

On the Design of LQR Kernels for Efficient Controller Learning

Alonso Marco¹, Philipp Hennig¹, Stefan Schaal^{1,2} and Sebastian Trimpe¹

Abstract—Finding optimal feedback controllers for nonlinear dynamic systems from data is hard. Recently, Bayesian optimization (BO) has been proposed as a powerful framework for direct controller tuning from experimental trials. For selecting the next query point and finding the global optimum, BO relies on a probabilistic description of the latent objective function, typically a Gaussian process (GP). As is shown herein, GPs with a common kernel choice can, however, lead to poor learning outcomes on standard quadratic control problems. For a first-order system, we construct two kernels that specifically leverage the structure of the well-known Linear Quadratic Regulator (LQR), yet retain the flexibility of Bayesian nonparametric learning. Simulations of uncertain linear and nonlinear systems demonstrate that the LQR kernels yield superior learning performance.

I. INTRODUCTION

A core problem of learning control is to determine optimal feedback controllers for (partially) unknown nonlinear systems from experimental data. Reinforcement learning (RL) [1], [2] is a promising framework for this, yet often requires performing many experiments on the physical system to even find suitable controllers, which limits the applicability of such techniques. Therefore, a lot of research effort has been invested into data efficiency of RL aiming at learning controllers from as few experiments as possible. Recently, Bayesian optimization (BO) has been proposed for RL as a promising approach in this direction. BO employs a probabilistic description of the latent objective function (typically a Gaussian process (GP)), which allows for selecting next control experiments in a principled manner, e.g., to maximize information gain [3] or perform safe exploration [4].

While BO provides a promising framework for learning controllers in fairly general settings, the full power of Bayesian learning is often not exploited. A key advantage of Bayesian methods is that they allow for combining prior problem knowledge with learning from data in a principled manner. In case of GP models, this concerns specifically the choice of the *kernel*, which captures the covariance between function values at different inputs and is thus the core component to model prior knowledge about the function shape. By choosing standard kernels, however, naive BO approaches do often not exploit this opportunity to improve learning performance.

¹Max Planck Institute for Intelligent Systems, 72076 Tübingen, Germany. {amarco, phennig, sschaal, strimpe}@tuebingen.mpg.de

²Computational Learning and Motor Control Lab, University of Southern California, Los Angeles, USA.

This work was supported in part by the Max Planck Society, the Max Planck ETH Center for Learning Systems, National Science Foundation grants IIS-1205249, IIS-1017134, EECS-0926052, the Office of Naval Research, and the Okawa Foundation.

In this paper, we show how structural knowledge about the optimal control problem at hand can be leveraged for designing specific kernels in order to improve data efficiency in learning control. For a first-order nonlinear quadratic optimal control problem, we propose two *LQR kernels* that leverage the structure of the famous Linear Quadratic Regulator (LQR) problem given in form of an *approximate linear model* of the *true nonlinear dynamics*. The proposed kernels leverage the structure of the LQR problem, while retaining the flexibility of nonparametric GP regression.

Contributions: In detail, this paper makes the following contributions:

- 1) We discuss how the structure of the well-known LQR problem can be leveraged for efficient learning of controllers for nonlinear systems.
- 2) This discussion leads to the proposal of two new kernels for GP regression in the context of learning control: the *parametric* and *nonparametric LQR kernels*.
- 3) The improved learning performance achieved with these kernels over a standard kernel is demonstrated through numerical simulations.

Related work: BO for learning controllers has recently been considered, for example in [3]–[7], in different flavors. While [3] and [5] focus on data efficiency by maximizing the information gain in each iteration, [4] and [6] propose methods for safe exploration. Different BO algorithms are compared in [7]. These works all consider tuning of state-feedback controllers with a quadratic cost (possibly saturated) similar to the setting herein, but using standard kernels.

The design of customized kernels for GP regression has been considered before in the context of control and robotics for related problems. A kernel for bipedal locomotion, which captures typical gait characteristics, is proposed in [8]. In [9], an impedance-based model is incorporated as prior knowledge for improved predictions in human-robot interaction. For the problem of maximizing power generation in photovoltaic plants, the authors in [10] incorporate explicit basis functions about known power curves in the kernel. In [5], a kernel is designed to model information from simulation and physical experiments, in order to leverage both sources of information for RL.

None of the above references considers the problem of incorporating the structure of the LQR problem for improving data-efficiency in learning control with BO.

Outline: We introduce the considered learning control problem in Sec. II, and summarize BO for RL in Sec. III, together with the necessary background on GPs. While the BO framework for learning control is introduced for general multivariate systems, we focus on the special case of a scalar

problem thereafter and develop the LQR kernels in Sec. IV. Numerical results in Sec. V illustrate the improved learning performance of the proposed kernels over a standard kernel. The paper concludes with remarks in Sec. VI.

Notation: A Gaussian random variable x with mean m and variance V is denoted by $x \sim \mathcal{N}(m, V)$. The expected value of x is denoted by $\mathbb{E}[x]$, while $\mathbb{V}[x_1, x_2]$ denotes the covariance of x_1 and x_2 . We also use $\mathbb{V}[x_1] := \mathbb{V}[x_1, x_1]$.

II. LEARNING CONTROL PROBLEM

We consider regulation of the nonlinear stochastic system

$$x_{t+1} = g(x_t, u_t, v_t) \quad (1)$$

with state $x_t \in \mathbb{R}^{n_x}$, control input $u_t \in \mathbb{R}^{n_u}$, and random noise $v_t \sim \mathcal{N}(0, V)$. We assume that the state x_t can be measured or otherwise estimated. The system dynamics $g(\cdot)$ are unknown. The control objective is to find a state-feedback controller

$$u_t = Fx_t \quad (2)$$

with controller gain $F \in \mathbb{R}^{n_u \times n_x}$ such that the quadratic cost

$$J = \lim_{T \rightarrow \infty} \frac{1}{T} \mathbb{E} \left[\sum_{t=0}^{T-1} x_t^T Q x_t + u_t^T R u_t \right] \quad (3)$$

with symmetric weights $Q \geq 0$ and $R > 0$ is minimized. A quadratic cost is a very common choice in optimal control [11] expressing the designer's preference in the fundamental trade-off between control performance and control effort.

While the solution of the above problem is standard when the dynamics (1) are known and linear [11], here, we face the problem of optimal control under *unknown* and *nonlinear* dynamics.

