Safe Exploration for Active Learning with Gaussian Processes

Jens Schreiter[1(✉)], Duy Nguyen-Tuong[1], Mona Eberts[1], Bastian Bischoff[1], Heiner Markert[1], and Marc Toussaint[2]

[1] Robert Bosch GmbH, 70442 Stuttgart, Germany
jens.schreiter@de.bosch.com
[2] University of Stuttgart, MLR Laboratory, 70569 Stuttgart, Germany

Abstract. In this paper, the problem of safe exploration in the active learning context is considered. Safe exploration is especially important for data sampling from technical and industrial systems, e.g. combustion engines and gas turbines, where critical and unsafe measurements need to be avoided. The objective is to learn data-based regression models from such technical systems using a limited budget of measured, i.e. labelled, points while ensuring that critical regions of the considered systems are avoided during measurements. We propose an approach for learning such models and exploring new data regions based on Gaussian processes (GP's). In particular, we employ a problem specific GP classifier to identify safe and unsafe regions, while using a differential entropy criterion for exploring relevant data regions. A theoretical analysis is shown for the proposed algorithm, where we provide an upper bound for the probability of failure. To demonstrate the efficiency and robustness of our safe exploration scheme in the active learning setting, we test the approach on a policy exploration task for the inverse pendulum hold up problem.

1 Introduction

Active learning (AL) deals with the problem of selective and guided generation of labeled data. In the AL setting, an agent guides the data generation process by choosing new informative samples to be labeled based on the knowledge obtained so far. Providing labels for new data points, e.g. image labels as by Lang and Baum [1992] or measurements of the system output in case of physical systems, like by Hans et al. [2008], can be very costly and tedious. The overall goal of AL is to create a data-based model, without having to supply more data than necessary and, thus, reducing the agent annotation effort or the measurements on machines. For regression tasks, the AL concept is sometimes also referred to optimal experimental design, see Fedorov [1972].

In this paper, we consider the problem of safe data selection while jointly learning a data-based regression model on the explored input space. Given failure conditions, the goal is to actively select a budget of measurement points for approximating the model, and keeping the probability of measurement failures

A. Bifet et al. (Eds.): ECML PKDD 2015, Part III, LNAI 9286, pp. 133–149, 2015.
DOI: 10.1007/978-3-s319-23461-8_9

to a minimum at the same time. In practice, safe data selection is highly relevant, especially, when measurements are performed on technical systems, e.g. combustion engines and test benches. For such technical systems, it is important to avoid critical points, where the measurements can damage the system. Thus, the main objective is (i) to approximate the system model from sampled data, (ii) using a limited budget of measured points, and (iii) ensuring that critical regions of the considered system are avoided during measurements.

We consider active data selection problems from systems with compact input spaces $\mathbb{X} \subset \mathbb{R}^d$. Within this constrained area $\mathbb{X}$ we have regions, where sampling and measuring is undesirable and can damage the system. For technical systems, for example, operation of the system in specific regions can result in exceeding the allowed physical limits such as temperatures and pressures. If the agent chooses a sample in such a region, it is considered as failure. To anticipate a failure, we assume that the agent observes some feedback from the system for each data point. This feedback indicates the health status of the system. In case of the combustion engine, this feedback is given by the engine temperature, for example.

The safety of an actively exploring algorithm is defined over the probability of failure, i.e. we call an exploration scheme safe at the level of $1 - \delta$, if the failure probability when querying an instance is lower than a certain threshold δ. The user can define δ sufficiently small to achieve an acceptable risk of failure. However, reducing this probability of failure comes at the cost of decreased sample efficiency, i.e. more samples will be required, as the agent will take smaller steps and explore more carefully. Throughout the paper, we use the notations $\mathbb{X}_+$ for safe and $\mathbb{X}_-$ for unsafe regions of the confined input space $\mathbb{X}$ to distinguish between safe and hazardous input areas of the system.

In this work, we employ Gaussian processes (GP's) to learn the regression model from a limited budget of incrementally sampled data points. Our exploration strategy for determining the next query points is based on the differential entropy criterion, cf. Krause and Guestrin [2007]. Furthermore, we employ a problem specific GP classifier to identify safe and unsafe regions in $\mathbb{X}$. The basic idea of this discriminative GP is learning a decision boundary between two classes, preferably without sampling a point in the unsafe region $\mathbb{X}_-$. To the best of our knowledge, such a safe exploring scheme has never been considered before in the AL context. We further show a theoretical analysis of our proposed safe exploration scheme with respect to the AL framework.

