Data-Driven Reachability and Support Estimation with Christoffel Functions

Alex Devonport, Forest Yang, Laurent El Ghaoui, and Murat Arcak

***Abstract*— We present algorithms for estimating the forward reachable set of a dynamical system using only a finite collection of independent and identically distributed samples. The produced estimate is the sublevel set of a function called an empirical inverse Christoffel function: empirical inverse Christoffel functions are known to provide good approximations to the support of probability distributions. In addition to reachability analysis, the same approach can be applied to general problems of estimating the support of a random variable, which has applications in data science towards detection of novelties and outliers in data sets. In applications where safety is a concern, having a guarantee of accuracy that holds on finite data sets is critical. In this paper, we prove such bounds for our algorithms under the Probably Approximately Correct (PAC) framework. In addition to applying classical Vapnik-Chervonenkis (VC) dimension bound arguments, we apply the PAC-Bayes theorem by leveraging a formal connection between kernelized empirical inverse Christoffel functions and Gaussian process regression models.**

***Index Terms*— Randomized algorithms; Uncertain systems; Statistical learning; Estimation**

I. Introduction

Reachability analysis is a popular and effective way to guarantee the safety of a system in the face of uncertainty. The primary object of study is the reachable set, which characterizes all possible evolutions of a system under certain constraints on initial conditions and disturbances. Many algorithms in reachability analysis use detailed system information to compute a sound approximation to the reachable set, that is an approximation guaranteed to completely contain (or be contained in) the reachable set. However, in many important applications, such as complex cyber-physical systems that are only accessible through simulations or experiments, this detailed system information is not available, so these algorithms cannot be applied. Applications such as these motivate *data-driven* reachability analysis, which studies algorithms to estimate reachable sets using the type of data that can be obtained from experiments and simulations. These algorithms have the advantage of being able to estimate the reachable sets of any system whose behavior can be simulated or measured experimentally, without requiring any additional mathematical information about the system. The main disadvantage of data-driven reachability algorithms is that generally they cannot provide the same type of soundness guarantees as traditional reachability analysis algorithms; however, they can still guarantee accuracy of the estimates in a probabilistic sense with high confidence, as this article will show.

Data-driven reachability is a rapidly growing area of research within reachability analysis. Many recent developments focus on providing probabilistic guarantees of correctness for data-driven methods that estimate the reachable set directly from data, for instance using PAC analysis from statistical learning theory [8, 30] or by using the scenario approach [21, 35, 16, 28, 15, 9]. The work in [30] is particularly unique in that it provides bounds for a continuum of reachable sets by randomizing the initial set and time horizon. Other works incorporate data-driven elements into more traditional reachability approaches, for instance estimating entities such as discrepancy functions [14] or differential inclusions [11]. Further developments include incorporating data-driven reachability into verification tools for cyber-physical systems [14, 27].

This paper investigates a data-driven reachability algorithm that directly estimates the reachable set from data using the sublevel sets of an empirical inverse Christoffel function, and provides a probabilistic guarantee of accuracy for the method using statistical learning-theoretic methods. Christoffel functions are a class of polynomials defined with respect to measures on $\mathbb{R}^n$: a single measure defines a family of Christoffel function polynomials. When the measure in question is defined by a probability distribution on $\mathbb{R}^n$ the level sets of Christoffel functions are known empirically to provide tight approximations to the support. This support-approximating quality has motivated the use of Christoffel functions in several statistical applications, such as density estimation [19, 20] and outlier detection [2]. Additionally, the level sets have been shown, using the plug-in approach [6], to converge exactly to the support of the distribution (in the sense of Hausdorff measure) when the degree of the polynomial approaches infinity and when the true probability distribution is available [20]. When the true probability distribution is *not* known, as is typically the case in data analysis, the Christoffel function can be empirically estimated using a point cloud of independent and identically distributed (iid) samples from the distribution: this *empirical Christoffel function* still provides accurate estimates for the support, and some convergence results in this case are also known [26].

Submitted to the editors on January 21, 2021. This work is funded in part by the Air Force Office of Scientific Research grant FA9550-21-1-0288, National Science Foundation grant ECCS-1906164, and the Office of Naval Research grant N00014-18-1-2209.

A. Devonport, F. Yang, L. El Ghaoui, and M. Arcak are with the Department of Electrical Engineering and Computer Sciences, University of California, Berkeley. {alex_devonport,forestyang,elghaoui,arcak}@berkeley.edu

In contrast to the asymptotic analysis of Christoffel functions reviewed above, our interest is in developing error bounds that hold with a finite number of samples. This paper is an extension of a conference paper [10] that reported our preliminary work on support set estimation with polynomial Christoffel functions in the context of data-driven reachability. In [10], we investigated empirical inverse Christoffel functions constructed from iid trajectory simulation data, and provided a finite-sample guarantee of the probabilistic accuracy of reachable set estimates produced by sublevel sets of this function. The present paper significantly extends the theory of finite-sample error bounds for support set estimators derived from Christoffel functions by applying techniques from Bayesian PAC analysis, a variation of classical PAC analysis that has been successfully applied to Gaussian process classifiers [29], kernel support vector machines [18], and minimum-volume covering ellipsoids [12]. This extension leverages a formal connection between the kernel empirical inverse Christoffel function investigated by Askari *et al.* [2] and the posterior variance of a Gaussian process regression model. In conjunction with the PAC-Bayes theorem, the connection can be used to derive finite-sample bounds for kernelized empirical inverse Christoffel functions.

The application of Bayesian PAC analysis to the theory of Christoffel function support set estimators has two benefits. First, it allows for the construction of finite-sample guarantees for kernelized inverse Christoffel functions, which to our knowledge have not been proved before. Second, when applied to polynomial empirical inverse Christoffel function estimators, Bayesian PAC analysis can provide guarantees of probabilistic accuracy and confidence with much greater sample efficiency than the finite-sample bounds provided by classical VC dimension bound arguments.

II. Preliminaries

A. Probabilistic Reachability and Estimation of Support

Consider a dynamical system with a state transition function $\Phi(t_1; t_0, x_0, d)$ that maps an initial state $x(t_0) = x_0 \in \mathbb{R}^n$ at time t_0 to a unique final state at time t_1, under a disturbance $d : [t_0, t_1] \to \mathbb{R}^w$. For instance, when the system state dynamics $\dot{x}(t) = f(t, x(t), d(t))$ are known and have unique solutions on the interval $[t_0, t_1]$, then $\Phi(t_1; t_0, x_0, d)$ is the solution of the state dynamics equation at time t_1 with initial condition $x(t_0) = x_0$. For the problem of forward reachability analysis, we are also given an *initial set* $\mathcal{X}_0 \subset \mathbb{R}^n$, a set $\mathcal{D}$ of allowed disturbances and a time range $[t_0, t_1]$. The *forward reachable set* is then defined as the set of all states to which the system can transition in the time range $[t_0, t_1]$ with initial states in $\mathcal{X}_0$ and disturbances in $\mathcal{D}$, that is the set

$$R_{[t_0,t_1]} = \{\Phi(t_1; t_0, x_0, d) : x_0 \in \mathcal{X}_0, d \in \mathcal{D}\}. \quad (1)$$

To tackle the problem of estimating the forward reachable set by statistical means, we add probabilistic structure to the reachability problem by taking random variables X_0 and D supported on $\mathcal{X}_0$ and $\mathcal{D}$ respectively. These random variables then induce a random variable $X = \Phi(t_1; t_0, X_0, D)$, whose support is precisely $R_{[t_0,t_1]}$ and whose probability measure we

symbol	definition
Reachability Analysis	
$\Phi(t_1; t_0, x_0, d)$	State transition function, evolving a state x_0 at time t_0 under disturbance d to a state at time t_1
$\mathcal{X}_0$	Set of initial states
$\mathcal{D}$	Set of disturbances
t_0, t_1	Initial and final times
$R_{[t_0,t_1]}$	Forward reachable set
$\hat{R}_{[t_0,t_1]}$	Approximation of forward reachable set
Probability, Statistical Learning Theory	
$\mathbb{E}[\cdot]$	Expected value of a random variable
$\mathbb{P}(\cdot)$	Probability of an event defined in terms of random variables
$D_{KL}(P\|\|Q)$	Kullback-Leibler (KL) divergence from P to Q
$D_{ber}(p\|\|q)$	KL divergence between Bernoulli distributions with parameters p and q
X	Random variable whose support we wish to estimate
F_1	CDF of the chi-square distribution with 1 degree of freedom
$\mathcal{X}$	Domain of X
P_X	Probability measure of the distribution of X
P_X^N	Probability measure of N iid samples from X
PAC	Probably Approximately Correct
iid	Independent and Identically Distributed
ϵ,δ	accuracy and confidence parameters in PAC guarantees
$\mathcal{C}$	Concept class
$\bar{c}_Q$	"central concept" of the posterior measure Q
P,Q	Prior and posterior probability measures on $\mathcal{C}$
W_P, W_Q	Parametric representations of P and Q
C_P, C_Q	Stochastic estimators: random variables on $\mathcal{C}$ distributed according to P, Q
$\ell(c, x)$	statistical loss function comparing a concept c and a datum x
$r(c)$	risk: average of $\ell(c, x)$ for $x \sim X$
$\hat{r}(c)$	empirical estimate of $r(c)$ from data $x_1, \ldots, x_N$
r_Q	stochastic risk: average of $\ell(c, x)$ for $x \sim X$, $c \sim Q$
$\hat{r}_Q$	empirical estimate of r_Q from data $x_1, \ldots, x_N$
Christoffel Functions	
M_m, $\hat{M}_m$	Matrix of moments of degree $\leq m$ and its empirical estimate
$\hat{M}_{m,\sigma_0}$	Empirical moment matrix with diagonals modified by σ_0
$z_m(x)$	vector of monomials with degree $\leq m$ evaluated at point x
$\hat{\kappa}^{-1}(x)$	Polynomial empirical inverse Christoffel function evaluated at x
$\hat{\kappa}^{-1}(x)$	kernelized empirical inverse Christoffel function
$C(x)$	Christoffel-based support set estimator, output of Algorithms 1,2, and 3
Gaussian Processes	
m, k	prior mean and covariance functions
m_q, k_q	posterior mean and covariance functions
K	kernel Gramian matrix, $K_{ij} = k(x_i, x_j)$
k_D	vector of kernel evaluations on data, $(k_D(x))_i = k(x_i, x)$
$\mathcal{N}(\mu, \Sigma)$	Multivariate normal with mean μ and covariance Σ
$\mathcal{GP}(m, k)$	Gaussian process with mean and covariance functions m, k

TABLE I
SYMBOLS AND ACRONYMS USED IN THIS PAPER.

denote as P_X. A measure-theoretic interpretation $P_X(A)$ for a set A is a measure of overlap between $R_{[t_0,t_1]}$ and A: $P_X(A)$ is nonzero only if A has nonempty intersection with $R_{[t_0,t_1]}$, and $P_X(A) = 1$ only if $R_{[t_0,t_1]} \subseteq A$. A probabilistic interpretation of $P_X(A)$ is that if we take samples x_0 and d of the random variables X_0 and D, then the vector $\Phi(t_1; t_0, x_0, d)$ lies in A with probability $P_X(A)$. These interpretations motivate $P_X(A)$ as a measure of *probabilistic accuracy*: if a set $A \subseteq \mathbb{R}^n$ has a greater measure $P_X(A)$ than a set $B \subseteq \mathbb{R}^n$, then A is a more accurate approximation of the reachable set than B, in the sense that it "misses" less of the probability mass than B does. In the probabilistic version of the forward reachability problem, our goal is to find reachable set approximations $\hat{R}_{[t_0,t_1]}$ such that $P_X(\hat{R}_{[t_0,t_1]})$ is close to 1. In addition, we will seek $\hat{R}_{[t_0,t_1]}$ with low volume, in order to preclude trivial estimates such as $\hat{R}_{[t_0,t_1]} = \mathbb{R}^n$ and to generally minimize the conservatism of the approximation.

