

---

# Model-based Policy Search for Partially Measurable Systems

---

**Fabio Amadio\***  
amadiofa@dei.unipd.it

**Alberto Dalla Libera\***  
dallaliber@dei.unipd.it

**Ruggero Carli\***  
carlirug@dei.unipd.it

**Daniel Nikovski†**  
nikovski@merl.com

**Diego Romeres†**  
romeres@merl.com

## Abstract

In this paper, we propose a Model-Based Reinforcement Learning (MBRL) algorithm for Partially Measurable Systems (PMS), i.e., systems where the state can not be directly measured, but must be estimated through proper state observers. The proposed algorithm, named Monte Carlo Probabilistic Inference for Learning COntrol for Partially Measurable Systems (MC-PILCO4PMS), relies on Gaussian Processes (GPs) to model the system dynamics, and on a Monte Carlo approach to update the policy parameters. W.r.t. previous GP-based MBRL algorithms, MC-PILCO4PMS models explicitly the presence of state observers during policy optimization, allowing to deal PMS. The effectiveness of the proposed algorithm has been tested both in simulation and in two real systems.

## 1 Introduction

Reinforcement Learning (RL) [1] has achieved outstanding results in many different environments. MBRL algorithms seem a promising solution to apply RL to real systems, due to their data-efficiency w.r.t. model-free RL algorithms. In particular, remarkable results have been obtained relying on Gaussian Processes (GPs) [2] to model the systems dynamics, see for instance [3, 4, 5, 6]. In this paper, we consider the application of MBRL algorithms to PMS, i.e., systems where only a subset of the state components can be directly measured, and the remaining components can be estimated through proper state observer. PMS are particularly relevant in real world applications, think for example to mechanical systems, where, typically, only positions are measured, while velocities are estimated through numerical differentiation or more complex filters. The proposed algorithm, named MC-PILCO4PMS, relies on Gaussian Processes (GPs) to model the system dynamics, and on a Monte Carlo approach [7] to optimize the policy parameters. W.r.t. previous GP-based MBRL algorithms, such as [3, 4, 5], MC-PILCO4PMS models explicitly the presence of two different state observers during the two phases of model learning and of policy optimization. This improves the characterization of the PMS in the two phases and so the control performance. In the following we provide a description of the proposed algorithm, assuming that it is applied to mechanical systems where only positions measurement are available. However, the algorithm generalizes to any PMS.

## 2 Problem Setting

Consider a mechanical system with  $d_q$  degrees of freedom, and denote with  $x_t = [q_t^T, \dot{q}_t^T]^T$  its state, where  $q_t \in \mathbb{R}^{d_q}$  and  $\dot{q}_t \in \mathbb{R}^{d_q}$  are, respectively, the vector of the generalized coordinates and

---

\*Department of Information Engineering, University of Padova, Via Gradenigo 6/B, 35131 Padova, Italy

†Mitsubishi Electric Research Laboratories (MERL), Cambridge, MA 02139

its derivative w.r.t. time. Assume that joint positions can be directly measured, while  $\dot{\mathbf{q}}_t$  must be estimated from the history of  $\mathbf{q}_t$  measurements. Moreover, let the system be Markovian, and describe its discrete-time dynamics as  $\mathbf{x}_{t+1} = f(\mathbf{x}_t, \mathbf{u}_t) + \mathbf{w}_t$ , where  $f(\cdot)$  is an unknown transition function,  $\mathbf{u}_t \in \mathbb{R}^{d_u}$  represents the system input, while  $\mathbf{w}_t \sim \mathcal{N}(0, \Sigma_w)$  models uncertainty. The objective of RL algorithms is learning to accomplish a given task based on interaction data. The task is encoded in a cost function  $c(\mathbf{x}_t)$ , defined to characterize the immediate penalty for being in state  $\mathbf{x}_t$ . The system inputs are chosen in accordance with a policy  $\pi_\theta : \mathbf{x} \mapsto \mathbf{u}$  that depends on the parameter vector  $\theta$ . Then, the objective is to find the policy that minimizes the expected cumulative cost over a finite number of time steps  $T$ , with initial state distribution  $p(\mathbf{x}_0)$ , i.e.,  $J(\theta) = \sum_{t=0}^T \mathbb{E}_{\mathbf{x}_t} [c(\mathbf{x}_t)]$ .