The remainder of the paper is organized as follows. In Section 2, a brief overview on some existing work on safe exploration approaches is given. In Section 3, we introduce our setting first and, subsequently, describe the proposed algorithm while providing details about our exploration strategy and the employed safety constraint. Section 4 provides some theoretical results on our proposed safe exploration technique. Experiments on a toy example and on learning a control policy for the inverse pendulum are shown in Section 5. A discussion in Section 6 concludes the paper.

2 Related Work

Most existing work for safe exploration in unknown environments arose in the reinforcement learning setting. The strategy of Moldovan and Abbeel [2012] for safe exploration in finite Markov decision processes (MDP's) relies on the restriction of suitable policies which ensure ergodicity at a user-defined safety level, i.e. there exists a policy with high probability to get back to the initial state. In the risk-sensitive reinforcement learning approach by Geibel [2001], the ergodic assumption for MDP's is dropped by introducing fatal absorbing states. The risk of a policy is thereby defined over the probability for ending in a fatal state. Therefore, the authors present a model-free reinforcement learning algorithm which is able to find near-optimal policies under bounded risk. The work by Gillula and Tomlin [2011] provides safety guarantees via a reachability analysis, when an autonomous robot explores its environment online. Here, the robot is observed by an aerial vehicle which avoid it from taking unsafe actions in the state space. This hybrid control system assumes bounded actions and disturbances to ensure a safe behavior for all current observable situations. Instead of bounding the action space, Polo and Rebollo [2011] introduce learning from demonstrations for dynamic control tasks. To safely explore the continuous state space in their reinforcement learning setting, a previously defined safe policy is iteratively adapted from the demonstration with small additive Gaussian noise. This approach ensures a baseline policy behavior which is used for safe exploration of high-risk tasks, e.g. hovering control of a helicopter. Galichet et al. [2013] consider a multi-armed risk-aware bandit setting to prevent hazards when exploring different tasks, e.g. energy management problems. They introduce a reward-based framework to limit the exploration of dangerous arms, i.e. with a negative exploration bonus for risky actions. However, their approach is highly dependent on the designed reward function which has significant impact on the probability for damaging the considered system. Similarly, Hans et al. [2008] define a reward-based safety function to assess each state of the MDP and assume that there exists a safe return policy to leave critical states with non-fatal actions. Although this assumption may not hold generally, it allows the usage of dynamic programming to solve an adapted Bellman optimality equation to get a return policy.

Strategies for exploring unknown environments without considering safety issues has also been reflected in the framework of global optimization with GP's. For example, Guestrin et al. [2005] propose an efficient submodular exploration criterion for near-optimal sensor placements, i.e. for discrete input spaces. In Auer [2002], a framework is presented which yields a compromise between exploration and exploitation through confidence bounds. Srinivas et al. [2012] show, that under reasonable assumptions strong exploration guaranties can be given for Bayesian optimization with GP's. Due to the fact that the exploration tasks may lead to NP-hard problems, cf. Guestrin et al. [2005], the additional introduction of safety will increase the complexity which must be handled by the AL scheme on a higher level.

3 Safe Exploration for Active Learning

In this section, we introduce our safe exploration approach for active learning in detail. The basic idea is to jointly learn a discriminative function during exploration of the input space. Using the sampled data, we incrementally learn a model of the system. We employ Gaussian processes to learn the model, as well as to build the discriminative function. The overall goal is to learn an accurate model of the system, while avoiding data sampling from unsafe regions indicated by the discriminative function as much as possible. We assume that additional information is available, when our exploration scheme is getting close to the decision boundary, cf. Hans et al. [2008]. For real-world applications in combustion engine measurement, for example, the engine temperature is an indicator for the proximity to the decision boundary. Thus, it is possible to design a discriminative function for recognizing whether our active learner is getting close to unsafe input locations. For the safe region of the input space, we assume that it is compact and connected. In Figure 1, we illustrate the described setting for the input space $\mathbb{X} \subset \mathbb{R}^2$. Next, we give an introduction to Gaussian processes and proceed by describing our exploration scheme and defining the safety constraint. The derived algorithm will be summarized in Section 3.4.

