Tractable Bayesian Dynamics Priors from Differentiable Physics for Learning and Control

Joe Watson[1,2], Benedikt Hahner[2], Boris Belousov[1], Jan Peters[1,2] *Fellow, IEEE*

***Abstract*— Statistical model-based reinforcement learning methods should enable efficient data-driven policy optimization for robotic systems. However, the success of these approaches relies on how well the learned dynamics model generalizes outside of its training data distribution, which is difficult to ensure in practice with 'black-box' models unless inductive biases are incorporated. We demonstrate how a differentiable simulation model can be used to synthesize a tractable Gaussian process prior using the linearized Laplace approximation, a principled approximate inference technique for Gaussian posteriors. Using this statistical model as an inductive bias, we can perform exploration and decision-making in an informed way using posterior sampling, performing reinforcement learning for control, and active learning for system identification. In the case of simulation-to-reality mismatch, we empirically investigate how this physics-informed prior is best used, comparing architecture- and objective-based approaches.**

I. Problem Setting

We consider general episodic sequential decision-making, which we frame as a finite-horizon control task for a stochastic dynamical system, maximizing episode objective $\mathcal{J}_n$ given the dynamics $p(\boldsymbol{s}_{t+1} \mid \boldsymbol{s}_t, \boldsymbol{a}_t)$ and return R_n [1],

$$\mathcal{J}_n(\pi, p) = \mathbb{E}_{\boldsymbol{s}_{t+1}\sim p(\cdot|\boldsymbol{s}_t,\boldsymbol{a}_t),\, \boldsymbol{a}\sim\pi(\cdot|\boldsymbol{s}_t,t),\, \boldsymbol{s}_1\sim\mu(\cdot)}[R_n(\boldsymbol{S}_1^T, \boldsymbol{A}_1^T)],$$

where $\boldsymbol{X}_1^T = [\boldsymbol{x}_1, \dots, \boldsymbol{x}_T]$. This problem definition covers both decision-making for control and active model learning, where the control objective is a Bayesian regression model's expected information gain $\mathbb{E}_{\boldsymbol{y},\,\boldsymbol{\theta}\sim p(\cdot,\cdot|\boldsymbol{x})}[\log p(\boldsymbol{y}, \boldsymbol{\theta} \mid \boldsymbol{x}) - \log p(\boldsymbol{y} \mid \boldsymbol{x})]$. To incorporate uncertainty, we consider model-based methods that learn an approximate Bayesian dynamics model from data [2]–[7]. To improve the data efficiency of learning on a real-world system, we consider two directions. One is exploration, where we use posterior sampling [8], [9] of dynamics models as a principled model-based exploration strategy, which requires a posterior belief over model parameters. The other is inductive biases [10], where we seek to incorporate domain knowledge into the dynamics model to improve generalization. Both perspectives can be tackled using Bayesian methods, but require the use of approximate inference [11].

II. Physics-based Gaussian Process Priors

We obtain a Gaussian process (GP) [12] from a differentiable simulator [13] using the linearized Laplace approximation (LLA) [14]–[16]. The LLA linearizes the nonlinear model $\boldsymbol{f}$ about its maximum-a-posteriori (MAP) parameters, $\boldsymbol{\theta}_* = \arg\max_{\boldsymbol{\theta}} \log p(\mathcal{D}, \boldsymbol{\theta}) + \log p(\boldsymbol{\theta})$ with data $\mathcal{D}$,

$$\hat{\boldsymbol{f}}_{\boldsymbol{\theta}_*}(\boldsymbol{x}, \boldsymbol{\theta}) \approx \boldsymbol{f}(\boldsymbol{x}, \boldsymbol{\theta}_*) + \boldsymbol{J}(\boldsymbol{x}, \boldsymbol{\theta}_*)^\top (\boldsymbol{\theta} - \boldsymbol{\theta}_*). \tag{1}$$

[1]DFKI, Germany, [2]Technical University of Darmstadt, Germany

This approximation yields a Gaussian parameter posterior $q(\boldsymbol{\theta}) = \mathcal{N}(\boldsymbol{\mu}_q, \boldsymbol{\Sigma}_q)$ given a likelihood $\mathcal{N}(\boldsymbol{y}, \boldsymbol{\Sigma}_\epsilon)$, and a closed-form predictive distribution and posterior covariance,

$$q(\boldsymbol{y} \mid \boldsymbol{x}) = \mathcal{N}(\boldsymbol{f}(\boldsymbol{x}, \boldsymbol{\theta}_*), \boldsymbol{J}(\boldsymbol{x}, \boldsymbol{\theta}_*)\, \boldsymbol{\Sigma}_q\, \boldsymbol{J}(\boldsymbol{x}, \boldsymbol{\theta}_*)^\top + \boldsymbol{\Sigma}_\epsilon),$$
$$\boldsymbol{\Sigma}_q^{-1} = \boldsymbol{\Sigma}_0^{-1} - \nabla^2_{\boldsymbol{\theta}} \log p(\boldsymbol{y} \mid \boldsymbol{f}(\boldsymbol{x}, \boldsymbol{\theta}))|_{\boldsymbol{\theta}=\boldsymbol{\theta}_*}. \tag{2}$$