### 3 Method

MC-PILCO4PMS consists of the iteration of three phases: (i) model learning, (ii) policy optimization, and (iii) policy execution. In the first phase, MC-PILCO4PMS relies on GPR to estimate the one-step-ahead system dynamics, while for the optimization of the policy parameters, MC-PILCO4PMS implements a gradient-based strategy. In the following, we briefly discuss the two phases.

#### 3.1 Model Learning

**Dynamics model.** The proposed one-step-ahead GP model exploits the intrinsic correlation between the position and velocity. In our algorithm a distinct GP model is learned to predict the velocity change, while positions are obtained by integration. This approach is different from previous GP-based MBRL algorithms, such as [3, 4, 5], that learn one independent model for each state component. Let us indicate the components of  $\mathbf{q}_t$  and  $\dot{\mathbf{q}}_t$  with  $q_t^{(i)}$  and  $\dot{q}_t^{(i)}$ , respectively, where  $i \in \{1, \dots, d_q\}$ . Then, let  $\Delta_t^{(i)} = \dot{q}_{t+1}^{(i)} - \dot{q}_t^{(i)}$  be the difference between the value of the  $i$ -th velocity at time  $t+1$  and  $t$ , and  $y_t^{(i)}$  the noisy measurement of  $\dot{q}_t^{(i)}$ . For each velocity component  $i$ , we model  $\Delta_t^{(i)}$  with a distinct GP with zero mean and kernel function  $k(\cdot, \cdot)$ , which takes as input  $\mathbf{x}_t = [\mathbf{x}_t, \mathbf{u}_t]$ . Details on the kernel choice can be found in Appendix 6.1. In GPR the posterior distribution of  $\Delta_t^{(i)}$  given the data is Gaussian, with mean and covariance available in closed form, see [2]. Then, given the GP input  $\mathbf{x}_t = [\mathbf{x}_t, \mathbf{u}_t]$ , a prediction of the velocity changes  $\hat{\Delta}_t^{(i)}$  can be sampled from the aforementioned posterior distribution. When considering a sufficiently small sampling time  $T_s$ , it is reasonable to assume constant accelerations between two consecutive time-steps, and the predicted positions and velocities are obtained with the following equations,  $\hat{q}_{t+1}^{(i)} = q_t^{(i)} + T_s \dot{q}_t^{(i)} + \frac{T_s^2}{2} \hat{\Delta}_t^{(i)}$  and  $\hat{\dot{q}}_{t+1}^{(i)} = \dot{q}_t^{(i)} + \hat{\Delta}_t^{(i)}$  for  $i \in \{1, \dots, d_q\}$ .

**Training data computation.** As described before, velocities are not accessible and have to be estimated from measurements of positions. Notice that the velocity estimates used to train the GP models can be computed offline, exploiting the (past and future) history of measurements to improve accuracy. Well filtered data, that resemble the real states of the system, improve significantly the adherence between the learnt model and the real system. In our experiments, we computed offline the velocities used to train the GPs, using for example, the central difference formula, i.e.,  $\hat{\dot{\mathbf{q}}}_t = (\mathbf{q}_{t+1} - \mathbf{q}_{t-1}) / (2T_s)$ , which is an acausal filter. We would like to underline that these state estimates are different from the ones computed real-time and provided to the control policy during system interaction. Typically, due to real-time constraints, online estimates are less accurate and it is fundamental to keep this in consideration during policy optimization as we can see in the following.

#### 3.2 Policy optimization

MC-PILCO4PMS optimizes the policy parameters with a gradient-based strategy. At each optimization step the algorithm performs the following operations: (i) approximation of the cumulative cost relying on a Monte Carlo approximation; (ii) computation of the gradient and update of  $\theta$ . More precisely, the algorithm samples  $M$  particles from the initial state distribution  $p(\mathbf{x}_0)$ , and simulates their evolution for  $T$  steps. At each simulation step the inputs are selected in accordance with the current policy, and the next state is predicted with the GP models previously described. This procedure models the propagation of the model uncertainty for long-term predictions. Then, the Monte Carlo estimate of the cumulative cost is  $\hat{J}(\theta) = \sum_{t=0}^T \frac{1}{M} \sum_{m=1}^M c(\mathbf{x}_t^{(m)})$ , where  $\mathbf{x}_t^{(m)}$