3.1 Gaussian Processes

In this paper, we use GP's in the regression and classification context to design a safe active learning scheme. A GP is a collection of random variables, where any finite subset of them has a joint multivariate normal distribution. In the following, we adopt the notations by Rasmussen and Williams [2006]. For both, the regression and the classification case, we use a centered Gaussian prior given by $p(\cdot \mid \boldsymbol{X}) = \mathcal{N}(\cdot \mid \mathbf{0}, \boldsymbol{K})$, where the matrix $\boldsymbol{X}$ contains row-wise all associated input points $\boldsymbol{x}_i \in \mathbb{R}^d$ which induce the covariance matrix $\boldsymbol{K}$. The dot $\cdot$ acts as

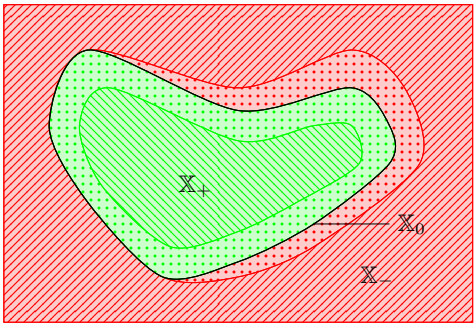

Fig. 1. Partition of the input space $\mathbb{X}$ into a safe explorable area $\mathbb{X}_+$ and an unsafe region $\mathbb{X}_-$ separated by the unknown decision boundary $\mathbb{X}_0$. Over the dotted area, a discriminative function is learned for recognizing whether the exploration becomes risky.

placeholder to distinguish between the latent target values f and the discriminative function g for the safety constraint. Throughout the paper, we use the stationary squared exponential covariance function

$$k(\boldsymbol{x}, \boldsymbol{z}, \boldsymbol{\theta}_{\cdot}) = \sigma_{\cdot}^2 \exp\left(-\tfrac{1}{2}(\boldsymbol{x}-\boldsymbol{z})^T \boldsymbol{\Lambda}_{\cdot}^{-2}(\boldsymbol{x}-\boldsymbol{z})\right), \tag{1}$$

where the signal magnitude $\sigma_{\cdot}^2$ and the diagonal matrix $\boldsymbol{\Lambda}_{\cdot} \in \mathbb{R}^{d\times d}$ of length-scales are summarized in the set $\boldsymbol{\theta}_{\cdot}$ of hyperparameters. We assume for both GP's that the hyperparameters are previously given. Here, recall from Krause and Guestrin [2007] that determining the hyperparameters in advance is similar to defining a grid over the whole input space $\mathbb{X} \subset \mathbb{R}^d$.

In the following, the cases for regression and classification are considered separately. For regression, the m data points are row-wise composed in $\boldsymbol{X} \in \mathbb{R}^{m\times d}$. The likelihood $p(\boldsymbol{y} \mid \boldsymbol{f}, \boldsymbol{X})$ is given through the model $y_i = f(\boldsymbol{x}_i) + \varepsilon_i$ with centered Gaussian noise $\varepsilon_i \sim \mathcal{N}(0, \sigma^2)$. The m inputs $\boldsymbol{X}$ and their associated targets $\boldsymbol{y} \in \mathbb{R}^m$ are summarized in $\mathcal{D}_m = (\boldsymbol{y}, \boldsymbol{X})$. Due to the fact that the likelihood in the regression case is Gaussian, exact inference is possible which yields a Gaussian posterior. The marginal likelihood of the regression model given the data then satisfies

$$p\left(\boldsymbol{y} \,\middle|\, \boldsymbol{X}\right) = \mathcal{N}\left(\boldsymbol{y} \,\middle|\, \boldsymbol{0}, \boldsymbol{K} + \sigma^2 \boldsymbol{I}\right), \tag{2}$$

cf. Rasmussen and Williams [2006]. For the predictive distribution of test points $\boldsymbol{x}_*$, we obtain