To address this problem, we follow a direct RL approach [12], where the optimal controller is learned by directly evaluating the performance of candidate controllers F_i on the real system (1) without learning an (intermediate) dynamics model. We employ Bayesian optimization (BO) as a popular approach for sequential global optimization, which leverages the information from previous trials in order to propose the next candidate controller F_{i+1} so as to eventually find the optimum of (3). Because evaluations on the system are expensive, we seek to find the optimal controller with as few evaluations as possible. To this end, we address herein how to leverage information from a linear model that approximates the true system (1).

III. BAYESIAN OPTIMIZATION FOR LEARNING CONTROL

We briefly introduce necessary background material on Gaussian processes (GPs) and Bayesian optimization (BO), before describing reinforcement learning with BO.

A. Gaussian process regression

Let θ denote the free parameters of a feedback controller; that is, $\theta = \text{vec}(F)$ in (2), or any other parameterization. The dependence of the cost function J (3) on θ is unknown *a priori* because of the lack of knowledge of the system (1). We

use GP regression to approximate the function $J : \Theta \rightarrow \mathbb{R}$, $\theta \in \Theta$, from data (i.e., noisy function evaluations)

$$J_i = J(\theta_i) + \varepsilon_i, \quad \varepsilon_i \sim \mathcal{N}(0, \sigma^2). \quad (4)$$

Obtaining one data point J_i corresponds to performing a closed-loop experiment on the true system (1) with controller θ_i , recording the state-input trajectory, and computing the cost from these data (in case of (3) a finite approximation is used with a sufficiently long horizon T).

A GP can be defined as a probability distribution over the space of functions $J : \Theta \rightarrow \mathbb{R}$ whose restriction to any finite number of function values is jointly Gaussian [13, p. 13]. A GP is specified through its prior mean function $\mu : \Theta \rightarrow \mathbb{R}$ and covariance function $k : \Theta \times \Theta \rightarrow \mathbb{R}$. We write

$$\begin{aligned} J(\theta) &\sim \mathcal{GP}(\mu(\theta), k(\theta, \theta')) \quad \text{with} \\ \mu(\theta) &= \mathbb{E}[J(\theta)] \\ k(\theta, \theta') &= \mathbb{V}[J(\theta), J(\theta')]. \end{aligned}$$

Hence, μ is the expected function value, and k captures the covariance between any two function values and is used to model uncertainty. The latter is also called *kernel* and must satisfy the following property to give rise to a valid covariance function:

Definition 1 (according to [13]): Let $k : \Theta \times \Theta \rightarrow \mathbb{R}$ be a function, and $K_N \in \mathbb{R}^{N \times N}$ be the symmetric matrix whose entries are computed as $K_{ij} = k(\theta_i, \theta_j)$ from the collection $\{\theta_1, \theta_2, \dots, \theta_N\}$, $\theta_i \in \Theta$. The function k is called a *positive semidefinite kernel* if K is positive semidefinite for any finite collection $\{\theta_1, \theta_2, \dots, \theta_N\}$.

The matrix K_N is called the *Gram matrix*.

A kernel typically has its own parameters, called *hyperparameters*, such as length scales or output variance. By choosing the prior mean and the kernel, the user can specify prior knowledge about the function such as expected shape, length scales, and smoothness properties.

In GP regression, learning a function amounts to predicting the (normal) distribution of function values $J(\theta^*)$ at arbitrary inputs $\theta^* \in \Theta$ based on previous evaluations. Given N data points $\mathcal{D}_N = \{\theta_i, J_i\}_{i=1}^N$, the posterior mean and variance of $J(\theta^*)$ can be stated in closed form as

$$\begin{aligned} \mathbb{E}[J(\theta^*) | \mathcal{D}_N] &= \mu(\theta^*) - k_N^T(\theta^*)(K_N + \sigma^2 I)^{-1} \hat{J}_N \quad (5) \\ \mathbb{V}[J(\theta^*) | \mathcal{D}_N] &= k(\theta^*, \theta^*) \\ &\quad - k_N^T(\theta^*)(K_N + \sigma^2 I)^{-1} k_N(\theta^*) \quad (6) \end{aligned}$$

where $k_N(\theta^*) := [k(\theta^*, \theta_1), \dots, k(\theta^*, \theta_N)]^T$ and $\hat{J}_N := [J_1 - \mu(\theta_1), \dots, J_N - \mu(\theta_N)]^T$.

In addition to computing the posterior distribution, the data \mathcal{D}_N can also be used to optimize hyperparameters in order to improve the model fit. A popular approach is maximizing the marginal likelihood of the data, $p(J_N | \theta_1, \dots, \theta_N)$, which is given in logarithmic form as [13, p. 19]

$$-\frac{1}{2} \hat{J}_N^T (K_N + \sigma^2 I)^{-1} \hat{J}_N - \frac{1}{2} |K_N + \sigma^2 I| - \frac{N}{2} \log 2\pi. \quad (7)$$

Algorithm 1 Learning control with Bayesian optimization

- 1: Specify objective function J (e.g., (3) with finite horizon T)
 - 2: Specify GP prior (mean μ , kernel k)
 - 3: Initialize: controller θ_0 ; data set $\mathcal{D}_0 = \emptyset$
 - 4: **while** (not terminated) **do** ▷ e.g., fixed number of experiments
 - 5: perform closed-loop experiment with controller θ_i
 - 6: compute J_i from experimental data $\{x_t, u_t\}_{t=0}^{T-1}$
 - 7: add evaluation $\{\theta_i, J_i\}$ to data set \mathcal{D}_i
 - 8: update GP posterior (5), (6)
 - 9: [optional:] optimize hyperparameters (e.g., maximize (7))
 - 10: compute next controller θ_{i+1} via (8)
 - 11: **end while**
 - 12: determine ‘best guess’ θ_{bg} for the controller parameters, e.g., minimum of posterior mean (5)
 - 13: **return** θ_{bg}
-

B. Bayesian optimization

Bayesian optimization [14] denotes a class of global optimization methods, where a probabilistic description $p(J(\theta))$ of the latent objective function J is used for data-efficient optimization. Here, the probabilistic description is given by the GP $J \sim \mathcal{GP}(\mu, k)$. Given $p(J(\theta))$, a *utility* $U[p(J(\theta))]$ is defined, which is maximized in each iteration to find the next evaluation point

$$\theta_{i+1} = \arg \max_{\theta} U[p(J(\theta))]. \quad (8)$$

Different BO algorithms primarily distinguish in how the selection of the next evaluation (8) is formulated. Popular BO algorithms include *expected improvement* (EI) [15], *probability of improvement* (PI) [16], *GP upper confidence bound* (GP-UCB) [17], and *entropy search* (ES) [18].