The probabilistic relaxation of the forward reachability problem is a statistical problem of *support set estimation* based on a finite set of observations. The support of a random variable is the range of values it can assume: for example, if X admits a probability density function p_X, then the support of X is the closure of the set $\{x : p_X(x) \neq 0\}$. In addition to the control-theoretic application developed above, support set estimation has several applications in statistics and data science, such as outlier and novelty detection [26, 25, 2]. It is therefore useful to consider the problem for general random variables: we will do so for the theoretical developments in this paper, returning to the reachability application in the numerical examples of Section IV. Formally, we address the following problem.

Problem 1: Given accuracy and confidence parameters $\epsilon, \delta \in (0,1)$ and a random variable X whose support lies in a compact domain $\mathcal{X} \subseteq \mathbb{R}^n$, collect data $x_1, \ldots, x_N \overset{\text{i.i.d.}}{\sim} X$ and use them to find a set $c(\epsilon, \delta; x_1, \ldots, x_N) \subset \mathcal{X}$ such that the following bound holds:

$$P_X^N\left(\{x_1, \ldots, x_N : P_X(c(\epsilon, \delta; x_1, \ldots, x_N)) \geq 1 - \epsilon\}\right) \geq 1 - \delta. \tag{2}$$

The bound (2) is known as a Probably Approximately Correct (PAC) bound, which appears frequently in statistical learning theory. The two probability inequalities in (2) are interpreted as assertions of probabilistic accuracy and confidence:

- *accuracy*: the inner inequality $P_X(c(\epsilon, \delta; x_1, \ldots, x_N)) \geq 1 - \epsilon$ asserts that the probabilistic accuracy of the estimator is at least $1 - \epsilon$.
- *confidence*: the outer inequality asserts that the accuracy statement holds with probability $1 - \delta$ with respect to P_X^N. The probability, and hence the confidence, is with respect to the data: P_X^N is the probability measure corresponding to N iid observations drawn from X, so $P_X^N(A)$ for $A \subseteq \mathcal{X}^N$ denotes the probability that $x_1, \ldots, x_N \in A$. Thus the inequality $P_X^N(\{x_1, \ldots, x_N : \cdots\}) \geq 1 - \delta$ asserts that the observed data set $x_1, \ldots, x_N$ belongs, with probability at least $1 - \delta$, to the class of data sets sufficiently informative to yield an estimator c satisfying the accuracy assertion.

For brevity, we drop the arguments of the estimator $c(\epsilon, \delta; x_1, \ldots, x_N)$ from the notation, understanding that an estimator c is always constructed using a given set of data $x_1, \ldots, x_N$, with respect to given parameters ϵ and δ. The sample size N is a fixed problem parameter: indeed, finding a suitable N is part of solving the problem. In addition to the requirements given in Problem 1, we may also impose that the estimator c be drawn from a pre-specified class of admissible estimators. Such a condition allows us to restrict attention to computationally feasible sets, or sets with certain properties such as compactness for cases when the reachable set is known to be compact. In classical PAC analysis, the structure of the pre-specified class also plays a key role in determining an appropriate N.

B. Christoffel Functions

Our main tool for constructing estimators of support, and thereby of forward reachable sets, in this work are empirical inverse Christoffel functions. These functions are well-suited to the problem of support set estimation, because they can functionally encode the statistics of a dataset directly from a collection of independent samples: in the polynomial case this encoding is a sum-of-squares polynomial weighted by an empirical moment matrix, and in the kernel case this encoding is a positive definite kernel function weighted by a kernel Gramian matrix. Most other functional representations of support, such as splines or ellipsoids, follow in a less straightforward way from the data, typically as the solution to a minimum-volume covering problem. Additionally, Christoffel functions are particularly well-suited to Bayesian PAC analysis in a way that many other estimators are not, due to their formal relation to Bayesian Gaussian process regression models.

Given a finite measure P_X on $\mathbb{R}^n$ and a positive integer m, the Christoffel function of order m is defined as the ratio $\kappa(x) = 1/z_m(x)^\top M_m^{-1} z_m(x)$, where $z_m(x)$ is the vector of monomials of degree $\leq m$, and where M_m is the matrix of moments $M_m = \int_{\mathcal{X}} z_m(x) z_m(x)^\top dP_X(x)$. We assume throughout that M_m is positive definite, ensuring that M_m^{-1} exists. The Christoffel function has several important applications in approximation theory [24], where its asymptotic properties are used to prove the regularity and consistency of Fourier series of orthogonal polynomials [34]. For our purposes, it is more convenient to use the *inverse Christoffel function* $\kappa(x)^{-1} = z_m(x)^\top M_m^{-1} z_m(x)$, which is a polynomial of degree $2m$. In Problem 1, and more generally in the problem of estimating a probability distribution from samples, P_X is unknown. In this case, we instead use an empirical estimate for the moment matrix M_m, namely $\hat{M}_m = \frac{1}{N}\sum_{i=1}^N z_m(x_i) z_m(x_i)^\top$. The matrix $\hat{M}_m$ is positive semidefinite: it is additionally positive definite, and hence nonsingular, if $N \geq \binom{n+m}{n}$ and $x_1, \ldots, x_N$ do not all belong to the zero set of a single degree m polynomial. It is useful, both numerically and theoretically, to modify this empirical estimate adding a scaled identity perturbation: thus we take $\hat{M}_{m,\sigma} = \sigma^2 I + \frac{1}{N}\sum_{i=1}^N z_m(x_i) z_m(x_i)^\top$, as our empirical moment matrix in the sequel, where $\sigma^2 > 0$ is a term fixing the magnitude of the perturbation. In addition to its role in developing the kernel extension, the $\sigma^2 I$ term

generally improves the conditioning of the empirical moment matrix and ensures nonsingularity in all cases. The empirical moment matrix $\hat{M}_{m,\sigma}$ itself defines a Christoffel function, whose inverse

$$\hat{\kappa}^{-1}(x) = z_m(x)^\top \hat{M}_{m,\sigma}^{-1} z_m(x) \tag{3}$$

is called the *empirical inverse Christoffel function*.

The dyadic sum $\frac{1}{N}\sum_{i=1}^N z_m(x_i)z_m(x_i)^\top$ can be expressed as the matrix product $\frac{1}{N}ZZ^\top$, where $Z \in \mathbb{R}^{\binom{n+m}{n}\times N}$ is the matrix $Z = \begin{bmatrix} z_m(x_i) & \dots & z_m(x_N)\end{bmatrix}$ of polynomial features. By expressing the dyadic sum this way, we can apply the matrix inversion lemma to express the inverse of the empirical moment matrix as

$$\begin{aligned}\hat{M}_{m,\sigma} &= \left(\sigma^2 I + \tfrac{1}{N}ZZ^\top\right)^{-1} \\ &= \sigma^{-2}\left(I - Z\left(\sigma^2 N I + Z^\top Z\right)^{-1} Z^\top\right).\end{aligned} \tag{4}$$

This expression for $\hat{M}_{m\sigma}$ allows us to rewrite the empirical inverse Christoffel function as

$$\begin{aligned}\hat{\kappa}^{-1}(x) = {} & N\sigma_0^{-2} z_m(x)^\top z_m(x) \\ & - N\sigma_0^{-2} z_m(x)^\top Z\left(\sigma_0^2 I + Z^\top Z\right)^{-1} Z^\top z_m(x),\end{aligned} \tag{5}$$

where we have made the change of variables $\sigma^2 = \sigma_0^2/N$. The vector z_m enters (5) only through the inner products $z_m(x_i)^\top z_m(x_j)$: The matrix $Z^\top Z \in \mathbb{R}^{N\times N}$ has elements $(Z^\top Z)_{ij} = z_m(x_i)^\top z_m(x_j)$, and the matrix-vector product $Z^\top z_m(x)$ has elements $(Z^\top z_m(x))_i = z_m(x_i)^\top z_m(x)$. By replacing the inner product $z_m(x_i)^\top z_m(x_j)$ with an arbitrary positive definite[1] function $k : \mathbb{R}^n \times \mathbb{R}^n \to \mathbb{R}$ and rescaling by a factor of σ_0^2/N, we obtain the kernelized variant of the empirical inverse Christoffel function,

$$\hat{\kappa}^{-1}(x) = k(x,x) - k_D(x)^\top\left(\sigma_0^2 I + K\right)^{-1} k_D(x), \tag{6}$$

where $K \in \mathbb{R}^{N\times N}$ and $k_D(x) \in \mathbb{R}^N$ are defined as

$$K_{ij} = k(x_i, x_j), \quad (k_D(x))_i = k(x_i, x). \tag{7}$$

III. Christoffel Function Estimators of Support

Algorithms 1 and 2 are procedures to estimate the support of a random variable with a sublevel set of an empirical inverse Christoffel function, where the only information needed from the random variable is a collection of iid samples. Algorithm 1 is designed to satisfy a classical PAC bound. This has the advantages of providing an *a priori* sample bound, and of admitting a fairly direct proof, which is given in Section III-A. The essence of the proof is to demonstrate that the sublevel sets of a polynomial empirical inverse Christoffel function of a given order inhabit a concept class of known VC dimension. This argument is valid for polynomial Christoffel functions of any order, but it is generally not valid for kernelized Christoffel functions. Indeed, the classes of sublevel sets of certain kernelized empirical inverse Christoffel functions can have infinite VC dimension, so a classical PAC bound is not possible in general for kernelized empirical inverse Christoffel functions. Algorithm 2 is designed to satisfy a Bayesian PAC bound which is developed in Section III-B. Unlike the classical PAC bound provided for Algorithm 1, this Bayesian PAC bound is applicable to all kernelized empirical inverse Christoffel functions, including those whose sublevel sets have infinite VC dimension. When applied to polynomial empirical inverse Christoffel functions as a special case, we find that it is more sample-efficient than the classical PAC bound: in some of the examples in Section IV, the Bayesian PAC bound requires an order of magnitude fewer samples to achieve the same accuracy and confidence as that guaranteed by the classical PAC bound. The disadvantages of the Bayesian PAC approach is that the required number of samples is not known *a priori*, since certain terms in the bound depend on the data. Algorithm 2 therefore takes an iterative approach, taking samples in batches and re-evaluating the Bayesian PAC bound after each batch until it reaches the desired level of accuracy.

[1]Here, and throughout the paper, we mean positive definite in the sense of reproducing kernel Hilbert spaces and kernel machines, which is that a square matrix K with elements $(K)_{ij} = k(x_i, x_j)$ is a positive definite matrix.