denotes the state of the  $m$ -th particle at time  $t$ . The gradient is computed by backpropagation on the computational graph of  $\hat{J}(\theta)$ , exploiting the reparametrization trick [8] to propagate the gradient through the stochastic operations, i.e., the sampling from the GP posteriors distribution. Advantages of MC based long-term predictions w.r.t to e.g., moment matching [3] are that no assumptions on the state distribution and on the kernel function in the GP models have to be made. The policy parameters are updated using the Adam solver [9]. In the remainder of this section we describe the particles simulation, which is the main novelty introduced to deal with PMS.

**Particles simulation with PMS.** In order to deal with PMS we not only simulate the evolution of the system state, but also the evolution of the observed states, modeling the measurement system and the online state observers implemented in the real system. A block scheme of the particles generation is reported in Fig.1. Let  $\mathbf{x}_t^{(m)} = [\mathbf{q}_t^{(m)}, \dot{\mathbf{q}}_t^{(m)}]$  be the state of the  $m$ -th particle at the simulation step  $t$  predicted by the GP models. In order to transform the prediction of the system state to the observed state, firstly, we simulate the measurement system by corrupting positions with a zero mean Gaussian i.i.d noise  $e_t^{(m)}$ :  $\mathbf{q}_t^{(m)} = \mathbf{q}_t^{(m)} + e_t^{(m)}$ . Secondly, the measured states are used to compute an estimate of the observed states:  $\mathbf{z}_t^{(m)} = f_z(\mathbf{q}_t^{(m)} \dots \mathbf{q}_{t-m_q}^{(m)}, \mathbf{z}_{t-1}^{(m)} \dots \mathbf{z}_{t-1-m_z}^{(m)})$ , where  $f_z$  is the online state observer implemented in the real system, with memory  $m_q$  and  $m_z$ . Finally, the control inputs for each particle are computed as  $\pi(\mathbf{z}_t^{(m)})$ , the next particles states are sampled from the GP dynamics, and the procedure is iterated. The procedure aims at obtaining robustness w.r.t. delays and distortions introduced by measurement noise and online observers. Notice that selecting the inputs as  $\pi(\mathbf{z}_t^{(m)})$ , as done in several previous MBRL algorithms, is equivalent to assume full access to the system state, which is often an unrealistic assumption when dealing with real systems, since the difference between the system state and the observed state might be significant. This is a key differentiation of our method. Let us denote, MC-PILCO, the version of the proposed algorithm which assumes fully access to the system state during policy optimization. A numerical comparison between MC-PILCO and two state-of-the-art GP-based MBRL algorithms is reported in the Appendix 6.3. The results obtained show that MC-PILCO overperforms both the algorithms in terms of data-efficiency and accuracy.

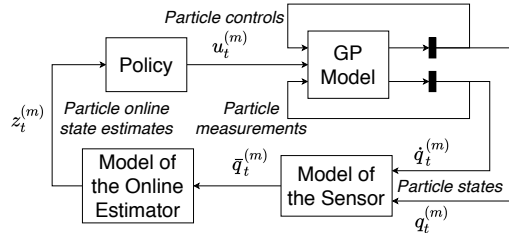


Figure 1: Block schemes illustrating particles generation in MC-PILCO4PMS.

## 4 Experiments

MC-PILCO4PMS has been tested both in simulation and in real systems. First, we validate in simulation the impact of taking into consideration the measurement system and the online filter during particle simulation. Second, MC-PILCO4PMS has been successfully applied to two real systems: a Furuta pendulum and a Ball-and-Plate system. Further details about the implementation of the algorithm on the presented systems can be found in Appendices 6.2, 6.4, 6.5.

