson, 1989] Atkeson, C. G. (1989). Learning arm kinematics and dynamics. *Annual review of neuroscience*, 12(1):157–183.

[Bishop, 2006] Bishop, C. M. (2006). *Pattern recognition and machine learning*. Springer.

[Bombois et al., 2011] Bombois, X., Gevers, M., Hildebrand, R., and Solari, G. (2011). Optimal experiment design for open and closed-loop system identification. *Communications in Information and Systems*, 11(3):197–224.

[Brockman et al., 2016] Brockman, G., Cheung, V., Pettersson, L., Schneider, J., Schulman, J., Tang, J., and Zaremba, W. (2016). Openai gym. *arXiv preprint arXiv:1606.01540*.

[Feldbaum, 1960] Feldbaum, A. (1960). Dual control theory. i. *Avtomatika i Telemekhanika*, 21(9):1240–1249.

[Huan and Marzouk, 2016] Huan, X. and Marzouk, Y. M. (2016). Sequential bayesian optimal experimental design via approximate dynamic programming. *arXiv preprint arXiv:1604.08320*.

[Kahn et al., 2015] Kahn, G., Sujan, P., Patil, S., Bopardikar, S., Ryde, J., Goldberg, K., and Abbeel, P. (2015). Active exploration using trajectory optimization for robotic grasping in the presence of occlusions. In *2015 IEEE International Conference on Robotics and Automation (ICRA)*, pages 4783–4790. IEEE.

[Larsson et al., 2013] Larsson, C. A., Annergren, M., Hjalmarsson, H., Rojas, C. R., Bombois, X., Mesbah, A., and Modén, P. E. (2013). Model predictive control with integrated experiment design for output error systems. In *2013 European Control Conference (ECC)*, pages 3790–3795. IEEE.

[Lindley et al., 1956] Lindley, D. V. et al. (1956). On a measure of the information provided by an experiment. *The Annals of Mathematical Statistics*, 27(4):986–1005.

[Mehra, 1974] Mehra, R. (1974). Optimal input signals for parameter estimation in dynamic systems–survey and new results. *IEEE Transactions on Automatic Control*, 19(6):753–768.

[Ryan et al., 2016] Ryan, E. G., Drovandi, C. C., McGree, J. M., and Pettitt, A. N. (2016). A review of modern computational algorithms for bayesian optimal design. *International Statistical Review*, 84(1):128–154.

[Stratonovich, 1968a] Stratonovich, R. (1968a). *Conditional Markov processes and their application to the theory of optimal control*. Elsevier.

[Stratonovich, 1968b] Stratonovich, R. (1968b). Is there a theory of synthesis of optimal adaptive, self learning and self adjusting systems? *Avtomat. i Telemekh*, 29(1):96–107.