Algorithm 1 To estimate a support set by a polynomial empirical inverse Christoffel function satisfying a classical PAC bound, from the authors' prior work [10]

Inputs: random variable X with support in $\mathcal{X}$; polynomial order $m \in \mathbb{N}_+$; PAC parameters $\epsilon, \delta \in (0,1)$; noise parameter $\sigma_0^2 \in \mathbb{R}_{++}$;
$N \leftarrow \lceil \frac{5}{\epsilon}\left(\log\frac{4}{\delta} + \binom{n+2m}{n}\log\frac{40}{\epsilon}\right)\rceil$
for $i \in \{1,\dots,N\}$ **do**
  sample $x_i \sim X$
end for
$\hat{M}_{m,\sigma_0} \leftarrow \sigma_0^2 I + \frac{1}{N}\sum_{i=1}^N z_m(x_i)z_m(x_i)^\top$
$\alpha \leftarrow \max_i z_m(x_i)^\top \hat{M}_{m,\sigma_0}^{-1} z_m(x_i)$
$C(x) = z_m(x)^\top \hat{M}_{m,\sigma_0}^{-1} z_m(x)$;
return $\mathbb{1}\{C(x) \le \alpha\}$;

Remark 1 (Subsets of state space and output observables): In some reachability problems, we are only interested in computing a reachable set for a subset of the state variables. For example, suppose the state is $(x_1,\dots,x_n) \in \mathbb{R}^n$, and we wish to verify a safety specification involving only the states $x_1,\dots,x_s$, where $s < n$: a reachable set for the states $x_1,\dots,x_s$ would suffice for this problem. In cases like this, the algorithms presented in this section can be modified to use only the first s elements of the samples. The output of the algorithm is then an empirical inverse Christoffel function with domain $\mathbb{R}^s$ whose sublevel set $\hat{R}_{[t_0,t_1]}$ estimates the reachable set for the reduced set of states. Additionally, we may also be interested solely in some measurable output of the system such that the space of the output is of lower dimension than the state space. In this case, we may apply the algorithms in this section directly to the support of the observed outputs. In the sequel, we refer to sch variations of the algorithms in this section as *reduced-state* variations. These variations are more data-efficient than the full-state algorithms since the supports being estimated are confined to a smaller space.

Remark 2 (Selecting m and k in practice): The algorithms that follow require the user to select either a polynomial order m or a kernel function k. Any choice of m or k is able to obtain an arbitrarily high probabilistic accuracy and confidence—that is, satisfying a PAC bound of the form (2) with ϵ, δ arbitrarily close to zero)—so the choice of m or k only affects the geometric fidelity of the estimator and the amount of data required to provably satisfy (2).

For the polynomial case, the trade-off is clearly defined: a higher order gives greater fidelity, but requires more data to be proven probabilistically accurate. A heuristic to select m in practice is to perform a "trial run" of the computation with a small pilot dataset: the trial estimator will have no guarantee, but may give insight into what order of Christoffel function is appropriate to use when running the full algorithm. Machine precision also places a practical upper bound on the order, as the empirical moment matrix becomes increasingly ill-conditioned: the details depend on the machine and the choice of σ_0, but in our experience, $m \geq 30$ is usually where trouble begins.

For the kernel case, there is generally not a clear relation between how the choice of kernel affects the fidelity and sample complexity of the algorithm. However, for kernels that admit a length-scale parameter, such as the squared-exponential kernel (26) discussed in Section IV, the length-scale plays a similar role to the order in the polynomial case, where a smaller length-scale gives a higher-fidelity estimate at a greater sample cost. The length-scale can be selected with a pilot test, similar to the polynomial order. However, if side information about the rough size of the reachable set is available, such as from an ellipsoidal or p-norm ball estimate of the reachable set (as in [9]), then this information can be incorporated directly into the length-scale.

A. Classical PAC Analysis

PAC bounds originate in study of empirical risk minimization problems in statistical learning theory. Our strategy to prove a PAC bound for Algorithm 1 is to express Problem 1 as an empirical risk minimization problem and to then apply the tools of statistical learning theory.

In empirical risk minimization, the objective is to match a concept $c \subseteq \mathcal{X}$ from a pre-specified concept class $\mathcal{C} \subseteq 2^{\mathcal{X}}$ to an unknown random variable X supported on $\mathcal{X}$ using only a finite set of iid observations $x_1, \ldots, x_N$ of X. How well a concept matches X is quantified by the statistical risk $r(c) = \mathbb{E}\left[\ell(c, X)\right]$ defined by a loss function $\ell : \mathcal{C} \times \mathcal{X} \to \mathbb{R}_+$ and the unknown measure P_X: a lower risk indicates a better match. Since we do not know P_X, we cannot directly evaluate the statistical risk. However, we can use the empirical risk $\hat{r}(c) = \frac{1}{N}\sum_{i=1}^{N} \ell(c, x_i)$ as a proxy for the true risk, and select a concept to match the data on the basis of minimizing the empirical risk.

Whether empirical risk minimization actually selects a concept with low risk depends on how much $\hat{r}(c)$ differs from $r(c)$. A classical PAC bound provides a bound on the difference $r(c) - \hat{r}(c)$, or the absolute difference, that holds with high probability. We use the following result from [1], which gives a quantitative sample bound that depends on the Vapnik-Chervonenkis (VC) dimension [32] of the concept class. The VC dimension of a concept is a combinatorial measure of its complexity based on the expressiveness of its concepts.

Lemma 1 ([1], Corollary 4): Let $\mathcal{C}$ be a concept class of sets with VC dimension $\leq d$, and let $\ell : \mathcal{C} \times \mathcal{X} \to \{0, 1\}$ denote a $\{0, 1\}$-valued loss function. If

$$N \geq \frac{5}{\epsilon}\left(\log\frac{4}{\delta} + d\log\frac{40}{\epsilon}\right), \tag{8}$$

and if $\hat{r}(c) = 0$, then $P_X^N\left(\{x_1, \ldots, x_N : r(c) \leq \epsilon\}\right) \geq 1 - \delta$. A concept class with higher VC dimension generally provides greater-fidelity estimates than one with lower VC dimension, but is also more prone to overfitting: informally, this is the reason why a concept class with higher VC dimension requires a larger sample bound for the same accuracy and confidence than one with lower VC dimension.

To apply Lemma 1, we must show that the sublevel sets of a polynomial empirical inverse Christoffel function belong to a concept class of bounded VC dimension. One such class is the class of superlevel sets of degree $2k$ polynomials: the following Lemma from [13], provides a bound on the VC dimension.

Lemma 2 ([13], Theorem 7.2): Let V be a vector space of functions $g : \mathbb{R}^n \to \mathbb{R}$ with dimension d. Then the class of sets $\mathrm{Pos}(V) = \{\ \{x : g(x) \geq 0\}, g \in V\}$ has VC dimension $\leq d$.

The PAC bound, and hence the validity of Algorithm 1 follows from Lemmas 2 and 1 by framing the support estimation problem as one of empirical risk minimization.

Theorem 1 ([10]): The support set estimate produced by Algorithm 1, that is the set $\{x \in \mathcal{X} : C(x) \leq \alpha\}$ where $C(x) = z_m(x)^\top \hat{M}_{m,\sigma_0}^{-1} z_m(x)$, $\alpha = \max_i C(x_i)$, satisfies the PAC bound $P_X^N(\{x_1, \ldots, x_N : P_X(\{x \in \mathcal{X} : C(x) \leq \alpha\}) \geq 1-\epsilon\}) \geq 1-\delta$, and thereby solves Problem 1 with parameters ϵ, δ.

Proof: Let $\mathcal{C} = \mathrm{Pos}(\mathbb{R}[x]_{2m}^n)$, and $\ell(c, x) = \mathbb{1}\{x \notin c\}$. Note that the set $\{x \in \mathbb{R}^n : C(x) \leq \alpha\}$ is a member of $\mathrm{Pos}(\mathbb{R}[x]_{2m}^n)$, since it can be expressed as $c = \{x \in \mathbb{R}^n : \alpha - C(x) \geq 0\}$. Since the dimension of $\mathbb{R}[x]_{2m}^n$ is $\binom{n+2m}{n}$, the VC dimension of $\mathrm{Pos}(\mathbb{R}[x]_{2m}^n)$ is $\leq \binom{n+2m}{n} = d$ by Lemma 2. For $\ell(c, x) = \mathbb{1}\{x \notin c\}$, the statistical risk is $r(c) = \mathbb{E}\left[\mathbb{1}\{x \notin c\}\right] = 1 - P_X(c)$, and its empirical counterpart is $\hat{r}(c) = \sum_{i=1}^{N} \mathbb{1}\{x_i \notin c\}$. The empirical risk is zero for any set c that encloses $x_1, \ldots, x_N$. The set $\{x \in \mathbb{R}^n : C(x) \leq \alpha\}$ encloses $x_1, \ldots, x_N$ by construction, meaning that $\hat{r}(\{x \in \mathbb{R}^n : C(x) \leq \alpha\}) = 0$. By applying Lemma 1 for this choice of $\mathcal{C}$, ℓ, and m, we find that if $N \geq \frac{5}{\epsilon}\left(\log\frac{4}{\delta} + \binom{n+2m}{n}\log\frac{40}{\epsilon}\right)$, then $P_X^N\left(\{x_1, \ldots, x_N\} : 1 - P_X(\{x \in \mathbb{R}^n : C(x) \leq \alpha\}) \leq \epsilon\right) \geq 1 - \delta$. Since Algorithm 1 selects N to be the smallest integer such that $N \geq \frac{5}{\epsilon}\left(\log\frac{4}{\delta} + \binom{n+2m}{n}\log\frac{40}{\epsilon}\right)$, it follows that the stated PAC bound holds for the output of Algorithm 1. ■

B. Bayesian PAC Analysis

Algorithm 2 To estimate a support set by a kernelized empirical inverse Christoffel function satisfying a Bayesian PAC bound.

Inputs: random variable X with support in $\mathcal{X}$; positive definite kernel function k; PAC parameters $\epsilon, \delta \in (0,1)$; noise parameter $\sigma_0^2 \in \mathbb{R}_{++}$; initial sample size N_0; batch size N_b; threshold η.
$N \leftarrow N_0$
$D \leftarrow (x_1, \ldots, x_N) \overset{\text{i.i.d.}}{\sim} X$
$i \leftarrow 0$
$\epsilon^0 \leftarrow 1$
while $\epsilon^i > \epsilon$ **do**
  $i \leftarrow i + 1$
  append $(x_{N+1}, \ldots, x_{N+N_b}) \overset{\text{i.i.d.}}{\sim} X$ to D
  $N \leftarrow N + N_b$
  $K_{\sigma_0} \leftarrow \sigma_0^2 I + K$
  define $C : \mathcal{X} \to \mathbb{R}_+$ to be $C(x) = k(x,x) - k_D(x) K_{\sigma_0}^{-1} k_D(x)$;
  Evaluate $\bar{r}$ as in (17)
  $\epsilon_i \leftarrow \frac{\bar{r} + \frac{2}{N}\log(\frac{\pi^2 i^2}{6\delta})}{1 - F_1(1)}$, F_1 as in (16)
end while
return $\mathbb{1}\{C(x) \le \eta\}$

Bayesian PAC analysis bounds the deviation of the expected values of the true and empirical risks with respect to a data-dependent probability measure. Given a prior measure P over $\mathcal{C}$ and a posterior measure Q derived from the prior and the observations, we define the expected risk $r_Q = \mathbb{E}[\ell(c, X)]$ and empirical expected risk $\hat{r}_Q = \mathbb{E}\left[\frac{1}{N}\sum_{i=1}^N \ell(c, x_i)\right]$ where $c \sim Q$. Equivalently, P and Q define random variables C_P, C_Q supported on $\mathcal{C}$, called the prior and posterior *stochastic estimators*: r_Q and $\hat{r}_Q$ are the true and empirical risks of C_Q. A Bayesian PAC bound is a bound on the deviation between r_Q and $\hat{r}_Q$. Bayesian PAC bounds can be used to provide an error bound for a single classifier which captures the central behavior of Q, which we call the *central concept* and denote as $\bar{c}_Q$. To verify that Algorithm 2 provides a valid solution to Problem 1, we show that its output is the central concept of a posterior stochastic estimator and use a Bayesian PAC bound to show that a bound of the form (2) holds.