**Simulation as proof of concepts.** Here, we test the relevance of modeling the presence of online observers on a simulated cart-pole system. The objective is to learn a policy able to swing-up the pole and stabilize it in the upwards equilibrium, while keeping the cart stationary. We assumed to be able to measure only the cart position and the pole angle. The online estimates of the velocities were computed by means of causal numerical differentiation followed by a first order low-pass filter. The velocities used to train the GPs were derived with the central difference formula. Two policy functions have been trained: the first has been derived with MC-PILCO, assuming direct access to the full state predicted by the model; the second policy has been derived using MC-PILCO4PMS. In Figure 2, we report the results of a Monte Carlo study with 400 runs. Even though the two policies perform similarly when applied to the learned models, the results obtained with the cart-pole system are significantly different. MC-PILCO4PMS solves the task in all 400 attempts. In contrast, in several attempts, the MC-PILCO policy does not solve the task, due to delays and discrepancies introduced by the online filter and not considered during policy optimization.

Figure 2: Comparison of 400 simulated particles rollout (left) and the trajectories performed applying repetitively the policy 400 times in the system (right) with the simulated cart-pole system. MC-PILCO results are on the top plots, while MC-PILCO4PMS are on the bottom.

Figure 3: Trajectories for the pendulum angle (left) and arm angle (right) obtained at each trial. For all the kernels, the angles are plotted up to the trial that solved the task.

Figure 4: Ten different ball trajectories obtained on the Ball-and-Plate under the policy learned by MC-PILCO4PMS. Steady-state positions are marked with black crosses. The dashed circle has the same diameter of the used ball.

Furuta Pendulum. The Furuta pendulum [10] is a popular under-actuated benchmark system that consists of a driven arm, revolving in the horizontal plane, with a pendulum attached to its end, which rotates in the vertical plane. Let  $\theta$  be the horizontal angle of the arm, and  $\phi$  the vertical angle of the pendulum. The objective is to learn a controller able to swing-up the pendulum and stabilize it in the upwards equilibrium ( $\phi = \pi$  [rad]) with  $\dot{\theta} = 0$  [rad]. Of the estimates of velocities for the GP model have been computed by means of central differences. Causal numerical differentiation were used for the online estimation. MC-PILCO4PMS managed to solve the task using the three different choices of kernel functions presented in Appendix 6.1. In Figure 3, we show the resulting trajectories for each trial. These experiments show the effectiveness of MC-PILCO4PMS and confirm the higher data efficiency of more structured kernels, which is one of the advantage that MC-PILCO4PMS offers by allowing any kernel function while in methods like PILCO the kernel choice is limited. For best of our knowledge, with 9 [s] of training data this algorithm is the most data-efficient to solve a FP.

Ball-and-Plate. The ball-and-plate system is composed of a square plate that can tilt in two orthogonal directions by means of two motors. On top of it, there is a camera to track the ball and measure its position on the plate. The objective of the experiment is to learn how to control the motor angles in order to stabilize the ball around the center of the plate. Measurements provided by the camera are very noisy, and cannot be used directly to estimate velocities from positions. We used a Kalman smoother [11] for the offline filtering of ball positions and associated velocities. Instead, in real-time we used a Kalman filter [12] to estimate online the ball state from noisy measures of positions. MC-PILCO4PMS learnt a policy to stabilize the ball around the center starting from any initial position after the third trial, 11.33 [s] of interaction with the system. We tested the learned policy starting from ten different points, see Figure 4. The mean steady-state error, i.e. the average distance of the final ball position from the center observed in the ten trials, was 0.0099 [m], while the maximum measured error was 0.0149 [m], which is lower than the ball radius of 0.016 [m].

## 5 Conclusions

We have presented a MBRL algorithm called, MC-PILCO4PMS, which does not assume that all the components of the state can be measured and we successfully applied it to robotic systems. The algorithm employs GPs to derive a probabilistic model of the system dynamics. Policy parameters are updated through a Monte Carlo gradient-based strategy: expected cumulative cost is estimated by averaging over hundreds of simulated rollouts, and policy gradient is computed by backpropagation on the resulting computational graph. We showed the importance of manipulating the measurements to both provide accurate state estimates to the model learning algorithm and to reproduce the measurement system together with the online state observer during policy optimization.