The Laplace approximation naturally transfers to differentiable simulators with only minor modification, as now $\boldsymbol{\theta}$ captures unknown physical parameters such as mass or length. In practice, we place our belief in an unconstrained *virtual* parameter space $\tilde{\boldsymbol{\theta}} \in \mathbb{R}^p$, and define a mapping that incorporates the parameter constraints, such as positivity. This can be done by combining any $\mathbb{R} \to \mathbb{R}_+$ bijection, such as the sigmoid function, with rectangular parameter constraints $\boldsymbol{\theta} \in [\boldsymbol{\theta}_{\text{min}}, \boldsymbol{\theta}_{\text{max}}]$, so $\boldsymbol{\theta}$ is reparameterized as $\boldsymbol{\theta} = \boldsymbol{\theta}_{\text{min}} + \boldsymbol{\theta}_{\text{max}}\, \text{Sigmoid}(\tilde{\boldsymbol{\theta}})$. Such reparameterizations have been adopted in prior works, e.g., [17]. LLA uses the virtual parameter space, so linearization combines both the dynamics and the bijections. This learning process is visualized for a two-link manipulator in Figure 1.

When the physical laws provide a reasonable model for the observed data, SIM2GP is naturally effective at control and model learning. However, in reality, there is likely a mismatch between the model and reality, so it is more useful to use SIM2GP as a prior for a more flexible function approximator, where we use the neural linear model (NLM) [18], [19]. We consider the residual approach [20], where SIM2GP is the mean function of an NLM (NLM-SIM2GP-RES),

$$\boldsymbol{f}_{\text{RES}}(\boldsymbol{x}, \boldsymbol{\theta}) = \boldsymbol{f}_{\text{SIM2GP}}(\boldsymbol{x}, \boldsymbol{\theta}_1) + \boldsymbol{f}_{\text{NLM}}(\boldsymbol{x}, \boldsymbol{\theta}_2), \tag{3}$$

and a function-space variational inference approach [19], [21], where SIM2GP is the prior $p(\boldsymbol{f})$ (NLM-SIM2GP-FS),

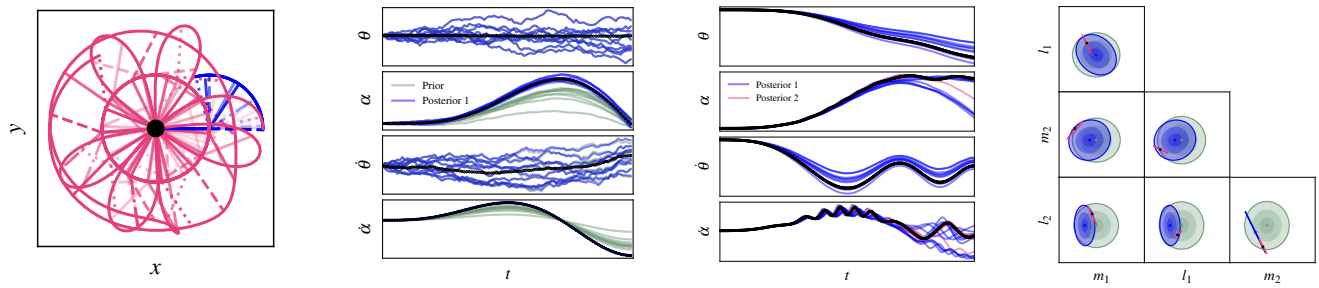
$$\max_q \mathbb{E}_{\boldsymbol{f}\sim q(\cdot)}[\log p(\mathcal{D} \mid \boldsymbol{f})] - \mathbb{D}_{\text{KL}}[q(\boldsymbol{f}) \,||\, p(\boldsymbol{f})]. \tag{4}$$

the function-space KL is intractable [21], so we replace it with an average KL between marginals under distribution ρ,

$$\hat{\mathbb{D}}_{\text{KL}}[q(\boldsymbol{f}) \,||\, p(\boldsymbol{f})] = \mathbb{E}_{\boldsymbol{x}\sim\rho(\cdot)}[\mathbb{D}_{\text{KL}}[q(\boldsymbol{y} \mid \boldsymbol{x}) \,||\, p(\boldsymbol{y} \mid \boldsymbol{x})]].$$

III. Experimental Results

We consider policy optimization using posterior sampling reinforcement learning (PSRL) [9], [22], [23] and active learning using 'posterior sampling active learning' (PSAL), with the information gain as the reward function [24], [25]. Figures 2 – 3 demonstrate performance on a simulated Furuta pendulum and cartpole, with and without simulator mismatch.