The method developed herein is agnostic to the type of BO algorithm used. For the examples in Sec. V, we employ EI, which uses

$$U[p(J(\theta))] = \mathbb{E}[\min\{0, J_{\text{low}} - J(\theta)\}] \quad (9)$$

for the optimization in (8), where J_{low} represents the lowest function from current data set \mathcal{D}_N . EI thus selects the next query point where the expected improvement upon J_{low} (under the GP of J) is maximal.

C. Reinforcement learning with Bayesian optimization

Algorithm 1 summarizes direct RL using Bayesian optimization. This framework has been successfully applied in experimental settings to learn feedback controllers for the control problem of Sec. II and related scenarios. Berkenkamp *et al.* [4] optimize a state-feedback controller (2) for quadrotor trajectory tracking using a safe BO algorithm [19], which builds on GP-UCB. Marco *et al.* [3] propose to parametrize the feedback gain F as LQR policy [20], whose optimal weights are learned using ES, which maximizes the information gain from each experiment. This algorithm has been extended in [5] to include information from different sources (such as simulation and physical experiments). Both methods [3], [5] were successfully applied to learn pole balancing controllers. Calandra *et al.* [7] compare PI, EI, and GP-UCB for learning the parameters of a discrete event controller for a walking robot.

IV. CONSTRUCTING LQR KERNELS

In this section, we present two kernels that are specifically designed for the control problem of Sec. II and incorporate approximate model knowledge in form of a linear approximation to (1). Because for a linear system, the solution to the optimal control problem of Sec. II is the well-known LQR, we term these kernels *LQR kernels*. The following derivations are presented for a first-order system (1), where all variables are scalars ($n_x = n_u = 1$). Small letters are used in place of the capital ones to emphasize scalar variables (e.g., v instead of V in (1) and f instead of F in (2)). We consider minimization of the cost (3), which is rewritten as

$$J = \lim_{T \rightarrow \infty} \frac{1}{T} \mathbb{E} \left[\sum_{t=0}^{T-1} qx_t^2 + ru_t^2 \right]. \quad (10)$$

We start by considering a learning example with a standard kernel choice for the GP, which motivates why a specifically designed kernel can be desirable.

A. Problems with standard kernel

The most common kernel choice in GP regression is arguably the squared exponential (SE) kernel [13]

$$k_{\text{SE}}(\theta, \theta') = \sigma_{\text{SE}}^2 \exp\left(-\frac{(\theta - \theta')^2}{2\ell^2}\right) \quad (11)$$

with signal variance σ_{SE}^2 and length-scale ℓ as its hyperparameters. Let us consider the problem of learning the cost function J (10) via GP regression with SE kernel for the following linear example:

Example 1: Let (1) be given by $x_{t+1} = 0.9x_t + u_t + v_t$ with $v_t \sim \mathcal{N}(0, 1)$.

Figure 1 shows the prior GP and the posterior GP after obtaining four data points (i.e., after four evaluations of controllers f_i). A few issues are apparent from the posterior: (i) the kernel has problems with the different length scales of the function (J is steep around $f = -0.1$, but rather flat in the center region); (ii) the GP does not generalize well to regions where no data has been seen (e.g., around $f = -1.6$, where the posterior mean resorts to the prior); and (iii) the overall fit is not very good.

Clearly, the fit will improve with more data, but, for efficient and fast learning of controllers, we are particularly interested in good fits from just few data points. Hence, we seek to improve the fitting properties of the GP by exploiting knowledge about the system (1) in terms of an approximate linear model.

B. Incorporating prior knowledge

In no practical situation, one has a perfect model of the system to be controlled. At the same time, it is often possible to obtain a rough model, e.g., from first principles modeling or some system identification procedure. Here, we assume that an uncertain linear model

$$x_{t+1} = ax_t + bu_t + v_t \quad (12)$$

$$a \in [a_{\min}, a_{\max}], b \in [b_{\min}, b_{\max}] \quad (13)$$

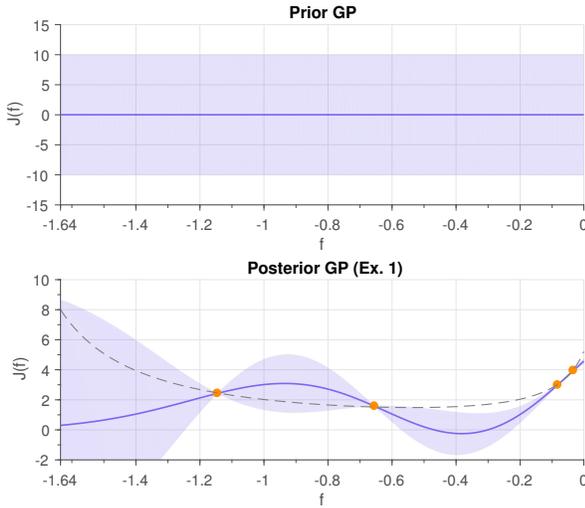


Fig. 1. Prior and posterior GPs of Example 1 using the squared exponential kernel k_{SE} (hyperparameters: $\sigma_{SE}^2 = 25$, $\ell = 0.4$). The thick line represents the GP mean, and the light blue area the GP variance (\pm two standard deviations). The true function is shown in the bottom plot in dashed gray, and data points in orange.

is available as an approximation to (1), e.g., from linearization of a first principles model with (possibly) some uncertainty about the physical parameters.

In the following, we will consider controller gains f such that the system (12) is guaranteed stable for all parameters (13). That is, we consider $f \in \mathcal{F}$ with

$$\mathcal{F} := \{f \in \mathbb{R} \mid |a + bf| < 1 \\ \forall a \in [a_{\min}, a_{\max}], b \in [b_{\min}, b_{\max}]\}. \quad (14)$$

This restriction makes sense, for example, in safety critical applications, where one wants to avoid the risk of trying an unstable controller based on the system knowledge available (i.e., (12), (13)). Moreover, the restriction to \mathcal{F} will ensure that subsequent calculations are well-defined.

If a and b were known, the functional dependence of the cost J on the controller gain f in (2) for the linear system (12) could be derived using standard control theory.