## References

- [1] Richard S Sutton and Andrew G Barto. Reinforcement learning: An introduction. MIT press, 2018.
- [2] Christopher KI Williams and Carl Edward Rasmussen. Gaussian processes for machine learning. MIT press Cambridge, MA, 2006.
- [3] Marc Deisenroth and Carl E Rasmussen. Pilco: A model-based and data-efficient approach to policy search. In Proceedings of the 28th International Conference on machine learning (ICML-11), pages 465–472, 2011.
- [4] Paavo Parmas, Carl Edward Rasmussen, Jan Peters, and Kenji Doya. Pippis: Flexible model-based policy search robust to the curse of chaos. In International Conference on Machine Learning pages 4065–4074, 2018.
- [5] Konstantinos Chatzilygeroudis, Roberto Rama, Rituraj Kaushik, Dorian Goepp, Vassilis Vassiliades, and Jean-Baptiste Mouret. Black-box data-efficient policy search for robots. In 2017 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), pages 51–58. IEEE, 2017.
- [6] Diego Romeres, Devesh K Jha, Alberto Dalla Libera, Bill Yezazunis, and Daniel Nikovski. Semiparametrical gaussian processes learning of forward dynamical models for navigating in a circular maze. In 2019 International Conference on Robotics and Automation (ICRA), pages 3195–3202. IEEE, 2019.
- [7] Russel E Ca isch et al. Monte carlo and quasi-monte carlo methods. Acta numerica 1998:1–49, 1998.
- [8] Diederik P Kingma and Max Welling. Auto-encoding variational bayes. arXiv preprint arXiv:1312.6114 2013.
- [9] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980 2014.
- [10] Benjamin Seth Cazzolato and Zebb Prime. On the dynamics of the furuta pendulum. Journal of Control Science and Engineering 2011, 2011.
- [11] Garry A Einicke. Optimal and robust noncausal lter formulations. IEEE Transactions on Signal Processing 54(3):1069–1077, 2006.
- [12] R. E. Kalman. A New Approach to Linear Filtering and Prediction Problems. Journal of Basic Engineering 82(1):35–45, 03 1960.
- [13] Alberto Dalla Libera, Ruggero Carli, and Gianluigi Pillonetto. A novel multiplicative polynomial kernel for volterra series identification. arXiv preprint arXiv:1905.07960 2019.
- [14] A. D. Libera and R. Carli. A data-efficient geometrically inspired polynomial kernel for robot inverse dynamics. IEEE Robotics and Automation Letters 5(1):24–31, 2020.
- [15] D. Nguyen-Tuong and J. Peters. Using model knowledge for learning inverse dynamics. In 2010 IEEE International Conference on Robotics and Automation, pages 2677–2682, 2010.

## 6 Appendix

### 6.1 Kernel functions

One of the advantages of the particle-based policy optimization method is the possibility of choosing any kernel functions without restrictions. Hence, we considered different kernel functions as examples to model the evolution of physical systems. But the reader can consider a custom kernel function appropriate for his application.

**Squared exponential (SE)** The SE kernel represents the standard choice adopted in many different works. It is defined as

$$k_{SE}(\mathbf{x}_{t_j}; \mathbf{x}_{t_k} | \mathbf{q}) = \sigma^2 e^{-\frac{1}{2} \mathbf{x}_{t_j}^T \mathbf{K}^{-1} \mathbf{x}_{t_k}}, \quad (1)$$

where the scaling factor and the matrix  $\mathbf{K}$  are kernel hyperparameters which can be estimated by marginal likelihood maximization. Typically,  $\mathbf{K}$  is assumed to be diagonal, with the diagonal elements named lengthscales.

**SE + Polynomial (SE+P<sup>pdq</sup>)**. Recalling that the sum of kernels is still a kernel [2], we considered also a kernel given by the sum of a SE and a polynomial kernel. In particular, we used the Multiplicative Polynomial (MP) kernel, which is a refinement of the standard polynomial kernel, introduced in [4]. The MP kernel of degree  $d$  is defined as the product of linear kernels, namely,

$$k_P^{pdq}(\mathbf{x}_{t_j}; \mathbf{x}_{t_k} | \mathbf{q}) = \prod_{r=1}^d \frac{1}{\beta_r} \mathbf{x}_{t_j}^T \mathbf{P}_r \mathbf{x}_{t_k}.$$

where the  $\mathbf{P}_r$ ,  $\beta_r > 0$  matrices are distinct diagonal matrices. The diagonal elements of  $\mathbf{P}_r$  together with the  $\beta_r$  elements are the kernel hyperparameters. The resulting kernel is

$$k_{SE+P}^{pdq}(\mathbf{x}_{t_j}; \mathbf{x}_{t_k} | \mathbf{q}) = k_{SE}(\mathbf{x}_{t_j}; \mathbf{x}_{t_k} | \mathbf{q}) k_P^{pdq}(\mathbf{x}_{t_j}; \mathbf{x}_{t_k} | \mathbf{q}) \quad (2)$$

The idea motivating this choice is the following: the MP kernel allows capturing possible modes of the system that are polynomial functions, which are typical in mechanical systems [14], while the SE kernel models more complex behaviors not captured by the polynomial kernel.