The most common tool to construct Bayesian PAC bounds is the PAC-Bayes theorem developed by McAllester [22], Seeger [29], and others [17]. We use the variation due to Seeger. This theorem assumes that the concept class admits a parameterization which can be infinite-dimensional.

Theorem 2 (PAC-Bayes Theorem, adapted from [29, 17]): Consider a concept class $\mathcal{C}$ admitting a parametrization by $w \in \mathcal{W}$. Let the loss function be zero-one valued, that is $\ell : \mathcal{C} \times \mathcal{X} \to \{0,1\}$. The following bound holds for all measures P, Q over the concept class $\mathcal{C}$ defined by measures W_P and W_Q over $\mathcal{W}$ such that W_Q is absolutely continuous with respect to W_P:

$$P_X^N\left(\{x_1, \ldots, x_N : D_{ber}(\hat{r}_Q || r_Q) \le \gamma\}\right) \ge 1 - \delta, \quad (9)$$

where $\gamma = (D_{KL}(W_Q||W_P) + \log\frac{N+1}{\delta})/N$. Here, $D_{KL}(W_P||W_Q)$ denotes the Kullback-Leibler (KL) divergence between W_P and W_Q, and $D_{ber}(q||p)$ denotes the KL divergence between two Bernoulli distributions with parameters q and p, given by the formula

$$D_{ber}(q||p) = q\log\frac{q}{p} + (1-q)\log\frac{1-q}{1-p}. \quad (10)$$

For a given set of data $x_1, \ldots, x_N$, confidence parameter δ, and a prior measure P chosen independently of the data, the inequality (9) provides a family of Bayesian PAC bounds, one for each posterior measure Q.

We use the PAC-Bayes theorem in the proof of Theorem 3, which asserts the validity of Algorithm 2. First, we construct prior and posterior stochastic estimators C_P and C_Q, corresponding to measures P, Q over a concept class, which admit a sublevel set of the empirical inverse Christoffel function as a central concept; namely $\bar{c}_Q = \{x : \kappa^{-1}(x) \le \eta\}$ for a given positive η. Next, we express a formula to compute the empirical stochastic risk $\hat{r}_Q$ of C_Q from the data. Then, we establish a bound on the true stochastic r_Q in terms of $\hat{r}_Q$ using the PAC-Bayes theorem. Finally, we prove a bound on the true risk $r(\bar{c}_Q)$ of the central concept in terms of r_Q. This sequence of bounds combines to yield a bound of the form (2) computable in terms of known data.

Theorem 3: Denote C^i as the inverse Christoffel function constructed during the ith iteration of Algorithm 3.2. We have the following PAC bound on all the inverse Christoffel functions constructed during the algorithm:

$$\mathbb{P}(\forall i \ge 1,\, P_X(\{x : C^i(x) \le \eta\}) \ge 1 - \epsilon^i) \ge 1 - \delta. \quad (11)$$

Thus, with confidence δ, upon the termination condition of Algorithm 3.2, we are left with a support set estimate of probability mass $\ge 1 - \epsilon$.

In addition to verifying the validity of the terminal output of Algorithm 2, Theorem 3 justifies the use of Algorithm 2 in an "any time algorithm" fashion, that is as an algorithm whose output is verified even if execution is stopped prematurely. The execution of Algorithm 2 will terminate as long as the growth of $D_{KL}(\mathcal{N}\left(0, (\sigma_0^{-1}I + K^{-1})^{-1}\right)||\mathcal{N}(0,K))$ is $o(N)$: determining the conditions under which this growth condition holds is a topic for future research.

We now develop the constructions used in the proof, starting with the prior and posterior stochastic estimators for the kernel case. We take

$$C_P = \{x : g_p(x)^2 \le \eta\}, \quad C_Q = \{x : g_q(x)^2 \le \eta\}, \quad (12)$$

where g_p and g_q are the prior and posterior of a general Gaussian process regression model with prior kernel k, conditioned on the observations $x_1, \ldots, x_N$, $y_1 = \ldots = y_N = 0$ with observation noise level σ_0^2. [2] The corresponding concept class is the class of η-sublevel sets of functions in the support of

[2]Appendix I provides background on the theory of Gaussian process regression models.

g_p, which depends on the choice of kernel. According to (28), g_q has posterior mean $m_q = 0$ and variance

$$\mathrm{Var}_{g_q}(x) = k(x,x) - k(X,x)^\top \left(\sigma^2 I_N + K(X,X)\right)^{-1} k(X,x). \tag{13}$$

We take the posterior central concept to be $\bar{c}_Q = \{x : \mathbb{E}\left[g_q(x)^2\right] \leq \eta\}$. Since $\mathbb{E}\left[g_q(x)\right] = m_q(x) = 0$ for all $x \in \mathcal{X}$, we know $\mathbb{E}\left[g_q(x)^2\right] = \mathrm{Var}_{g_q}(x)$. This means that the posterior central concept is

$$\begin{aligned}\bar{c}_Q &= \{x : k(x,x) \\ &\quad - k(X,x)^\top \left(\sigma^2 I_N + K(X,X)\right)^{-1} k(X,x) \leq \eta\} \\ &= \{x : \kappa^{-1}(x) \leq \eta\}\end{aligned} \tag{14}$$

as desired.

Next, we construct the sequence of bounds, starting with the formula for the empirical stochastic risk of C_Q in terms of known data.

Lemma 3: For the zero-one membership loss $\ell(c,x) = \mathbb{1}\{x \notin c\}$, the empirical stochastic risk of the posterior stochastic estimators C_Q defined in (12) is

$$\hat{r}_Q = \frac{1}{N}\sum_{i=1}^{N} 1 - F_1\left(\frac{\eta}{\kappa^{-1}(x_i)}\right), \tag{15}$$

where F_1 is the CDF of the chi-square distribution with one degree of freedom, that is

$$F_1(x) = \mathbb{P}\left(Z^2 \leq x\right) \text{ where } Z \sim \mathcal{N}(0,1). \tag{16}$$

The proofs of this Lemma and the other Lemmas in this section are deferred to Appendix II.

Next, we use the PAC-Bayes theorem to bound the stochastic risk r_Q by the empirical stochastic risk $\hat{r}_Q$.

Lemma 4: Let $x_1, \ldots, x_N \overset{\text{i.i.d.}}{\sim} X$ denote a set of observations used to construct C_Q from C_P in (12). The stochastic risk r_Q is bounded by $\overline{r} \in (0,1)$, where

$$\overline{r} = \sup\left\{\beta : D_{\text{ber}}(\hat{r}_Q || \beta) \leq \gamma_k\right\} \tag{17}$$

with confidence $1-\delta$, where

$$\gamma_k = \frac{(D_{KL}(\mathcal{N}\left(0, (K^{-1}+\sigma_0^{-2}I)^{-1}\right)||\mathcal{N}(0,K)) + \log\frac{N+1}{\delta}}{N}.$$

Since $D_{ber}(q||p)$ is convex in (q,p) and equal to zero for $q = p$, the set in (17) is an interval containing $\hat{r}_Q$. Once $\hat{r}_Q$ and the right-hand side of the inequality in (17) are evaluated, the supremum $\overline{r}$ can be computed using a scalar root-finding procedure to solve $D_{\text{ber}}(\hat{r}_Q||\beta) - (D_{KL}(\mathcal{N}\left(0, (K^{-1}+\sigma_0^{-2}I)^{-1}\right)||\mathcal{N}(0,K)) + \log\frac{N+1}{\delta})/N = 0$ over the interval $\beta \in [\hat{r}_Q, 1)$.

Finally, we relate the statistical risk of $r(\bar{c}_Q)$ to r_Q.

Lemma 5: The statistical risk $r(\bar{c}_\eta)$ of the posterior central concept and the stochastic risk r_Q of the posterior stochastic estimator satisfy the bound $r(\bar{c}_Q) \leq \frac{1}{1-F_1(1)} r_Q \approx 3.15 r_Q$.

When combined, the sequence of bounds above provide a bound of the form (2) that holds independently for each iteration of Algorithm 3. Applying a union bound argument to provide a guarantee that holds uniformly over iterations forms the central argument of the proof of Theorem 3.

Proof: [Proof (of Theorem 3)] The bound is trivially satisfied at the beginning of execution, since $\epsilon^0 \leftarrow 1$. Next, let $i > 0$, and let C_Q^i denote the stochastic classifier $\{g_Q^i(x)^2 \leq \eta\}$, where $g_Q^i(x) \sim \mathcal{N}(0, k(x,x) - k_{D^i}(x)^\top(\sigma_0^2 I + K^i)k_{D^i}(x))$, with the i superscripts signifying using the dataset accumulated so far at iteration i. Let r_Q^i denote the risk of C_Q^i. By Lemma 3.7, we have $\forall i \geq 1$, $\mathbb{P}(r_Q^i > (1-F_1(1))\epsilon^i) \leq \frac{6\delta}{\pi^2 i^2}$. By a union bound, $\mathbb{P}(\exists i,\, r_Q^i > (1-F_1(1))\epsilon^i) \leq \sum_{i\geq 1}\frac{6\delta}{\pi^2 i^2} = \delta$. Thus, with probability at least $1-\delta$, every $r_Q^i \leq \epsilon^i$. On this event, by Lemma 3.8, we have $\forall i \geq 1$, $P_X(\{x : C^i(x) > \eta\}) \leq \frac{r_Q^i}{1-F_1(1)} = \epsilon^i$ as desired. ∎

Algorithm 3 To estimate a support set by a polynomial empirical inverse Christoffel function satisfying a Bayesian PAC bound.

Inputs: random variable X with support in $\mathcal{X}$; Christoffel function order m; PAC parameters $\epsilon, \delta \in (0,1)$; noise parameter $\sigma_0^2 \in \mathbb{R}_{++}$; initial sample size N_0; batch size N_b;
$N \leftarrow N_0$
$D \leftarrow (x_1, \ldots, x_N) \overset{\text{i.i.d.}}{\sim} X$
$i \leftarrow 0$
$\epsilon^0 \leftarrow 1$
while $\epsilon^i > \epsilon$ **do**
  $i \leftarrow i+1$
  append $(x_{N+1}, \ldots, x_{N+N_b}) \overset{\text{i.i.d.}}{\sim} X$ to D
  $N \leftarrow N + N_b$
  define $C : \mathcal{X} \to \mathbb{R}_+$ to be $C(x) = z_m(x)^\top \hat{M}_{m,\sigma_0}^{-1} z_m(x)$;
  evaluate $\overline{r}$ as in (20)
  $\epsilon_i \leftarrow \frac{\overline{r} + \frac{2}{N}\log(\frac{\pi^2 i^2}{6\delta})}{1-F_1(1)}$, F_1 as in (16)
end while
return $\mathbb{1}\{C(x) \leq \eta\}$

C. Bayesian PAC Analysis: the Polynomial Case

With the general kernel case settled, we now consider the polynomial case in particular. Since the kernel case reduces to the polynomial case by the kernel $k(x,y) = z_m(x)^\top z_m(y)$, we have in a sense already provided a bound for the polynomial empirical inverse Christoffel function by means of Bayesian PAC analysis. However, we can construct a prior and posterior stochastic estimator for the polynomial case which avoids direct use of the $N \times N$ kernel Gramian, which can be computationally advantageous. The special prior and posterior stochastic estimators are

$$\begin{aligned} C_P &= \{x : (W_P^\top z_m(x))^2 \leq \eta\}, \\ C_Q &= \{x : (W_Q^\top z_m(x))^2 \leq \eta\}, \end{aligned} \tag{18}$$

where $W_P \sim \mathcal{N}\left(0, \sigma_0^{-2} I\right)$, $W_Q \sim \mathcal{N}\left(0, \hat{M}_{m,\sigma_0}^{-1}\right)$.