Fact 1: Consider the system (12) with known parameters a and b , and let $|a + bf| < 1$. Then, the cost (10) is given by¹

$$J = v \frac{q + rf^2}{1 - (a + bf)^2} =: \phi_{(a,b)}(f). \quad (15)$$

Proof: The controlled process $x_{t+1} = (a + bf)x_t + v_t$ is stable by assumption and thus converges to a stationary process with zero mean and variance $\mathbb{E}[x_t^2] = P$, where P is the unique positive solution to $P - P(a + bf)^2 = v$, [21, Theorem 3.1]. For stationary x_t , (10) resolves into

$$J = \mathbb{E}[qx_t^2 + ru_t^2] = (q + rf^2) \mathbb{E}[x_t^2] = (q + rf^2)P.$$

Equation (15) then follows. \blacksquare

¹In the notation $\phi_{(a,b)}(f)$, we omit the parametric dependence on v , q and r since v is a multiplicative constant, which does not play a role in the later optimization, and we assume that q and r are fixed.

If we are uncertain about a and b , a collection of possible costs (15) emerge from all the possible combinations of a and b within their ranges (13). We assume this collection of costs to be explained by a Gaussian process J_{LQR} . While any arbitrary choice of a and b in (12) is only an approximation to (1), it yields a cost (15) that contains useful structural information, which can be leveraged for faster learning controllers from data. In the next sections, we show how this prior knowledge can be exploited to construct LQR kernels.

C. Parametric LQR kernel

A reasonable choice for $J_{LQR}(f)$ is

$$J_{LQR}(f) = w \phi_{(\bar{a}, \bar{b})}(f), \quad w \sim \mathcal{N}(\bar{w}, \sigma_w^2) \quad (16)$$

where $\bar{a} := (a_{\min} + a_{\max})/2$ and $\bar{b} := (b_{\min} + b_{\max})/2$ are the midpoints of the uncertainty intervals (13).

Equation (16) is a standard parametric model with a single feature $\phi_{(\bar{a}, \bar{b})}(f)$ and Gaussian prior. It is well-known (see e.g. [13]) that $J_{LQR}(f)$ is a GP,

$$J_{LQR}(f) \sim \mathcal{GP}(0, k_{pLQR}(f, f')) \quad (17)$$

where we have assumed $\bar{w} = 0$, with kernel

$$k_{pLQR}(f, f') := \sigma_w^2 \phi_{(\bar{a}, \bar{b})}(f) \phi_{(\bar{a}, \bar{b})}(f') \\ = \sigma_p^2 \frac{v^2 (q + rf^2)(q + rf'^2)}{(1 - (\bar{a} + \bar{b}f)^2)(1 - (\bar{a} + \bar{b}f')^2)} \quad (18)$$

and hyperparameters $\sigma_p := \sigma_w$, \bar{a} , and \bar{b} . We refer to (18) as the *parametric LQR kernel* because it captures the cost function for the linear system (\bar{a}, \bar{b}) with quadratic cost, and thus the structure of the LQR problem.

To illustrate the performance of the parametric LQR kernel k_{pLQR} , we revisit Example 1 using this kernel instead of the SE kernel. The top two graphs of Fig. 2 show the prior and posterior GP for the same data points as in Fig. 1 when using k_{pLQR} with hyperparameters $\bar{a} = 0.9$ and $\bar{b} = 1$. We see that the posterior fit from only four data points is almost perfect and much better than the one in Fig. 1. This is, of course, not a big surprise because the hyperparameters match the true underlying system of Example 1 perfectly. The fitting performance deteriorates, however, if the hyperparameters are off. This can be seen in the bottom graph of Fig. 2, which shows the posterior GP with the same hyperparameters, but for the system

Example 2: $x_{t+1} = 0.8x_t + 0.9u_t + v_t$ with $v_t \sim \mathcal{N}(0, 1)$.

To improve the fit in this situation, one can employ hyperparameter optimization (see Sec. III-A) to find improved parameters \bar{a} and \bar{b} that better explain the data. The simulation results in Sec. V will show that this can be a viable approach. An alternative is to design more flexible and expressive kernels, which allow for fitting more general models. This we discuss next.

D. Nonparametric LQR kernel

The kernel (18) captures the structure of the cost function for the LQR problem with one specific linear model (\bar{a}, \bar{b}) . A straightforward way to increase the flexibility of the kernel

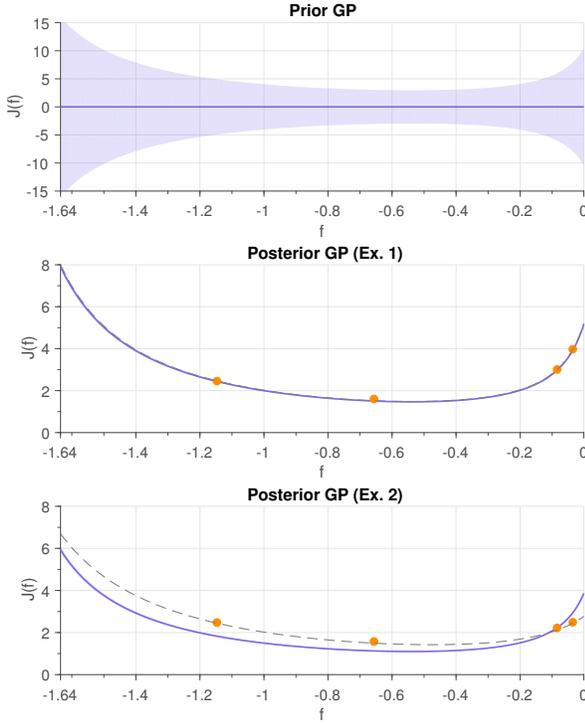


Fig. 2. GP fit using the parametric LQR kernel k_{pLQR} (hyperparameters: $\sigma_p^2 = 1$, $\bar{a} = 0.9$, $\bar{b} = 1$). The color code is the same as in Fig. 1. The hyperparameters are exact for Example 1, while they are off by about 10% for Example 2.