**Semi-Parametrical (SP)** When prior knowledge about the system dynamics is available, for example given by physics first principles, the so called physically inspired (PI) kernel can be derived. The PI kernel is a linear kernel defined on suitable basis functions  $\mathbf{p}(\mathbf{x}; \mathbf{q})$ , see for instance [6]. More precisely,  $\mathbf{p}(\mathbf{x}; \mathbf{q}) \in \mathbb{R}^d$  is a, possibly nonlinear, transformation of the GP input  $\mathbf{x}$  determined by the physical model. Then we have

$$k_{PI}(\mathbf{x}_{t_j}; \mathbf{x}_{t_k} | \mathbf{q}) = \mathbf{p}(\mathbf{x}_{t_j}; \mathbf{q})^T \mathbf{P}_I \mathbf{p}(\mathbf{x}_{t_k}; \mathbf{q})$$

where  $\mathbf{P}_I$  is a  $d \times d$  positive-definite matrix, whose elements are the hyperparameters; to limit the number of hyperparameters, a standard choice consists in considering  $\mathbf{P}_I$  to be diagonal. To compensate possible inaccuracies of the physical model, it is common to combine an SE kernel, obtaining so called semi-parametric kernels [15, 6], expressed as

$$k_{SP}(\mathbf{x}_{t_j}; \mathbf{x}_{t_k} | \mathbf{q}) = k_{PI}(\mathbf{x}_{t_j}; \mathbf{x}_{t_k} | \mathbf{q}) + k_{SE}(\mathbf{x}_{t_j}; \mathbf{x}_{t_k} | \mathbf{q})$$

The rationale behind this kernel is the following:  $k_{PI}$  encodes the prior information given by the physics, and  $k_{SE}$  compensates for the dynamical components unmodelled in

### 6.2 Simulated Cart-pole

The physical properties of the cart-pole system considered are the following: the masses of both cart and pole are 0.5 [kg], the length of the pole is 0.5 [m], and the coefficient of friction between cart and ground is 0.1. The state at each time step is defined as  $\mathbf{x}_t = [r; \dot{r}; \theta; \dot{\theta}]^T$ , where  $r$  represents the position of the cart and  $\theta$  the angle of the pole. The target states corresponding to the swing-up of the pendulum is given by  $\mathbf{x}^{des} = [0; 0; \pi; 0]^T$  [m] and  $[\text{rad}]$ . The downward stable equilibrium point is defined at  $\theta = 0$  [rad]. As done in [3], in order to avoid singularities due to the angles,  $\theta$  is replaced with the state representation  $[\sin \theta; \cos \theta]^T$  inside GP inputs. For the GP models SE kernels have been chosen. The control action is the force that pushes the cart horizontally. We considered white measurement noise with standard deviation of  $10^{-3}$ , and as initial state distribution  $\mathcal{N}(\mathbf{p}; \mathbf{0}; \mathbf{0}; \text{diag}(10^{-4}; 10^{-4}; 10^{-4}; 10^{-4}))$  in order to obtain reliable

|             | Trial 1 | Trial 2 | Trial 3 | Trial 4 | Trial 5 |
|-------------|---------|---------|---------|---------|---------|
| PILCO       | 2%      | 4%      | 20%     | 36%     | 42%     |
| Black-DROPS | 0%      | 4%      | 30%     | 68%     | 86%     |
| MC-PILCO    | 0%      | 14%     | 78%     | 94%     | 100%    |

Table 1: Success rate per trial obtained with PILCO, Black-DROPS and MC-PILCO.

estimates of the velocities, samples were collected at 30 [Hz]. The number of particles has been set to  $M = 400$  in all the tests.