(a) A kinematic visualization of the datasets created for the two-link manipulator.

(b) Comparing the prior and first SIM2GP posterior on the first trajectory dataset.

(c) Comparing the first and second SIM2GP posterior on the second trajectory dataset.

(d) SIM2GP priors and posteriors, shown in the virtual parameterization space.

Fig. 1: Applying SIM2GP to identify a simulated, horizontal two-link manipulator across two datasets. The first dataset seeks to identify the second link, while the second dataset allows the whole manipulator to be learned. This is captured by the variance of the first and second posterior, as the first posterior remains uncertain about the first link, while the second posterior is converges about the true value.

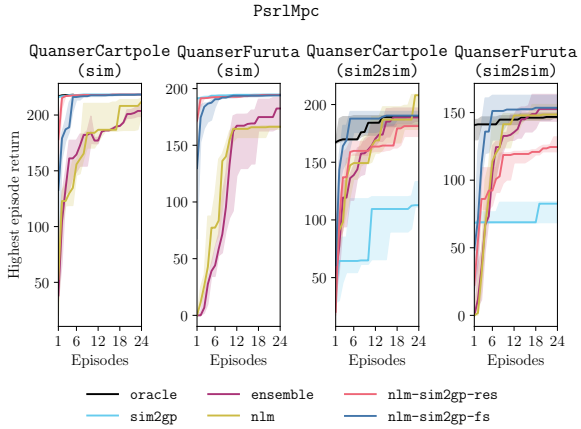


Fig. 2: Best-seen episodic return with posterior sampling reinforcement learning, using sample-based model predictive control (MPC) for optimal control. While the hybrid residual model performs best when the prior is well-specified, the function-space model performs best under mismatch. Uncertainty intervals are the quartiles over five seeds.

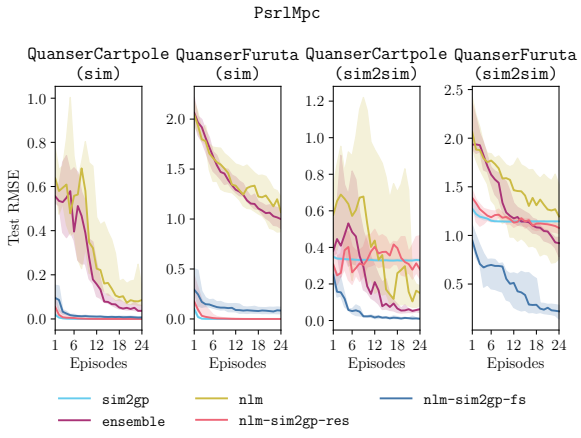


Fig. 3: Test root mean square error in the posterior sampling active learning setting. The test dataset is the replay buffer of the swing-up task, so low error indicates the agent has managed to swing up through exploration. Uncertainty intervals are the quartiles over five seeds.

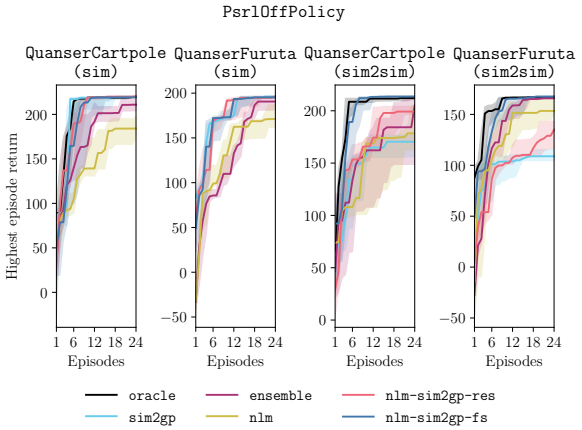


Fig. 4: Best-seen episodic return with posterior sampling reinforcement learning, using off-policy deep reinforcement learning. As with MPC, the function-space approach performs best under mismatch, and the off-policy oracle closely matches the performance with the oracle dynamics model. Uncertainty intervals are the quartiles over five seeds.

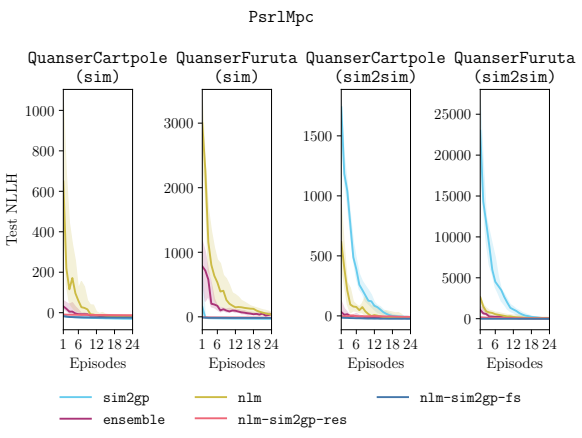


Fig. 5: Test negative log-likelihood in the posterior sampling active learning setting. Uncertainty intervals are the quartiles over five seeds.