in order to fit more general problems is to use m basis functions (or features) $\phi_{(a_i, b_i)}$ corresponding to m models $(a_1, b_1), (a_2, b_2), \dots, (a_m, b_m)$,

$$\begin{aligned} J_{\text{LQR}}(f) &= \underbrace{[\phi_{(a_1, b_1)}(f) \phi_{(a_2, b_2)}(f) \cdots \phi_{(a_m, b_m)}(f)]}_{{=: \Phi^T(f)}} w \\ &= \Phi^T(f) w \end{aligned} \quad (19)$$

with $w \in \mathbb{R}^m$, $w \sim \mathcal{N}(\bar{w}, \Sigma_w)$. The derivation of the corresponding kernel is analogous to (18) (see [13, Sec. 2.7]) and yields

$$k_{\text{pLQR}, m}(f, f') = \Phi^T(f) \Sigma_w \Phi(f'). \quad (20)$$

Same as (16), the model (19) represents a parametric model for the LQR cost J_{LQR} . That is, its flexibility is essentially limited to the number of explicit features $\phi_{(a_i, b_i)}$. Employing powerful kernel techniques [22], the parametric model can be turned into a nonparametric one, which includes an infinite number of features while retaining finite computational complexity. The key idea is to consider the kernel (20) in the limit of infinitely many features corresponding to models $a \in [a_{\min}, a_{\max}]$ and $b \in [b_{\min}, b_{\max}]$. The derivation follows ideas similar to how the standard SE kernel can be derived from basic features, [13, p. 84].

Consider the partitions of $[a_{\min}, a_{\max}]$ and $[b_{\min}, b_{\max}]$ into m equidistant intervals, and let $\{a_i\}_{1:m}$ and $\{b_i\}_{1:m}$ be the lower (or upper) interval limits. Consider the model (19) with feature vector

$$\Phi^T(f) = [\phi_{(a_1, b_1)}(f) \cdots \phi_{(a_i, b_j)}(f) \cdots \phi_{(a_m, b_m)}(f)]$$

which includes all combinations $\phi_{(a_i, b_j)}$ for $i, j \in \{1, \dots, m\}$, and the parametric prior $\bar{w} = 0$ and $\Sigma_w = \frac{\sigma_n^2 (a_{\max} - a_{\min})(b_{\max} - b_{\min})}{m^2} I$ for some $\sigma_n \in \mathbb{R}$. The kernel (20) then becomes

$$\begin{aligned} k_{\text{pLQR}, m}(f, f') &= \frac{\sigma_n^2 (a_{\max} - a_{\min})(b_{\max} - b_{\min})}{m^2} \\ &\quad \times \sum_{j=1}^m \sum_{i=1}^m \phi_{(a_i, b_j)}(f) \phi_{(a_i, b_j)}(f'). \end{aligned} \quad (21)$$

Since $\phi_{(a_i, b_j)}$ is continuous on \mathcal{F} , the finite sum (21) converges to the Riemann integral in the limits as $m \rightarrow \infty$. We can thus define the *nonparametric LQR kernel*

$$\begin{aligned} k_{\text{nLQR}}(f, f') &= \lim_{m \rightarrow \infty} k_{\text{pLQR}, m}(f, f') \\ &= \sigma_n^2 \int_{b_{\min}}^{b_{\max}} \int_{a_{\min}}^{a_{\max}} \phi_{(a, b)}(f) \phi_{(a, b)}(f') da db \end{aligned} \quad (22)$$

for $f, f' \in \mathcal{F}$ with the signal variance σ_n^2 and the integration boundaries a_{\min} , a_{\max} , b_{\min} , and b_{\max} as hyperparameters. While the kernel in (22) represents the structure of the cost function (15) for an infinite number of models (all $a \in [a_{\min}, a_{\max}]$ and $b \in [b_{\min}, b_{\max}]$), its computation is finite consisting of solving the integral in (22). By contrast, the computational complexity of the parametric kernels (20) and (21) grows with the number of features m . We prove that the kernel (22) is indeed a valid covariance function.

Proposition 1: $k_{\text{nLQR}}(f, f')$ is positive semidefinite for all $f, f' \in \mathcal{F}$.

Proof: Take any collection $\{\theta_1, \theta_2, \dots, \theta_N\}$ and any $z \in \mathbb{R}^N$. Let K_N be the Gram matrix for the kernel k_{nLQR} . Then

$$\begin{aligned} z^T K_N z &= \sum_{j=1}^N z_j \left(\sum_{i=1}^N z_i k_{\text{nLQR}}(\theta_i, \theta_j) \right) \\ &= \sigma_n^2 \sum_{j=1}^N z_j \sum_{i=1}^N z_i \int_{b_{\min}}^{b_{\max}} \int_{a_{\min}}^{a_{\max}} \phi_{(a, b)}(\theta_i) \phi_{(a, b)}(\theta_j) da db \\ &= \sigma_n^2 \int_{b_{\min}}^{b_{\max}} \int_{a_{\min}}^{a_{\max}} \sum_{j=1}^N z_j \phi_{(a, b)}(\theta_j) \sum_{i=1}^N z_i \phi_{(a, b)}(\theta_i) da db \\ &= \sigma_n^2 \int_{b_{\min}}^{b_{\max}} \int_{a_{\min}}^{a_{\max}} \left(\sum_{i=1}^N z_i \phi_{(a, b)}(\theta_i) \right)^2 da db \geq 0 \end{aligned}$$

which completes the proof. \blacksquare

The above derivation corresponds to the *kernel trick* [13], [22], which is a core idea of kernel methods and behind many powerful learning algorithms. In essence, the kernel trick means to write a learning algorithm solely in terms of inner products of features and replacing those by a kernel². In particular, this allows for considering an infinite number of features, while retaining finite computation.

Figure 3 shows the prior and posterior GP for the nonparametric LQR kernel (22). Because the kernel is more flexible

²All computations for the GP regression, i.e., equations (5), (6), and (7), are expressed solely in terms of the kernel k (and the mean μ).