The cost function optimized in MC-PILCO is the following,

$$c_{p,t,q} = 1 - \exp\left(-\frac{|t|}{l} - \frac{p_t}{l_p}\right)^2; \quad (3)$$

where  $l$  and  $l_p$  are named lengthscales. Notice that the lengthscales define the shape of the cost function goes to its maximum value more rapidly with small lengthscales. Therefore, higher cost is associated to the same distance from the target state with lower  $l_p$ . The lower the lengthscales the more selective the cost function. The absolute value is needed to allow different swing-up solutions to both the equivalent target angles of the pole. The selected lengthscales were  $l = 3$  and  $l_p = 1$ .

The policy adopted is an RBF network policy with outputs limited by an hyperbolic tangent function, properly scaled. We call this function squashed-RBF-network and it is defined as

$$p_{t,q} = u_{\max} \tanh\left(\frac{1}{u_{\max}} \sum_{i=1}^{n_b} w_i e^{-a_i \|x_t\|^2}\right). \quad (4)$$

parameters are  $t, w, A, u$ , where  $r, w_1, \dots, w_{n_b}$  and  $A, t, a_1, \dots, a_{n_b}, u$  are, respectively, the weights and the centers of the Gaussian basis functions, while  $\sigma$  determines their shapes; in all experiments we assumed  $\sigma$  to be diagonal.  $u_{\max}$  is the maximum control action applicable. For this experiment we choose  $n_b = 200$  basis functions and  $u_{\max} = 10$  [N]. The exploration trajectory is obtained by applying at each time step a random control action sampled from  $u_{\max}; u_{\max} q$ .

### 6.3 Comparison with state of the art algorithms

We tested PILCO [5], Black-DROPS [5] and MC-PILCO on the simulated cart-pole system. The setup is equal to the one described in Appendix 6.2, with the only two difference that here the samples are collected at 20 [Hz] and the noise standard deviation is  $10^{-2}$ . In PILCO and Black-DROPS, we considered their original cost function,

$$c^{\text{pilco}}_{p,t,q} = 1 - \exp\left(-\frac{1}{2} \frac{d_t^2}{0.25}\right); \quad (5)$$

where  $d_t^2 = p_t^2 + 2p_t L \sin p_{t,q} + 2L^2 p_1^2 \cos p_{t,q}$  is the squared distance between the tip of the pole and its position at the unstable equilibrium point with  $p_t = 0$  [m]. This last cost is also adopted as a common metric to compare the results obtained by the three algorithms. Results of the cumulative cost are reported in Figure 5.

Figure 5: Median and confidence interval of the cumulative cost  $c^{\text{pilco}}_{p,q}$  per trial obtained with PILCO, Black-DROPS and MC-PILCO. MC-PILCO achieved the best performance both in transitory and at convergence, by trial 5, it learned how to swing up the cart-pole with a success rate of 100%. In each and every trial, MC-PILCO obtained cumulative costs with lower median and less variability. On the other hand, the policy in PILCO showed poor convergence

properties with only 42% of success rate after all the 5 trials. Black-DROPS outperforms PILCO, but it obtained worse results than MC-PILCO in each and every trial, with a success rate of only 86% at trial 5.

#### 6.4 Furuta Pendulum

The Furuta pendulum (FP) [10] is a popular benchmark system used in nonlinear control and reinforcement learning. The system is composed of two revolute joints and three links. The first link, called the base link, is fixed and perpendicular to the ground. The second link, called arm, rotates parallel to the ground, while the rotation axis of the last link, the pendulum, is parallel to the principal axis of the second link, see Figure 6. The FP is an under-actuated system as only the first joint is actuated. In particular, in the FP considered the horizontal joint is actuated by a DC servomotor, and the two angles are measured by optical encoders with 4096 [ppr]. The control variable is the motor voltage. Let the state at time step  $t$  be  $\mathbf{x}_t = [\theta_t^h, \dot{\theta}_t^h, \theta_t^v, \dot{\theta}_t^v]^T$ , where  $\theta_t^h$  is the angle of the horizontal joint and  $\theta_t^v$  the angle of the vertical joint attached to the pendulum. The objective is to learn a controller able to swing-up the pendulum and stabilize it in the upwards equilibrium ( $\theta_t^v = \pm\pi$  [rad]) with  $\theta_t^h = 0$  [rad]. The trial length is 3 [s] with a sampling frequency of 30 [Hz]. The cost function is defined as