Notice that $W_P^\top z_m$ and $W_Q^\top z_m$ are Gaussian processes: indeed, they correspond to the prior and posterior of a general Gaussian process regression model with prior kernel $k(x,y) = z_m(x)^\top z_m(y)$, conditioned on the observations $x_1, \ldots, x_N$,

$y_1 = \ldots = y_N = 0$ with observation noise level σ_0^2. We take the central concept $\bar{c}_Q$ of C_Q to be the η-sublevel set

$$\begin{aligned}\bar{c}_Q &= \{x : \mathbb{E}\left[(W_Q^\top z_m(x))^2\right] \leq \eta\} \\ &= \{x : z_m(x)^\top \hat{M}_{m,\sigma_0}^{-1} z_m(x) \leq \eta\},\end{aligned} \tag{19}$$

that is the η-sublevel set of the polynomial empirical inverse Christoffel function. Applying the PAC-Bayes theorem to this construction yields the following alternative to Lemma 4.

Lemma 6: Let $x_1, \ldots, x_N \overset{\text{i.i.d.}}{\sim} X$ denote a set of observations used to construct C_Q from C_P in (18). The stochastic risk r_Q is bounded by $\overline{r} \in (0,1)$, where

$$\overline{r} = \sup\left\{\beta : D_{\text{ber}}(\hat{r}_Q || \beta) \leq \gamma_p\right\}, \tag{20}$$

where

$$\gamma_p = \frac{1}{N}\left(D_{KL}(\mathcal{N}\left(0, (\sigma_0^2 I + \hat{M}_{m,\sigma_0})^{-1}\right) || \mathcal{N}\left(0, \sigma_0^{-2} I\right)) + \log\frac{N+1}{\delta}\right)$$

Using this alternative lemma, we obtain a validation for Algorithm 3.

Corollary 1: At each stage i of execution, the empirical inverse Christoffel function constructed in Algorithms 3 satisfies the PAC bound (11).

Proof: The argument to verify Algorithm 1 is identical to that used in the proof of Theorem 3, except that Lemma 6 is used instead of Lemma 4. ∎

Remark 3: Algorithms 2 and 3 require that a threshold parameter η be selected *a priori* based on the kernel. For instance, if a squared exponential kernel $k(x,y) = \exp(-\|x-y\|^2/(2\ell)^2)$ is used in Algorithm 2, the resulting empirical inverse Christoffel function will always have values in $[0,1]$, with values generally smaller close to data points: thus choosing a value between 0 and 1 is a suitable choice, with smaller values yielding finer approximations of the support set. For Algorithm 3, a reasonable heuristic is to select $\eta = \binom{n+2m}{n}/\epsilon$: one can show that the expected value of the true inverse Christoffel function of order m is $\binom{n+2m}{n}$ when the input is distributed according to X, so by Markov's inequality the probability mass of the $\binom{n+2m}{n}/\epsilon$-sublevel set of the true inverse Christoffel function is at least $1-\epsilon$.

D. Numerical Considerations for Large Datasets

As the sample size N grows, the calculations in Algorithm 2 involving the kernel matrix K can become computation- and memory-intensive. In particular, evaluating $\kappa^{-1}(x)$ to compute the support set estimate and computing the KL divergence that appears in (17) both require the construction of an $N \times N$ matrix and an $O(N^3)$ matrix inversion. Computational difficulties related to the size of the K matrix are well known in the field of kernel machines; in response, a wealth of approximation techniques have been developed to reduce compute and memory requirements at the cost of fidelity. These approximation techniques can be used to improve the efficiency of evaluating the kernelized empirical inverse Christoffel function and its construction via Algorithm 2.

For example, to reduce the speed and memory requirements of evaluating $\kappa^{-1}(x)$, we can replace the kernel matrix K with its rank-r Nyström approximation [33]. The Nyström approximation is a method to construct low-rank approximations of Gramian matrices, such as the kernel matrix K, which has a simple expression in terms of block submatrices of the original matrix. Specifically, the rank-r Nyström approximation of the kernel matrix K has the form

$$\tilde{K} = K_{Nr} K_{rr}^{-1} K_{Nr}, \tag{21}$$

where $K_{Nr} \in \mathbb{R}^{N\times r}$, $K_{rr} \in \mathbb{R}^{r\times r}$ are submatrices of K whose i,j elements are $k(x_i, x_j)$. Making the substitution $K \mapsto \tilde{K}$ and applying the matrix inversion lemma to $\kappa^{-1}(x)$ yields

$$\begin{aligned}\tilde{\kappa}^{-1}(x) &= k(x,x) \\ &\quad - k_D(x)^\top (\sigma_0^{-2} I + K_{Nr} K_{rr}^{-1} K_{Nr})^{-1} k_D(x) \\ &= k(x,x) \\ &\quad - \sigma_0^{-2} k_D(x)^\top k_D(x) - k_D(x,x)^\top V k_X(x),\end{aligned} \tag{22}$$

where

$$V = K_{Nr}(\sigma_0^2 K_{rr} + K_{rN} K_{Nr})^{-1} K_{rN}$$

To numerically compute the final expression, we need only invert an $r \times r$ matrix instead of an $N \times N$ one; indeed, we do not need to explicitly construct an $N \times N$ matrix at all.

Next, we consider a method to over-approximate the KL divergence based on the r largest eigenvalues of K. Since the KL divergence $D_{KL}(Z_0||Z_1)$ between N-dimensional normal random variables $Z_0 \sim \mathcal{N}(\mu_0, \Sigma_0)$ and $Z_1 \sim \mathcal{N}(\mu_1, \Sigma_1)$ has the expression

$$\begin{aligned}D_{KL}(Z_0||Z_1) &= \tfrac{1}{2}\log\det\Sigma_1\Sigma_0^{-1} \\ &\quad + \tfrac{1}{2}\text{tr}\Sigma_1^{-1}\left((\mu_0-\mu_1)(\mu_0-\mu_1)^\top + \Sigma_0\right) \\ &\quad - \tfrac{N}{2}.\end{aligned} \tag{23}$$

For $\Sigma_0 = (\sigma_0^{-2} I + K^{-1})^{-1}$, $\Sigma_1 = K$, $\mu_0 = \mu_1 = 0$, (23) reduces to

$$\tfrac{1}{2}\log\det(I + \sigma_0^{-2} K) + \tfrac{1}{2}\text{tr}\left((I + \sigma_0^{-2} K)^{-1}\right) - \tfrac{N}{2}. \tag{24}$$

Since $\log(1+\sigma_0^{-2}x)$ and $1/(1+\sigma_0^{-2}x)$ are analytic for $x \geq 0$, we can apply the spectral mapping theorem [4, Sec. 4.7] to (24) to obtain an expression for the KL divergence in terms of the eigenvalues $\lambda_1, \ldots, \lambda_N$ of K, namely

$$= \frac{1}{2}\sum_{i=1}^{N}\left(\log(1+\sigma_0^{-2}\lambda_i) + \frac{1}{1+\sigma_0^{-2}\lambda_i} - 1\right). \tag{25}$$

Numerically computing the KL divergence with the expression (24) requires an explicit construction of the K matrix, and the inverse of an $N \times N$ matrix: this requires $O(N^3)$ operations and $O(N^2)$ memory. Using (25) instead of (24) to compute the KL divergence with the full set of eigenvalues does not generally yield an improvement, since computing the eigenvalues of K is also $O(N^3)$. However, since K is a symmetric positive definite matrix, the eigenvalues are all positive, and the m largest eigenvalues can be computed in less than $O(N^3)$ time, for instance by a Lanczos-type algorithm [31, ch. 9]. Let λ_p denote the p^{th} largest eigenvalue: Since (25) is a nondecreasing function in each λ_i, the approximation $\lambda_i \approx \lambda_p$ for λ_i such that $\lambda_i < \lambda_p$ yields an upper bound on the KL divergence that can be computed in less than $O(N^3)$ time.

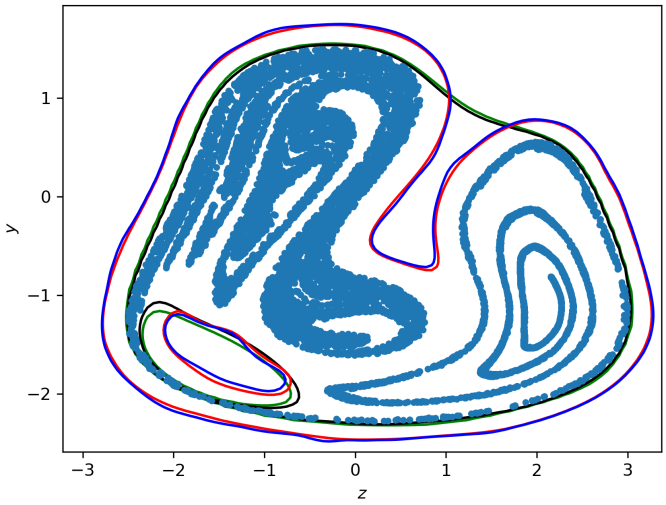

Fig. 1. Results of Algorithms 1, 2 and 3 on the Duffing oscillator reachability problem. Black contour: output of Algorithm 1. Green contour: output of Algorithm 3. Red contour: output of Algorithm 2. Blue contour: output of Algorithm 2, over-approximated using the Nyström approximation with 1,000 samples. Blue dots: samples used in Algorithm 3.

IV. Examples

This section demonstrates how Algorithms 1, 2, and 3 can be used to make accurate estimates of forward reachable sets. These examples were run on Savio, a high-performance computing cluster managed by the University of California at Berkeley. Specifically, each experiment used a single `savio2_bigmem` node comprising 20 CPUs running at 2.3 GHz and 128 GB of memory. In all experiments, we use the parameters $\epsilon = 0.1$, $\delta = 10^{-9}$ for all three algorithms, and in Algorithm 2, we use the squared exponential kernel

$$k(x, y) = \exp(-\|x - y\|^2/(2\ell)^2). \tag{26}$$

The values for m and ℓ used in experiments is listed in Table II. To select thresholds in Algorithms 2 and 3, we follow the advice of Remark 3, using $\eta = 0.15$ for Algorithm 2 and $\eta = \binom{n+2m}{n}/\epsilon$ for Algorithm 3. For Algorithm 1, we use an initial sample size of $20,000$ and a batch size of $5,000$ samples. For Algorithm 3, we use an initial sample size and batch size of $1,000$ samples.

To validate the accuracy bounds promised by Theorems 1 and 3 and Corollary 1, we performed an *a posteriori* analysis of the empirical error of the estimator constructed in each experiment. To calculate this empirical error, we simulated an additional $N_{ap} = 46{,}052$ samples, and took the empirical error $\hat{\epsilon}$ as the fraction of the new samples lying outside of the estimator. The number N_{ap} was chosen so that the empirical error estimate satisfies a *Chernoff bound* ensuring that, with confidence 0.9999, $\hat{\epsilon}$ differs from the true error by no more than 0.01. The computed empirical accuracies are displayed in Table II: since all are below 0.01, the *a posteriori* analysis ensures with high confidence that the guaranteed error bounds have been met.