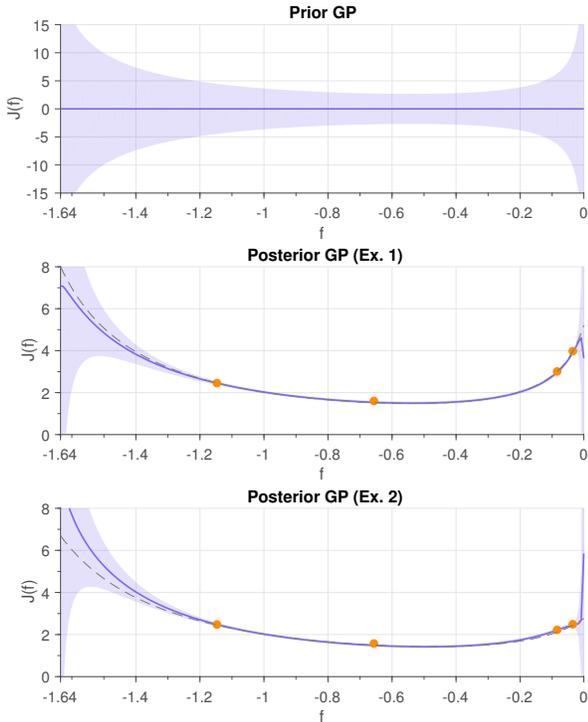


Fig. 3. GP fit using the nonparametric LQR kernel k_{nLQR} (hyperparameters: $\sigma_n^2 = 20$, $a_{\min} = 0.8$, $a_{\max} = 1.0$, $b_{\min} = 0.9$, and $b_{\max} = 1.1$). Colors are the same as in Fig. 1. Both examples are fitted well.

than (18), it can fit the cost functions for the two different models of Example 1 and Example 2 well.

E. A combined kernel

System (12) is an approximation to the true system (1). It is thus not desirable to fully commit to the model for solving the optimal control problem. This would mean to directly minimize (15) (which would result in the well-known LQR solution). On the other hand, the linear problem contains information about the structure of the optimization problem, which shall be useful also for optimization of the true nonlinear system (1), as long as we believe (12) to be a reasonable approximation thereof. In other words, we can expect the true cost function J for the nonlinear problem to bear some similarity to (15).

We model the cost function (3) for the nonlinear system (1) as being composed of a part that stems from the approximation as LQR problem and an error term,

$$J(f) = J_{\text{LQR}}(f) + J_{\Delta}(f). \quad (23)$$

The term $J_{\Delta}(f)$ captures the error in the model that stems from the fact that the true problem is nonlinear. We model it as a standard GP, e.g., employing the SE kernel (11): $J_{\Delta}(f) \sim \mathcal{GP}(0, k_{\text{SE}}(f, f'))$. We can model $J_{\text{LQR}}(f)$ as a GP (17) using either the parametric LQR kernel (18) or the nonparametric (22) LQR kernel. Then, since the sum of two independent Gaussians is also Gaussian, it follows from (23) that also J is a GP (see [13, Sec. 2.7]), with

$$J(f) \sim \mathcal{GP}(0, k_{\text{LQR}}(f, f') + k_{\text{SE}}(f, f')).$$

where k_{LQR} can be replaced by (18) or (22). By choosing the hyperparameters of the kernels, the designer can express how much he or she trusts the LQR versus the SE model. For example, $\sigma_{\text{SE}} = 0$ means to fully rely on the LQR kernel.

V. SIMULATIONS

In this section, we show statistical comparisons of the LQR kernels proposed in Sec. IV against the commonly used SE kernel, in two different settings. In the first setting, we evaluate the performance of each kernel in the context of GP regression. Specifically, we quantify the mismatch between the GP posterior mean, computed from a set of random evaluations, and the underlying cost function. In the second setting, we evaluate each kernel in the context of BO by comparing the learned minimum to the true global minimum.

The GP regression and BO experiments are presented in Sec. V-B and Sec. V-C, respectively, considering a linear system (1). In addition, we also evaluate the BO setting for a nonlinear system in Sec. V-D.

A. Experimental choices

For the simulations in Sec. V-B and Sec. V-C, we consider the true system (1) to be linear (i.e., (12)) with uncertain parameters $a \in [0.8, 1.0]$ and $b \in [0.9, 1.1]$. We consider the optimal control problem as in Sec. II with $q = r = v = 1$. Feedback controllers (2) are considered to be in the range (14), $\mathcal{F} = [-1.64, -0.001]$.

For each controller f , the corresponding infinite-horizon LQR cost $J(f)$ is given by (15). In practice, only finite-horizon simulations can be realized. Therefore, the outcome of an experiment is noisy, as modeled in (4), with $\sigma = 0.05$.

In each simulation, a different linear model (a, b) is obtained by uniformly sampling the ranges above, which yields a different underlying cost function (15). Each simulation is repeated for four different kernels:

- **SE kernel:** As described in (11).
- **LQR kernel I:** Parametric LQR kernel (18) with fixed parameters (\bar{a}, \bar{b}) (midpoints of uncertainty intervals).
- **LQR kernel II:** Parametric LQR kernel (18), with (\bar{a}, \bar{b}) optimized from evaluations by maximizing (7).
- **LQR kernel III:** Nonparametric LQR kernel (22).

The parametric LQR kernel (18) is constructed taking the middle points of the uncertainty intervals, i.e., $\bar{a} = 0.9$ and $\bar{b} = 1$. For the nonparametric LQR kernel (22), the intervals serve as integration domains. The length scale of the SE kernel is computed as one fifth of the input domain, i.e., $\ell = 0.327$. The signal variance of the parametric and nonparametric LQR kernel, i.e., σ_p^2 and σ_n^2 , are normalized such that $k_{\text{pLQR}}(\bar{f}, \bar{f}) = k_{\text{nLQR}}(\bar{f}, \bar{f}) = 10$, where $\bar{f} = -0.82$ is the midpoint of \mathcal{F} . Since the variance of these two kernels grow fast toward the corners of the domain, we set for the SE kernel $\sigma_{\text{SE}}^2 = 50$, for a fair comparison.

B. GP regression setting

For this statistical comparison, we run 1000 simulations. In each simulation, we compute the GP posterior conditioned

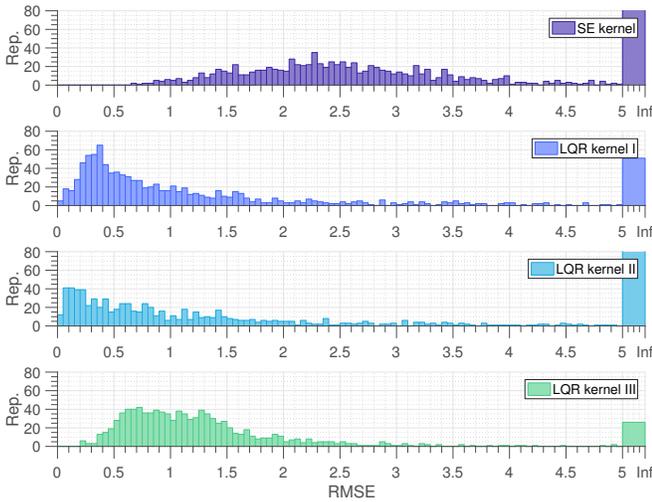


Fig. 4. Histograms of the root mean squared error (RMSE) between the true cost function and the GP posterior mean conditioned on two evaluations. The histograms are computed out of 1000 simulations.

on two evaluations randomly sampled from the underlying cost function. We assess the quality of the regression by computing the root mean squared error (RMSE) between the true cost function and the GP posterior mean, both computed on a grid of 100 points over \mathcal{F} .