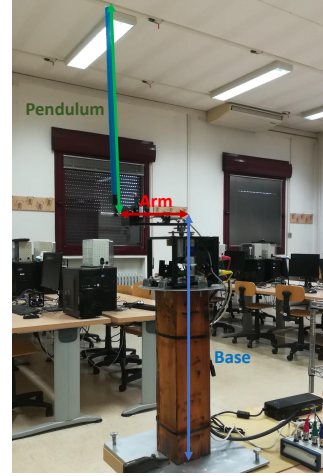


Figure 6: Furuta pendulum used in the experiment while being controlled in the upward equilibrium point by the learned policy.

$$c(\mathbf{x}_t) = 1 - \exp\left[-\frac{\theta_t^h}{2}\right]^2 - \frac{|\theta_t^v| - \pi}{2}\left[|\theta_t^v| - \pi\right]^2 + c_b(\mathbf{x}_t), \quad (6)$$

with

$$c_b(\mathbf{x}_t) = \frac{1}{1 + \exp\left[-10\left(-\frac{3}{4}\pi - \theta_t^h\right)\right]} + \frac{1}{1 + \exp\left[-10\left(\theta_t^h - \frac{3}{4}\pi\right)\right]}.$$

The first part of the function in (6) aims at driving the two angles towards  $\theta_t^h = 0$  and  $\theta_t^v = \pm\pi$ , while  $c_b(\mathbf{x}_t)$  penalizes solutions where  $\theta_t^h \leq -\frac{3}{4}\pi$  or  $\theta_t^h \geq \frac{3}{4}\pi$ . We set those boundaries to avoid the risk of damaging the system if the horizontal joint rotates too much. Offline estimates of velocities for the GP model have been computed by means of central differences. For the online estimation, we used causal numerical differentiation:  $\dot{\mathbf{q}}_t = (\mathbf{q}_t - \mathbf{q}_{t-1})/T_s$ , where  $T_s$  is the sampling time. Instead of  $\mathbf{x}_t$ , we considered the extended state  $\mathbf{x}_t = [\dot{\theta}_t^h, \dot{\theta}_t^v, \sin(\theta_t^h), \cos(\theta_t^h), \sin(\theta_t^v), \cos(\theta_t^v)]^T$  in GP input. The policy is a *squashed-RBF-network* with  $n_b = 200$  basis functions that receives as input  $\mathbf{z}_t = [(\theta_t^h - \theta_{t-1}^h)/T_s, (\theta_t^v - \theta_{t-1}^v)/T_s, \sin(\theta_t^h), \cos(\theta_t^h), \sin(\theta_t^v), \cos(\theta_t^v)]^T$ . We used 400 particles to estimate the policy gradient from model predictions. The exploration trajectory has been obtained using as input a sum of ten sine waves of random frequencies and same amplitudes. The initial state distribution is assumed to be  $\mathcal{N}([0, 0, 0, 0]^T, \text{diag}([5 \cdot 10^{-3}, 5 \cdot 10^{-3}, 5 \cdot 10^{-3}, 5 \cdot 10^{-3}]))$ .  $M = 400$  particles were used for gradient estimation.

#### 6.5 Ball-and-Plate

The ball-and-plate system is composed of a square plate that can be tilted in two orthogonal directions by means of two motors. On top of it, there is a camera to track the ball and measure its position on the plate. Let  $(b_t^x, b_t^y)$  be the position of the center of the ball along X-axis and Y-axis, while  $\theta_t^{(1)}$  and  $\theta_t^{(2)}$  are the angles of the two motors tilting the plate, at time  $t$ . So, the state of the system is defined as  $\mathbf{x}_t = [b_t^x, b_t^y, \dot{b}_t^x, \dot{b}_t^y, \theta_t^{(1)}, \theta_t^{(2)}, \dot{\theta}_t^{(1)}, \dot{\theta}_t^{(2)}]^T$ . The drivers of the motors allow only position control, and do not provide feedback about the motors angles. To keep track of the motor angles, we defined the control actions as the difference between two consecutive reference values sent to