A. Chaotic Nonlinear Oscillator

The first example is a reachable set estimation problem for the nonlinear, time-varying system with dynamics $\dot{z} = y, \dot{y} = -\alpha y + z - z^3 + \gamma\cos(\omega t)$, with states $x = (z, y) \in \mathbb{R}^2$ and parameters $\alpha, \gamma, \omega \in \mathbb{R}$. This system is known as the *Duffing oscillator*, a nonlinear oscillator which exhibits chaotic behavior for certain values of α, γ, and ω, for instance $\alpha = 0.05$, $\gamma = 0.4$, $\omega = 1.3$. The initial set is the interval such that $z(0) \in [0.95, 1.05]$, $y(0) \in [-0.05, 0.05]$, and we take X_0 to be uniform over this interval. The time range is $[t_0, t_1] = [0, 100]$.

We use Algorithms 1 and 3 to compute reachable set estimates using an order $k = 10$ empirical inverse Christoffel function with accuracy and confidence parameters $\epsilon = 0.10$, $\delta = 10^{-9}$. Additionally, we use Algorithm 2 to compute a kernelized empirical inverse Christoffel function using the squared exponential kernel $k(x, y) = \exp(\|x - y\|^2/(2\ell^2))$ with $\ell = 0.25$. Figure 1 shows the reachable set estimate for the Duffing oscillator system with the problem data given above produced by all three algorithms: for Algorithm 2, both the full kernelized Christoffel function estimator and its Nyström approximation with $r = 2000$. The cloud of points are the $11,000$ samples used in Algorithm 3. The reachable set estimate is neither convex nor simply connected, closely following the boundaries of the cloud of points and excluding an empty region. In particular, all estimates exhibit a hole in a region of the state space devoid of samples.

B. Planar Quadrotor

The next example is a reachable set estimation problem for horizontal position and altitude in a nonlinear model of the planar dynamics of a quadrotor used as an example in [23, 3]. The dynamics for this model are $\ddot{p_x} = u_1 K\sin(\theta), \ddot{p_h} = -g + u_1 L\cos(\theta), \ddot{\theta} = -d_0\theta - d_1\dot{\theta} + n_0 u_2$, where p_x and p_h denote the quadrotor's horizontal position and altitude in meters, respectively, and θ denotes its angular displacement (so that the quadrotor is level with the ground at $\theta = 0$) in radians. The system has 6 states, which we take to be x, h, θ, and their first derivatives. The two system inputs u_1 and u_2 (treated as disturbances for this example) represent the motor thrust and the desired angle, respectively. The parameter values used (following [3]) are $g = 9.81$, $L = 0.64$, $d_0 = 70$, $d_1 = 17$, and $n_0 = 55$. The set of initial states is the interval such that $p_x(0) \in [-1.7, 1.7]$, $\dot{p}_x(0) \in [-0.8, 0.8]$, $p_h(0) \in [0.3, 2.0]$, $\dot{p}_h(0) \in [-1.0, 1.0]$, $\theta(0) \in [-\pi/12, \pi/12]$, $\dot{\theta}(0) \in [-\pi/2, \pi/2]$, the set of inputs is the set of constant functions $u_1(t) = u_1$, $u_2(t) = u_2$ $\forall t \in [t_0, t_1]$, whose values lie in the interval $u_1 \in [-1.5 + g/L, 1.5 + g/L], u_2 \in [-\pi/4, \pi/4]$, and we take X_0 and D to be the uniform random variables defined over these intervals. The time range is $[t_0, t_1] = [0, 5]$. We take probabilistic parameters $\epsilon = 0.10$, $\delta = 10^{-9}$. Since the goal of this example is to estimate a reachable set for the horizontal position and altitude only, we are interested in a reachable set for a subset of the state variables, namely p_x and p_h. Following Remark 1, we use the reduced-state variations of Algorithms 1, 2, to compute reachable set estimates using only data for the (p_x, p_h) states, effectively reducing the dimension of the problem from 6 to 2. Figure 2 shows the reachable set estimate for the planar quadrotor system with the problem data given above produced by all three algorithms and the Nyström-approximated Algorithm 2 with $r = 2000$. The reachable set

Example	Alg. 1				Alg. 3				Alg. 2			
	m	time	N	$\hat{\epsilon}$	m	time	N	$\hat{\epsilon}$	ℓ	time	N	$\hat{\epsilon}$
Oscillator	10	60	70307	0	10	13	11000	0	1/4	506	30000	0
Quadrotor	4	3	14587	0	4	4	6000	10^{-3}	1/4	788	35000	0
Traffic	10	16	70307	0	10	11	10000	2×10^{-4}	1/4	504	30000	0

TABLE II

COMPUTATION TIMES, SAMPLE SIZES, PARAMETERS, AND *a posteriori* EMPIRICAL ERRORS FOR NUMERICAL EXPERIMENTS. ALL TIMES IN SECONDS. ALGORITHMS 1 AND 3 USED POLYNOMIAL ORDER $\boldsymbol{m}$, AND ALGORITHM 2 USED $\boldsymbol{k(x,y) = \exp(-\|x-y\|^2/(2\ell)^2)}$, WITH $\boldsymbol{m}$, $\boldsymbol{\ell}$ AS GIVEN IN THE TABLE. ALL EXPERIMENTS USE $\boldsymbol{\epsilon = 0.1}$, $\boldsymbol{\delta = 10^{-9}}$.

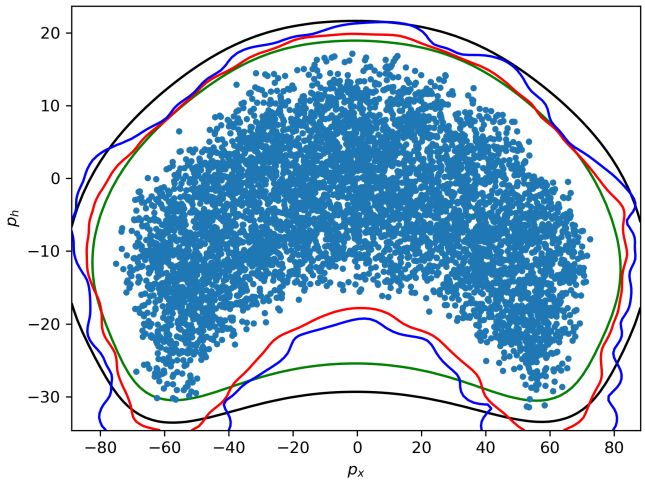

Fig. 2. Results of Algorithms 1, 2, and 3 on the planar quadrotor reachability problem, restricting the reachability analysis to the $(\boldsymbol{p_x}, \boldsymbol{p_h})$ plane. Black contour: output of Algorithm 1. Green contour: polynomial Christoffel function of order $\boldsymbol{k = 10}$. Blue contour: kernelized inverse Christoffel function with squared exponential kernel. Red contour: Nyström approximation ($\boldsymbol{m = 10,000}$) of the kernelized inverse Christoffel function with squared exponential kernel.

estimates displayed in Figure 2, and the computation times reported in Table II, use the reduced-state variation.

C. Monotone Traffic

This example is a special case of a continuous-time road traffic analysis problem used as a reachability benchmark in [5]. This problem investigates the density of traffic on a single lane over a time range over four periods of duration T using the Cell Transmission Model [7] that divides the road into n equal segments. The spatially discretized model is an n-dimensional dynamical system with states $x_1, \ldots, x_n$, where x_i represents the density of traffic in the i^{th} segment. Traffic enters segment through x_1 and flows through each successive segment before leaving through segment n. The system dynamics (27) are monotone, i.e. order-preserving: this property allows us to compute an interval containing the reachable set by evaluating the dynamics at the extreme points of the intervals defining the initial set and the set of disturbances. While this interval over-approximation is easy to compute, and is the best possible over-approximation by an interval, it is in general a conservative over-approximation because the reachable set may only occupy a small volume of the interval. Since the empirical Inverse Christoffel function method can accurately detect the geometry of the reachable set, we use this method to compare the shape of the reachable set to the best interval over-approximation.

The state dynamics are

$$\begin{aligned}
\dot{x}_1 &= \frac{1}{T}\left(d - \min(c, vx_1, w(\bar{x} - x_2))\right) \\
\dot{x}_i &= \frac{1}{T}\big(\min(c, vx_{i-1}, w(\bar{x} - x_i)) \\
&\quad - \min(c, vx_i, w(\bar{x} - x_{i+1}))\big), \quad (i = 2, \ldots, n-1) \\
\dot{x}_n &= \frac{1}{T}\left(\min(c, vx_{n-1}, w(\bar{x} - x_n)/\beta) - \min(c, vx_n)\right),
\end{aligned} \tag{27}$$

where v represents the free-flow speed of traffic, c the maximum flow between neighboring segments, $\bar{x}$ the maximum occupancy of a segment, and w the congestion wave speed. The input u represents the influx of traffic into the first node. For the reachable set estimation problem, we use a model with $n = 6$ states, and take $T = 30$, $v = 0.5$, $w = 1/6$, and $\bar{x} = 320$. The initial set is the interval such that $x_i(0) \in [100, 200]$, $i = 1, \ldots, n$, the set of disturbances is the set of constant disturbances with values in the range $d \in [40/T, 60/T]$, and X_0 and D are the uniform random variables over these sets. The time range is $[t_0, t_1] = [0, 4T]$.

We use the reduced-state variant of Algorithms 1, 2, and 3 to compute a reachable set for the traffic densities x_5 and x_6 at the end of the road, using an order $k = 10$ empirical inverse Christoffel function with accuracy and confidence parameters $\epsilon = 0.10$, $\delta = 10^{-9}$. Figure 3 compares the reachable set estimates for the traffic system produced by all three algorithms, and the Nyström-approximated Algorithm 2 with $r = 2000$, with the projection of the tight interval over-approximation computed using the monotonicity property of the traffic system. The figure indicates that the tight interval over-approximation of the reachable set is a somewhat conservative over-approximation, since the reachable set has approximately the shape of a parallelotope whose sides are not axis-aligned.

V. CONCLUSION

This paper advances the non-asymptotic theory of support set estimation by empirical Christoffel functions by applying the formal connection between Christoffel functions and Gaussian process regression models to a Bayesian PAC analysis of the estimator. The numerical examples demonstrate that the Bayesian PAC give a large improvement in sample efficiency

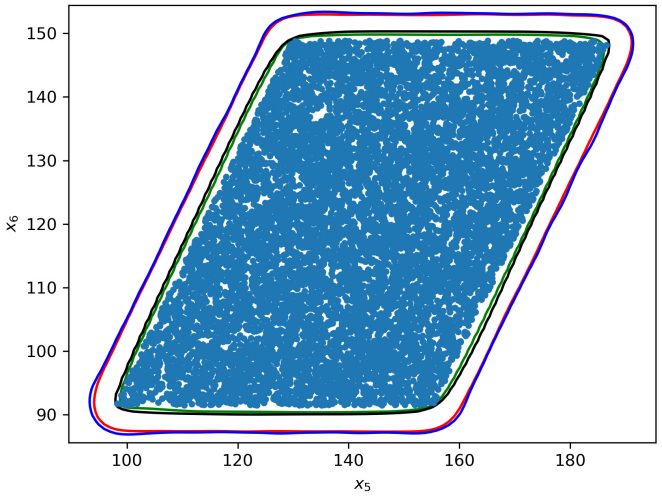

Fig. 3. Results of Algorithms 1, 2, and 3 on the six-state monotone traffic reachability problem, restricting the reachability analysis to the (x_5, x_6) plane. Black contour: output of Algorithm 1. Green contour: polynomial Christoffel function of order $k = 10$. Blue contour: kernelized inverse Christoffel function with squared exponential kernel. Red contour: Nyström approximation ($m = 10,000$) of the kernelized inverse Christoffel function with squared exponential kernel.

over classical PAC bounds. Additionally, Bayesian PAC arguments endow the kernelized inverse Christoffel function with PAC bounds, a development not possible with classical VC dimension bounds.