$$p\left(y_* \middle|\, \boldsymbol{x}_*, \mathcal{D}_m\right) = \mathcal{N}\left(y_* \,\middle|\, \boldsymbol{k}_*^T \boldsymbol{\alpha}, k_{**} - \boldsymbol{k}_*^T \left(\boldsymbol{K} + \sigma^2 \boldsymbol{I}\right)^{-1} \boldsymbol{k}_* + \sigma^2\right), \tag{3}$$

where $k_{**} = k(\boldsymbol{x}_*, \boldsymbol{x}_*, \boldsymbol{\theta}_f) = \sigma_f^2$ is the covariance value, $\boldsymbol{k}_* \in \mathbb{R}^m$ is the corresponding covariance vector induced by the test point and the training points, and with the prediction vector $\boldsymbol{\alpha} = \left(\boldsymbol{K} + \sigma^2 \boldsymbol{I}\right)^{-1} \boldsymbol{y} \in \mathbb{R}^m$.

To distinguish between safe and unsafe regions in our active learning framework, we use GP classification. Here, the latent discriminative function $g\colon \mathbb{X} \to \mathbb{R}$ is learned and, subsequently, mapped to the unit interval to describe the class likelihood for each point. This mapping is realized by means of a point symmetric sigmoid function, where we use the cumulative Gaussian $\Phi(\cdot)$ throughout the paper. For the class affiliation probability, we thus obtain $\Pr[c = +1 \mid \boldsymbol{x}] = \Phi\big(g(\boldsymbol{x})\big)$. Given the discriminative function g, the class labels $c_i \in \{-1, +1\}$, $i \in \{1, \ldots, k\}$, of the data points $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_k$ are independent random variables, so that the likelihood factorizes over the data points. This results in the non-Gaussian likelihood

$$p\left(\boldsymbol{c} \,|\, \boldsymbol{g}, \boldsymbol{X}\right) = \prod_{i=1}^{k} \Phi\big(c_i \, g(\boldsymbol{x}_i)\big), \tag{4}$$

where the matrix $\boldsymbol{X} \in \mathbb{R}^{k\times d}$ is row-wise composed of the input points. Due to the non-Gaussian likelihood (4), the posterior is not analytically tractable for this

probit model. To solve this problem, we use the efficient and accurate Laplace approximation to calculate an approximate Gaussian posterior $q\left(\boldsymbol{g} \mid \boldsymbol{c}, \boldsymbol{X}\right)$ similar to Nickisch and Rasmussen [2008]. The approximate posterior distribution forms the basis for the subsequent prediction of class affiliations for test points. Note that our classification model is an extended variant of the general classification task and will be explained in more detail in Section 3.3. For more details regarding GP classification, we refer to Rasmussen and Williams [2006].

3.2 Exploration Strategy

The exploration strategy applied in this paper is a selective sampling approach based on the posterior entropy of the GP model for the functional relationship $f(\boldsymbol{x})$. This entropy criterion has been frequently used in the active learning literature, e.g. in Seo et al. [2000] or Guestrin et al. [2005]. To the best of our knowledge, safe exploration in combination with this criterion has not been considered before. The main goal for this uncertainty sampling task, cf. Settles [2010], is to find an optimal set of feasible data points $\boldsymbol{X}_{\text{opt}}$ of given size m and determined hyperparameters $\boldsymbol{\theta}_f$ such that the differential entropy of the model evidence (2) is maximal, i.e.

$$\boldsymbol{X}_{\text{opt}} = \underset{\boldsymbol{X} \subset \mathbb{X},\ |\boldsymbol{X}|=m}{\arg\max}\ \mathrm{H}\left[\, \boldsymbol{y} \,\middle|\, \boldsymbol{X} \,\right]. \tag{5}$$

Here, the differential entropy of the evidence (2) depends on the determinant of the associated covariance matrix, i.e.

$$\mathrm{H}\left[\, \boldsymbol{y} \,\middle|\, \boldsymbol{X} \,\right] = \frac{1}{2} \log \left| 2\pi e \left(\boldsymbol{K} + \sigma^2 \boldsymbol{I} \right) \right|, \tag{6}$$

see e.g. Cover and Thomas [2006]. Note that Ko et al. [1995] showed that the problem of finding a finite set $\boldsymbol{X}_{\text{opt}}$ is NP-hard. Nevertheless, the following lemma summarizes some nice results about the differential entropy in our GP setting.