REFERENCES

[1] M. P. Deisenroth, G. Neumann, J. Peters *et al.*, "A survey on policy search for robotics," *Foundations and Trends® in Robotics*, 2013.

[2] J. Schneider, "Exploiting model uncertainty estimates for safe dynamic control learning," *Advances in Neural Information Processing Systems (NeurIPS)*, 1996.

[3] J.-A. Ting, M. N. Mistry, J. Peters, S. Schaal, and J. Nakanishi, "A Bayesian approach to nonlinear parameter identification for rigid body dynamics." in *Robotics: Science and Systems (RSS)*, 2006.

[4] M. Deisenroth and C. E. Rasmussen, "PILCO: A model-based and data-efficient approach to policy search," in *International Conference on Machine Learning (ICML)*, 2011.

[5] K. Chua, R. Calandra, R. McAllister, and S. Levine, "Deep reinforcement learning in a handful of trials using probabilistic dynamics models," in *Advances in Neural Information Processing Systems (NeurIPS)*, 2018.

[6] F. Ramos, R. Possas, and D. Fox, "BayesSim: Adaptive domain randomization via probabilistic inference for robotics simulators," in *Robotics: Science and Systems (RSS)*, 2019.

[7] E. Heiden, C. E. Denniston, D. Millard, F. Ramos, and G. S. Sukhatme, "Probabilistic inference of simulation parameters via parallel differentiable simulation," in *IEEE International Conference on Robotics and Automation (ICRA)*, 2022.

[8] W. R. Thompson, "On the likelihood that one unknown probability exceeds another in view of the evidence of two samples," *Biometrika*, 1933.

[9] D. Russo and B. Van Roy, "Learning to optimize via posterior sampling," *Mathematics of Operations Research*, 2014.

[10] J. Baxter, "A model of inductive bias learning," *Journal of Artificial Intelligence Research*, 2000.

[11] D. J. MacKay, *Information theory, inference and learning algorithms*. Cambridge University Press, 2003.

[12] C. Rasmussen and C. Williams, *Gaussian Processes for Machine Learning*. MIT Press, 2006.

[13] R. Newbury, J. Collins, K. He, J. Pan, I. Posner, D. Howard, and A. Cosgun, "A review of differentiable simulators," *IEEE Access*, 2024.

[14] D. J. MacKay, "A practical Bayesian framework for backpropagation networks," *Neural Computation*, 1992.

[15] M. E. E. Khan, A. Immer, E. Abedi, and M. Korzepa, "Approximate inference turns deep networks into Gaussian processes," *Advances in Neural Information Processing Systems (NeurIPS)*, 2019.

[16] A. Immer, M. Korzepa, and M. Bauer, "Improving predictions of Bayesian neural nets via local linearization," in *Artificial Intelligence and Statistics (AISTATS)*, 2021.

[17] M. Lutter, J. Silberbauer, J. Watson, and J. Peters, "A differentiable Newton-Euler algorithm for real-world robotics," *arXiv preprint arXiv:2110.12422*, 2021.

[18] M. Lázaro-Gredilla and A. R. Figueiras-Vidal, "Marginalized neural network mixtures for large-scale regression," *IEEE Transactions on Neural Networks*, 2010.

[19] J. Watson, J. A. Lin, P. Klink, and J. Peters, "Neural linear models with functional gaussian process priors," in *Advances in Approximate Bayesian Inference (AABI)*, 2021.

[20] M. Saveriano, Y. Yin, P. Falco, and D. Lee, "Data-efficient control policy search using residual dynamics learning," in *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 2017.

[21] S. Sun, G. Zhang, J. Shi, and R. Grosse, "Functional variational Bayesian neural networks," in *International Conference for Learning Representations (ICLR)*, 2019.

[22] M. J. Strens, "A Bayesian framework for reinforcement learning," in *International Conference on Machine Learning (ICML)*, 2000.

[23] Y. Fan and Y. Ming, "Model-based reinforcement learning for continuous control with posterior sampling," in *International Conference of Machine Learning (ICML)*, 2021.

[24] M. Schultheis, B. Belousov, H. Abdulsamad, and J. Peters, "Receding horizon curiosity," in *Conference on Robot Learning (CoRL)*, 2020.

[25] T. Schneider, B. Belousov, G. Chalvatzaki, D. Romeres, D. K. Jha, and J. Peters, "Active exploration for robotic manipulation," in *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 2022.