Fig. 4 shows the histograms of the RMSE obtained with the different kernels. The LQR kernels clearly outperform the SE kernel in these experiments because they contain structural knowledge about the true cost, which contributes to a better GP fit, even with only two data points. The nonparametric kernel makes good predictions because it inherently contains information about all possible functions within their uncertainty ranges of (a, b) . However, poorly specified integration bounds will decrease its performance. The RMSE statistical analysis confirms this when the integration intervals on (a, b) are a 50% larger than their uncertainties. The parametric kernel with fixed hyperparameters (\bar{a}, \bar{b}) has a significant number of outliers since, in many cases, the data is queried from a cost function whose sampled (a, b) are far away from the ones of the kernel. The parametric kernel with hyperparameter optimization also leads to a better fit than SE, but has most outliers (about 75% of the simulations) since hyperparameter optimization is not reliable with just two data points.

We have repeated these experiments with $N = 1, 5,$ and 10 evaluations. Table I shows the averaged RMSE over 1000 simulations for each kernel, and the corresponding standard

TABLE I
RMSE AVERAGED OVER 1000 SIMULATIONS.

| N | SE ker. | SE ker. (*) | LQR ker. I | LQR ker. II | LQR ker. III |
|-----|-------------|-------------|--------------------|--------------------|--------------|
| 1 | 2.76 (0.74) | 2.39 (0.81) | 1.01 (0.95) | 1.98 (1.29) | 1.31 (0.71) |
| 2 | 2.49 (0.85) | 2.42 (0.93) | 1.02 (0.96) | 1.09 (1.07) | 1.22 (0.67) |
| 5 | 1.83 (1.02) | 1.31 (0.95) | 1.10 (1.05) | 0.45 (0.70) | 1.20 (0.76) |
| 10 | 1.13 (0.99) | 0.98 (0.86) | 0.98 (0.96) | 0.20 (0.46) | 1.11 (0.74) |

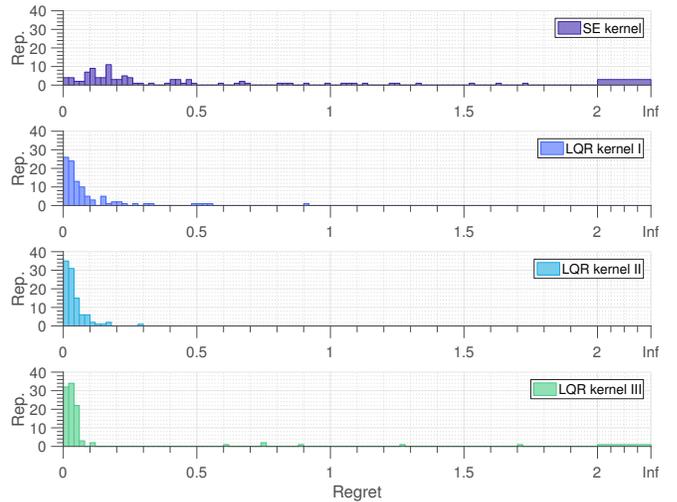


Fig. 5. Histogram of the regret incurred by stopping BO after three evaluations. The system considered is the linear system (12). The histogram is computed over 100 BO runs.

deviation (in parentheses). The outliers (i.e., any RMSE above 5) were excluded from these computations. In general, we see that the LQR kernel optimized from data performs better than the others for more than 2 evaluations.

For a fair comparison, we also include the SE kernel with hyperparameter optimization in the table (marked with an asterisk). Because it performs similar to the SE kernel, and it does not improve upon the LQR kernels, we leave it out of the discussion for the rest of the paper.

Remark: The hyperparameters (a, b) of the parametric LQR kernel are optimized from data by maximizing the marginal likelihood (7). In a sense, optimizing the hyperparameters (a, b) of the LQR kernel from data can be considered similar to doing system identification on the linear system (a, b) using (7) as performance metric.

C. BO setting

In this section, we evaluate the performance of each kernel in the context of BO. For each BO run, the first evaluation is decided randomly within the range of controllers \mathcal{F} . Subsequent evaluations are acquired using the expected improvement (EI) method (8), (9). We stop the exploration after three evaluations and compute the instantaneous regret (i.e., the absolute error between the true minimum and the minimum of the GP posterior mean) as the outcome of each experiment.

Fig. 5 shows the histogram of the regret for each kernel over 100 BO runs. The LQR kernels consistently outperform the SE kernel. The nonparametric kernel shows some outliers because in some cases the GP posterior mean grows large toward negative values, and so does its minimum. This is a wrong prediction of the underlying cost, defined positive. However, this minor issue can easily be detected, and the optimization procedure continued with a randomly sampled controller.

TABLE II

REGRET AVERAGED OVER 100 BO RUNS FOR A NONLINEAR SYSTEM

| N | SE kernel | LQR kernel I | LQR kernel II | LQR kernel III |
|-----|-------------|--------------------|--------------------|--------------------|
| 2 | 1.34 (0.33) | 0.30 (0.21) | 0.32 (0.20) | 0.35 (0.18) |
| 3 | 0.49 (0.39) | 0.33 (0.38) | 0.31 (0.19) | 0.32 (0.18) |
| 4 | 0.35 (0.27) | 0.36 (0.44) | 0.32 (0.18) | 0.32 (0.17) |
| 5 | 0.36 (0.21) | 0.31 (0.36) | 0.32 (0.20) | 0.32 (0.18) |

D. BO setting for a nonlinear system

In this section, we use the same BO setting as in Sec. V-C, but consider now a nonlinear system (1), namely

$$x_{t+1} = \tilde{a} \sin(x_t) + \tilde{b}u_t + v_t \quad (24)$$

with $v_t \sim \mathcal{N}(0, 1)$ and uncertain parameters $\tilde{a} \in [0.9, 1.1]$, $\tilde{b} \in [0.9, 1.1]$. We control this system from $x_0 = 1$ to zero, using the same controller structure as the one described in Sec. V-A. In this case, the considered range of controllers is $\mathcal{F} = [-1.57, -0.27]$, which corresponds to (14), reduced by a 20%. The LQR kernels are built up using the linearized version of (24) around the zero equilibrium point, i.e., $x_{t+1} = \tilde{a}x_t + \tilde{b}u_t + v_t$. Table II shows the regret average and standard deviation. As can be seen, the LQR kernels perform better than the SE kernel.