Improvements to the general theory can advance in step with advances in Bayesian PAC analysis. For instance, there are new results in theory of *derandomizing* Bayesian PAC bounds, which could offer sample efficiency improvements over the argument used in Lemma 5 to apply the Bayesian PAC bound to the central concept. Furthermore, domain-specific knowledge could be applied to the GP prior used to construct the Christoffel functions. For instance, in reachability problems and estimate of the system sensitivity matrix could be used to intelligently select length-scales in the kernel, along with other algorithm hyper-parameters such as the initial sample size and batch size.

There are also several numerical improvements to make in certain aspects of Algorithm 2, namely in the computation of the KL divergence. The implementation used in the examples of this paper requires the full kernel Gramian to be stored in memory to compute the KL divergence. For large datasets this becomes a prohibitive memory bottleneck. This bottleneck can be overcome with by computing the eigenvalues with a method that requires only an implicit representation, such as a Lanczos algorithm: an efficient implementation of this approach remains as future work.

REFERENCES

[1] Teodoro Alamo, Roberto Tempo, and Eduardo F Camacho. “Randomized strategies for probabilistic solutions of uncertain feasibility and optimization problems”. In: *IEEE Transactions on Automatic Control* 54.11 (2009), pp. 2545–2559.

[2] Armin Askari, Forest Yang, and Laurent El Ghaoui. “Kernel-based outlier detection using the inverse Christoffel function”. In: *arXiv preprint arXiv:1806.06775* (2018).

[3] Patrick Bouffard. “On-board Model Predictive Control of a Quadrotor Helicopter: Design, Implementation, and Experiments”. In: (2012). URL: http://www.eecs.berkeley.edu/Pubs/TechRpts/2012/EECS-2012-241.html.

[4] Frank M Callier and Charles A Desoer. *Linear system theory*. Springer Science & Business Media, 1991.

[5] Samuel Coogan and Murat Arcak. “A benchmark problem in transportation networks”. In: *arXiv preprint arXiv:1803.00367* (2018).

[6] Antonio Cuevas and Ricardo Fraiman. “A plug-in approach to support estimation”. In: *The Annals of Statistics* 25.6 (1997), pp. 2300–2312.

[7] Carlos F Daganzo. “The cell transmission model: A dynamic representation of highway traffic consistent with the hydrodynamic theory”. In: *Transportation Research Part B: Methodological* 28.4 (1994), pp. 269–287.

[8] Alex Devonport and Murat Arcak. “Data-driven reachable set computation using adaptive Gaussian process classification and Monte Carlo methods”. In: *2020 American Control Conference (ACC)*. IEEE. 2020, pp. 2629–2634.

[9] Alex Devonport and Murat Arcak. “Estimating Reachable Sets with Scenario Optimization”. In: ed. by Alexandre M. Bayen et al. Vol. 120. Proceedings of Machine Learning Research. PMLR, 2020, pp. 75–84.

[10] Alex Devonport et al. “Data-Driven Reachability Analysis with Christoffel Functions”. In: *2021 60th IEEE Conference on Decision and Control (CDC)*. IEEE. 2021, pp. 5067–5072.

[11] Franck Djeumou et al. “On-the-fly control of unknown smooth systems from limited data”. In: *arXiv preprint arXiv:2009.12733* (2020).

[12] Alexander N Dolia et al. “The minimum volume covering ellipsoid estimation in kernel-defined feature spaces”. In: *European Conference on Machine Learning*. Springer. 2006, pp. 630–637.

[13] Richard M Dudley. “Central limit theorems for empirical measures”. In: *The Annals of Probability* (1978), pp. 899–929.

[14] Chuchu Fan et al. “DryVR: data-driven verification and compositional reasoning for automotive systems”. In: *International Conference on Computer Aided Verification*. Springer. 2017, pp. 441–461.

[15] Lukas Hewing and Melanie N Zeilinger. “Scenario-based probabilistic reachable sets for recursively feasible stochastic model predictive control”. In: *IEEE Control Systems Letters* 4.2 (2019), pp. 450–455.

[16] Daniele Ioli et al. “A smart grid energy management problem for data-driven design with probabilistic reachability guarantees”. In: *4th International Workshop on Applied Verification of Continuous and Hybrid Systems*. Vol. 48. 2017, pp. 2–19.

[17] John Langford and Robert Schapire. “Tutorial on Practical Prediction Theory for Classification”. In: *Journal of Machine Learning Research* 6.3 (2005).

[18] John Langford and John Shawe-Taylor. “PAC-Bayes & margins”. In: *Advances in Neural Information Processing Systems* (2003), pp. 439–446.

[19] Jean B Lasserre and Edouard Pauwels. “The empirical Christoffel function in statistics and machine learning”. In: *arXiv preprint arXiv:1701.02886* (2017).

[20] Jean B Lasserre and Edouard Pauwels. “The empirical Christoffel function with applications in data analysis”. In: *Advances in Computational Mathematics* 45.3 (2019), pp. 1439–1468.

[21] Giuseppe Roberto Marseglia et al. “A hybrid stochastic-deterministic approach for active fault diagnosis using scenario optimization”. In: *IFAC Proceedings Volumes* 47.3 (2014), pp. 1102–1107.

[22] David A McAllester. “Some PAC-Bayesian theorems”. In: *Machine Learning* 37.3 (1999), pp. 355–363.

[23] Ian M Mitchell, Jacob Budzis, and Andriy Bolyachevets. “Invariant, viability and discriminating kernel under-approximation via zonotope scaling”. In: *Proceedings of the 22nd ACM International Conference on Hybrid Systems: Computation and Control*. 2019, pp. 268–269.

[24] Paul Nevai. “Géza Freud, orthogonal polynomials and Christoffel functions. A case study”. In: *Journal of approximation theory* 48.1 (1986), pp. 3–167.

[25] Edouard Pauwels and Jean-Bernard Lasserre. “Sorting out typicality with the inverse moment matrix SOS polynomial”. In: *Advances in Neural Information Processing Systems* 29 (2016), pp. 190–198.

[26] Edouard Pauwels, Mihai Putinar, and Jean-Bernard Lasserre. “Data analysis from empirical moments and the Christoffel function”. In: *Foundations of Computational Mathematics* 21.1 (2021), pp. 243–273.

[27] Bolun Qi et al. “DryVR 2.0: a tool for verification and controller synthesis of black-box cyber-physical systems”. In: *Proceedings of the 21st International Conference on Hybrid Systems: Computation and Control (part of CPS Week)*. 2018, pp. 269–270.

[28] Hossein Sartipizadeh et al. “Voronoi partition-based scenario reduction for fast sampling-based stochastic reachability computation of linear systems”. In: *2019 American Control Conference (ACC)*. IEEE. 2019, pp. 37–44.

[29] Matthias Seeger. “PAC-Bayesian generalisation error bounds for Gaussian process classification”. In: *Journal of Machine Learning Research* 3.Oct (2002), pp. 233–269.

[30] Dawei Sun and Sayan Mitra. “NeuReach: Learning Reachability Functions from Simulations”. In: *International Conference on Tools and Algorithms for the Construction and Analysis of Systems*. Springer. 2022, pp. 322–337.

[31] Charles F Van Loan and G Golub. *Matrix computations*. The Johns Hopkins University Press, 1996.

[32] Mathukumalli Vidyasagar. *Learning and Generalisation: With Applications to Neural Networks*. Springer Science & Business Media, 2002.

[33] Christopher Williams and Matthias Seeger. “Using the Nyström method to speed up kernel machines”. In: *Proceedings of the 14th Annual Conference on Neural Information Processing Systems*. 2001, pp. 682–688.

[34] Yuan Xu. “Christoffel functions and Fourier series for multivariate orthogonal polynomials”. In: *Journal of Approximation Theory* 82.2 (1995), pp. 205–239.

[35] Yang Yang et al. “Multi-aircraft conflict detection and resolution based on probabilistic reach sets”. In: *IEEE Transactions on Control Systems Technology* 25.1 (2016), pp. 309–316.

APPENDIX I
BACKGROUND ON GAUSSIAN PROCESS MODELS

A Gaussian process g is a stochastic process such that vectors $(g(x_1), \ldots, g(x_m))$ of point evaluations are multivariate Gaussian distributions. Similar to how a Gaussian random variable is completely characterized by its mean and variance, a Gaussian process is completely characterized by a mean function m, defined pointwise as $m(x) = \mathbb{E}[g(x)]$, and a positive semidefinite covariance function k, defined on all pairs of points $x, y \in \mathcal{X}$ as $k(x, y) = \mathbb{E}[g(x)g(y)]$.

Gaussian processes can also be defined according to a finite set of basis functions, admitting a direct construction as a finite weighted sum. For an m-dimensional space of functions with basis $b_1, \ldots, b_m : \mathcal{X} \to \mathbb{R}$, we form the stochastic weighted average $\sum_{i=1}^{m} w_i b_i$, where $w = (w_1, \ldots, w_m) \sim \mathcal{N}(0, \Sigma)$. This weighted average is a Gaussian process whose support is the span of $b_1, \ldots, b_m$, with mean $m(x) = 0$ and covariance $k(x, y) = \sum_{i=1}^{m} b(x)^\top \Sigma b(y)$, where $b(\cdot) = (b_1(\cdot), \ldots, b_m(\cdot))^\top$.

The Gaussian process regression model is Bayesian regression model that uses a Gaussian process as the prior over regression functions. In our case, we take the mean of the prior process to be zero. The data is assumed to be of the form $g(x_i) = h_i + \varepsilon$, where ε is a Gaussian noise term with variance σ^2. Under these conditions, the posterior for the unknown function is also a Gaussian process, whose mean and covariance are given by the formulas

$$m_q(x) = k_D(x)^\top \left(\sigma^2 I_N + K\right)^{-1} h, \tag{28}$$

$$k_q(x, y) = k(x, y) - k_D(x)^\top \left(\sigma^2 I_N + K\right)^{-1} k_D(y). \tag{29}$$

From the expression for the posterior covariance, we get the posterior variance

$$\begin{aligned} \mathrm{Var}_{g_q}(x) &= k_q(x, x) \\ &= k_p(x, x) \\ &\quad - k_D(x)^\top \left(\sigma^2 I_N + K\right)^{-1} k_D(X, x), \end{aligned} \tag{30}$$

which is precisely the kernelized empirical inverse Christoffel function with kernel k for the data $D = (x_1, \ldots, x_N)$ evaluated at the point x. In the finite-dimensional case, the

posterior process has mean and covariance functions

$$m_q(x) = \sigma^{-2} b(x)^\top \left(\Sigma^{-1} + \sigma^2 BB^\top\right)^{-1} By \tag{31}$$

$$k_q(x,y) = b(x)^\top (\Sigma^{-1} + \sigma^{-2} BB^\top)^{-1} b(x), \tag{32}$$

where $B \in \mathbb{R}m \times N$ is the matrix formed by evaluating the basis functions on the data, that is $B = [b(x_1) \quad \cdots \quad b(x_N)]$. Taking $b = z_k$, $\Sigma = \sigma_0^{-2} I$, $\sigma = N^{-1/2}$, yields the posterior variance $\mathrm{Var}_{g_q}(x) = z_m(x)^\top \left(\sigma_0^2 I + \frac{1}{N}\sum_{i=1}^N z_m(x_i) z_m(x_i)\top\right)^{-1} z_m(x)$, which is precisely the polynomial empirical inverse Christoffel function of order k for the data $x_1, \ldots, x_n$ evaluated at the point x.