VI. CONCLUDING REMARKS

In this paper, we discussed how prior knowledge about the structure of an optimal control problem can be leveraged for data-efficient learning control. Specifically, for a nonlinear quadratic optimal control problem, we showed how an uncertain linear model approximating the true nonlinear dynamics can be exploited in a Bayesian setting. This led to the proposal of two novel kernels, a parametric and a nonparametric version of an LQR kernel, which incorporate the structure of the well-known LQR problem as prior knowledge. Numerical simulations herein demonstrate improved data efficiency over standard kernels, i.e., good controllers are learned from fewer experiments. We hope that the discussion and analysis herein also motivate further development of kernels tailored for other learning control problems.

Approaching the nonlinear quadratic optimal control problem presented herein with pure model-based methods can lead to superior performance for very accurate models compared to the proposed data-based approach. However, this paper shares the motivation of [3], which proposes a data-based approach when only poor models are available, and extends it by incorporating the LQR structure into the kernel.

The results herein are preliminary in the sense that the LQR kernels are developed for a first-order system. While this is the natural first step, future work will concern the extension of the ideas and derivations to multivariate systems in order to develop this into a powerful framework for learning control in practice. Moreover, we plan to validate the benefit of the new kernels in experiments on more realistic nonlinear examples or physical hardware such as those in [3] and [5], settings that pose challenging issues, like imperfect states measurement, among others.

REFERENCES

- [1] R. S. Sutton and A. G. Barto, *Reinforcement learning: An introduction*. MIT press, 1998.
- [2] J. Kober, J. A. Bagnell, and J. Peters, "Reinforcement learning in robotics: A survey," *The International Journal of Robotics Research*, vol. 32, no. 11, pp. 1238–1274, 2013.
- [3] A. Marco, P. Hennig, J. Bohg, S. Schaal, and S. Trimpe, "Automatic LQR tuning based on Gaussian process global optimization," in *IEEE International Conference on Robotics and Automation*, 2016, pp. 270–277.
- [4] F. Berkenkamp, A. P. Schoellig, and A. Krause, "Safe controller optimization for quadrotors with Gaussian processes," in *IEEE International Conference on Robotics and Automation*, 2016, pp. 491–496.
- [5] A. Marco, F. Berkenkamp, P. Hennig, A. P. Schoellig, A. Krause, S. Schaal, and S. Trimpe, "Virtual vs. real: Trading off simulations and physical experiments in reinforcement learning with Bayesian optimization," in *IEEE International Conference on Robotics and Automation*, 2017, pp. 1557–1563.
- [6] J. Schreiter, D. Nguyen-Tuong, M. Eberts, B. Bischoff, H. Markert, and M. Toussaint, "Safe exploration for active learning with gaussian processes," in *Machine Learning and Knowledge Discovery in Databases: European Conference*, 2015, pp. 133–149.
- [7] R. Calandra, A. Seyfarth, J. Peters, and M. P. Deisenroth, "Bayesian optimization for learning gaits under uncertainty," *Annals of Mathematics and Artificial Intelligence*, pp. 1–19, 2015.
- [8] R. Antonova, A. Rai, and C. G. Atkeson, "Sample efficient optimization for learning controllers for bipedal locomotion," in *IEEE International Conference on Humanoid Robots*, 2016, pp. 22–28.
- [9] J. R. Medina, S. Endo, and S. Hirche, "Impedance-based Gaussian processes for predicting human behavior during physical interaction," in *IEEE International Conference on Robotics and Automation*, 2016, pp. 3055–3061.
- [10] H. Abdelrahman, F. Berkenkamp, J. Poland, and A. Krause, "Bayesian optimization for maximum power point tracking in photovoltaic power plants," in *European Control Conference*, 2016, pp. 2078–2083.
- [11] B. D. O. Anderson and J. B. Moore, *Optimal Control: Linear Quadratic Methods*. Mineola, New York: Dover Publications, 2007.
- [12] S. Schaal and C. Atkeson, "Learning control in robotics," *IEEE Robotics Automation Magazine*, vol. 17, no. 2, pp. 20–29, Jun. 2010.
- [13] C. Rasmussen and C. Williams, *Gaussian Processes for Machine Learning*. MIT Press, 2006.
- [14] J. Mockus, *Bayesian approach to global optimization: theory and applications*, ser. Mathematics and its applications. Kluwer Academic Publishers, 1989, vol. 37.
- [15] D. R. Jones, M. Schonlau, and W. J. Welch, "Efficient global optimization of expensive black-box functions," *Journal of Global Optimization*, vol. 13, no. 4, pp. 455–492, 1998.
- [16] H. J. Kushner, "A new method of locating the maximum point of an arbitrary multipeak curve in the presence of noise," *Journal of Basic Engineering*, vol. 86, no. 1, pp. 97–106, 1964.
- [17] N. Srinivas, A. Krause, S. M. Kakade, and M. Seeger, "Gaussian process optimization in the bandit setting: No regret and experimental design," in *International Conference on Machine Learning*, 2010, pp. 1015–1022.
- [18] P. Hennig and C. J. Schuler, "Entropy search for information-efficient global optimization," *The Journal of Machine Learning Research*, vol. 13, no. 1, pp. 1809–1837, 2012.
- [19] Y. Sui, A. Gotovos, J. W. Burdick, and A. Krause, "Safe exploration for optimization with gaussian processes," in *International Conference on Machine Learning*, 2015, pp. 997–1005.
- [20] S. Trimpe, A. Millane, S. Doessegger, and R. D'Andrea, "A self-tuning LQR approach demonstrated on an inverted pendulum," in *IFAC World Congress*, Cape Town, South Africa, Aug. 2014, pp. 11 281–11 287.
- [21] B. D. O. Anderson and J. B. Moore, *Optimal Filtering*. Mineola, New York: Dover Publications, 2005.
- [22] B. Schölkopf and A. J. Smola, *Learning with kernels*. MIT Press, 2002.