APPENDIX II
PROOFS OF LEMMAS IN SECTION III-B

A. Proof of some Lemma 3

We consider the kernel case, since the polynomial case follows by the appropriate choice of kernel function. Recall that $\kappa^{-1}(x)$ is the variance of g_p by construction. Evaluating g_p at a single point x yields the normal random variable $g_p(x) \sim \mathcal{N}\left(0, \kappa^{-1}(x)\right)$. It follows that $g_p(x)/\sqrt{\kappa^{-1}(x)} \sim \mathcal{N}(0,1)$, and that $g_p(x)^2/\kappa^{-1}(x) \sim \chi_1^2$, that is that $g_p(x)^2/\kappa^{-1}(x)$, is a chi-square random variable with one degree of freedom. The average loss over C_P for a fixed point x is then

$$\begin{aligned}
\mathbb{E}\left[\ell(C_Q, x)\right] &= \mathbb{E}\left[\mathbb{1}\{x \in C_Q\}\right] \\
&= 1 - \mathbb{P}\left(g_p(x)^2 \leq \eta\right) \\
&= 1 - \mathbb{P}\left(\frac{g_p(x)^2}{\kappa^{-1}(x)} \leq \frac{\eta}{\kappa^{-1}(x)}\right) \\
&= 1 - F_1\left(\frac{\eta}{\kappa^{-1}(x)}\right).
\end{aligned} \tag{33}$$

Averaging this expression over the data points yields (15).

B. Proof of Lemma 6

We apply the Seeger PAC-Bayes Theorem 2 to the prior and posterior measures P and Q induced by C_P and C_Q as defined in (18). Recall that these prior and posterior measures are defined by the random vectors $W_P \sim \mathcal{N}\left(0, \sigma_0^{-2} I\right)$, $W_Q \sim \mathcal{N}\left(0, (\sigma_0^{-2} I + \hat{M}_{m,\sigma_0})^{-1}\right)$, which act as parameters. Applying this choice of W_p and W_q to equation (9) of Theorem 2 yields the inequality

$$P_X^N\left(\{x_1, \ldots, x_N : D_{\text{ber}}(\hat{r}_Q || r_Q) \leq \gamma\}\right) \geq 1 - \delta, \tag{34}$$

where

$$\gamma = \frac{D_{KL}(\mathcal{N}\left(0, (\sigma_0^2 I + \hat{M}_{m,\sigma_0})^{-1}\right) || \mathcal{N}\left(0, \sigma_0^{-2} I\right)) + \log\frac{N+1}{\delta}}{N}$$

Suppose the data set $x_1, \ldots, x_N$ is one such that the inner inequality $D_{\text{ber}}(\hat{r}_Q || r_Q) \leq \gamma$ holds: then r_Q, the true stochastic risk, lies in the set $\{\beta : D_{ber}(\hat{r}_Q || \beta) \leq \gamma\}$. The function $D_{ber}(\hat{r}_Q || \beta)$ is convex in β and covers the range $[0, \infty)$, attaining 0 for $\beta = \hat{r}_Q$ and approaching ∞ for $\beta \to 0$ and $\beta \to 1$. By these properties, $\{\beta : D_{ber}(\hat{r}_Q || \beta) \leq \gamma\}$ is a closed convex subset of $(0,1)$ for any positive γ. As such, it attains a supremum, meaning that $\overline{r}$ as defined in (20) is well-defined. Thus we have, with confidence $1 - \delta$, that $\overline{r}$ is an upper bound on the stochastic risk r_Q.

C. Proof of Lemma 4

As in the proof of Lemma 6 we apply the Seeger PAC-Bayes Theorem 2, this time to the prior and posterior measures P and Q induced by C_P and C_Q as defined in (12). These measures are defined by the Gaussian processes g_p and g_q which act as the concept class parameters W_P and W_Q respectively in the statement of Theorem 2. To compute the KL divergence between W_P and W_Q, we use another result due to Seeger, described in Section 2.2 of [29], which states that the KL divergence between a prior Gaussian process g_p and the posterior Gaussian processes g_q obtained after conditioning on data $x_1, \ldots, x_N$ is equal to the KL divergence between the restriction of the two Gaussian processes to the data points, that is the KL divergence between the multivariate normal random vectors $(g_p(x_1), \ldots, g_p(x_N))$ and $(g_q(x_1), \ldots, g_q(x_N))$. The mean and covariance of these random variables are simply the restrictions of the mean and covariance functions of their defining processes to $(x_1, \ldots, x_N)$. Both random vectors have mean zero. The covariance matrix of the prior random vector $(g_p(x_1), \ldots, g_p(x_N))$ is $K_p(X,X) = K(X,X)$ as discussed in Section I. By (28) and an application of the matrix inversion lemma, the covariance of the posterior random vector $(g_q(x_1), \ldots, g_q(x_N))$ is

$$\begin{aligned}
K_q(X,X) &= K(X,X) \\
&\quad - K(X,X)\left(\sigma_0^2 I + K(X,X)\right)^{-1} K(X,X) \\
&= \left(K(X,X)^{-1} + \sigma_0^{-2} I\right)^{-1}.
\end{aligned} \tag{35}$$

D. Proof of Lemma 5

Consider a point $x \in \mathcal{X}$ outside of the central concept, that is such that $\bar{c}_\eta(x) = \mathbb{E}\left[(g(x)^2\right] > \eta$. The probability that $W_Q^\top z_m(x)$ also exceeds η is bounded as

$$\mathbb{P}\left((g(x)^2 \geq \eta\right) \leq \mathbb{P}\left((g(x)^2 \geq \mathbb{E}\left[(g(x)^2\right]\right) \tag{36}$$

$$= \mathbb{P}\left(\frac{(g(x)^2}{\mathbb{E}\left[(g(x)^2\right]} > 1\right) \tag{37}$$

$$= 1 - F_1(1). \tag{38}$$

Next, let us consider the risk of the stochastic estimator, that is $r_Q = \mathbb{P}\left((g(X)^2 > \eta\right)$. Applying the law of total probability with respect to the random variable X, we divide r_Q into two integrals according to whether the central concept exceeds η:

$$\mathbb{P}\left((g(X)^2 > \eta\right) = \int_{\mathcal{X}} \mathbb{P}\left((g(x)^2 > \eta\right) dP_x(x) \tag{39}$$

$$= \int_{\mathcal{X}} \mathbb{P}\left((g(x)^2 > \eta\right) \mathbb{1}\{\mathbb{E}\left[(g(x)^2\right] > \eta\} dP_x(x) \tag{40}$$

$$+ \int_{\mathcal{X}} \mathbb{P}\left((g(x)^2 > \eta\right) \mathbb{1}\{\mathbb{E}\left[(g(x)^2\right] \leq \eta\} dP_x(x). \tag{41}$$

We have that $\mathbb{P}\left((g(X)^2 > \eta\right) \geq \int_{\mathcal{X}} \mathbb{P}\left((g(x)^2 > \eta\right) \mathbb{1}\{\mathbb{E}\left[(g(x)^2\right] > \eta\} dP_x(x)$, since all three integrands are nonnegative. To find an upper bound on this probability in terms of the empirical classifier, we

combine the two inequalities above to find

$$\mathbb{P}\left((g(X)^2 > \eta\right) = \int_{\mathcal{X}} \mathbb{P}\left((g(x)^2 > \eta\right) dP_x(x) \tag{42}$$

$$\geq \int_{\mathcal{X}} \mathbb{P}\left((g(x)^2 > \eta\right) \mathbb{1}\{\mathbb{E}\left[(g(x)^2\right] > \eta\} dP_x(x) \tag{43}$$

$$\geq (1 - F_1(1)) \int_{\mathcal{X}} \mathbb{1}\{\mathbb{E}\left[(g(x)^2\right] > \eta\} dP_x(x) \tag{44}$$

$$= (1 - F_1(1))\mathbb{P}\left(\mathbb{E}\left[(g(x)^2\right] > \eta\right) \tag{45}$$

$$= (1 - F_1(1)) r(\hat{c}_\eta), \tag{46}$$

which we rearrange to yield $r(\bar{c}_\eta) \leq \frac{1}{1-F_1(1)} r_{Q_\eta}$.

Alex Devonport was born in Phoenix, AZ, USA, in 1993. He received the B.S. degree in electrical engineering from Arizona State University, Tempe, AZ, USA in 2016, and is pursuing the Ph.D. degree in electrical engineering at the University of California, Berkeley, USA.

In 2013 he worked as an engineering assistant at Summit Green manufacturing, Tempe, AZ, USA. From 2014 to 2016 he worked as a research assistant in the Goldwater Engineering center at Arizona State university, Tempe, AZ, USA, and from 2016 to 2017 as a research staff member. In 2019 he was a research intern at the Technical University of Munich, Munich, Germany. His research interests include data-driven methods in control theory, particularly for reachability analysis and abstraction-based controller design, with a focus on establishing probabilistic correctness guarantees.

Forest Yang was born in Riverside, CA, USA in 1996. He received the B.S. degree in electrical engineering and computer science in 2018 from the University of California, Berkeley, pursued a Ph.D. in computer science from 2018-2019 at the University of Illinois, Urbana-Champaign, and is currently pursuing a Ph.D. in computer science at the University of California, Berkeley.

From 2018 to 2019 he was a Machine Learning Engineer for SumUp Analytics, and in 2019 was a Research Intern at Google Brain, Accra, Ghana. He published On the Consistency of Top-k Surrogate Losses at ICML2020 and Fairness with Overlapping Groups at NeurIPS2020.

Laurent El Ghaoui is the Dean of the College of Engineering and Computer Sciences at Vin University in Hanoi, Vietnam. Prior to that, he was a professor at U.C. Berkeley in the Electrical Engineering and Computer Sciences Department, with a courtesy appointment in the Industrial Engineering and Operations Research department, and a teaching appointment in the Financial Engineering program at the Haas business school. Laurent graduated from Ecole Polytechnique (Palaiseau, France) in 1985, and obtained his Ph.D. in Aeronautics and Astronautics at Stanford University in March 1990. He was a faculty member of the Ecole Nationale Supérieure de Techniques Avancées (Paris, France) from 1992 until 1999, and held part-time teaching appointments at Ecole Polytechnique within the Applied Mathematics department and Université de Paris-I (La Sorbonne) in the Mathematics in Economy program.

Laurent's research is in robust optimization and sparse machine learning, with a recent interest in so-called implicit models in deep learning. He is the recipient of a Bronze Medal for Engineering Sciences, from the Centre National de la Recherche Scientifique (France), a CAREER award, and an Okawa research grant. He is also the co-recipient of a SIAM optimization prize.

Murat Arcak is a professor at U.C. Berkeley in the Electrical Engineering and Computer Sciences Department, with a courtesy appointment in Mechanical Engineering. He received the B.S. degree in Electrical Engineering from the Bogazici University, Istanbul, Turkey (1996) and the M.S. and Ph.D. degrees from the University of California, Santa Barbara (1997 and 2000). His research is in dynamical systems and control theory with applications to synthetic biology, multi-agent systems, and transportation.

Prior to joining Berkeley in 2008, he was a faculty member at the Rensselaer Polytechnic Institute. He received a CAREER Award from the National Science Foundation in 2003, the Donald P. Eckman Award from the American Automatic Control Council in 2006, the Control and Systems Theory Prize from the Society for Industrial and Applied Mathematics (SIAM) in 2007, and the Antonio Ruberti Young Researcher Prize from the IEEE Control Systems Society in 2014. He is a member of ACM and SIAM, and a fellow of IEEE and the International Federation of Automatic Control (IFAC).