Lemma 1. *Let $\mathcal{D}_m$ be a non-empty data set and $\sigma^2 \geq (2\pi e)^{-1}$. Then, for some stationary covariance function with magnitude σ_f^2, the differential entropy (6) of the GP evidence (2) is a non-negative, monotonically increasing, and submodular function.*

The detailed proof of the latter lemma uses known results from Nemhauser et al. [1978] and Cover and Thomas [2006] and is given in the supplemental material. In Lemma 1 the lower bound for the noise σ^2 can be easily achieved by additionally scaling the target values $\boldsymbol{y}$ and the magnitude σ_f^2 with some suitable constant, cf. the proof of Lemma 1. The properties of the entropy induced by Lemma 1 guarantee that the greedy selection scheme

$$\boldsymbol{x}_{i+1} = \underset{\boldsymbol{x}_* \in \mathbb{X}}{\arg\max}\ \mathrm{H}\left[\, y_* \,\middle|\, \boldsymbol{x}_*, \mathcal{D}_i \right] \tag{7}$$

for our agent yields a nearly optimal subset of input points, where the distribution of y_* results from (3). More precisely, as shown by Nemhauser et al. [1978],

it holds true that the greedy selection scheme (7) yields a model entropy which is greater than 63% of the optimal entropy value induced by $\boldsymbol{X}_{\text{opt}}$. This guarantee induces an efficient greedy algorithm and, with some foresight to the introduction of safety in Section 3.3, that it will generally not be possible to design an optimal and fast safe active learning algorithm, cf. Moldovan and Abbeel [2012].

As empirically shown in Ramakrishnan et al. [2005], the described entropy based sampling strategy tends to select input locations close to the border and induces a point set $\boldsymbol{X}$ which is nearly uniformly distributed in the confined input space $\mathbb{X}$. This is a favorable behavior for modeling with GP's, which additionally enables us to slightly extrapolate the borders of the technical system with our GP model.

For the visualization of the exploration process, we use theoretical results by Cover and Thomas [2006] and the bounds

$$\tfrac{m}{2}\log\bigl(2\pi e\sigma^2\bigr) \leq \mathrm{H}\bigl[\,\boldsymbol{y}\,\big|\,\boldsymbol{X}\,\bigr] \leq \tfrac{m}{2}\log\bigl(2\pi e(\sigma_f^2+\sigma^2)\bigr)\,,$$

derived in the proof of Lemma 1, to normalize the gain in the differential entropy (6). After sampling the m-th query point, we thus define the normalized entropy ratio NER by

$$NER(\boldsymbol{X}) = \frac{\frac{2}{m}\,\mathrm{H}\bigl[\,\boldsymbol{y}\,\big|\,\boldsymbol{X}\,\bigr] - \log\bigl(2\pi e\sigma^2\bigr)}{\log\Bigl(1+\frac{\sigma_f^2}{\sigma^2}\Bigr)}\,. \tag{8}$$

If this ratio is close to one when $\boldsymbol{X}$ enlarges during the exploration process, we gain nearly maximal entropy for the selected queries. Otherwise, if $NER \approx 0$, the current query $\boldsymbol{x}_m$ does not explore the input space very much.

3.3 Safety Constraint

Our approach to introduce safety is based on additional information to describe the discriminative function g, when getting close to the decision boundary. In practice, without any feedback from the physical system or some user-defined knowledge, it is not possible to explore the environment safely, cf. Valiant [1984]. We encode this additional feedback in a possibly noisy function $h\colon \mathbb{X} \to (-1,1)$ to train the discriminative function g around the decision boundary. To define a likelihood for this heterogeneous model, we assume that sampling from the system to get values $h_j = h(\boldsymbol{x}_j)$ or labels $c_i = c(\boldsymbol{x}_i)$ leads to consistent data. That is, depending on the location of the data point (cf. Figure 1), we obtain either labels or discriminative function values. For $\boldsymbol{c} \in \mathbb{R}^k$, $\boldsymbol{h} \in \mathbb{R}^l$, $\boldsymbol{g} \in \mathbb{R}^n$, and $k+l=n$, the model likelihood results in

$$p\,(\boldsymbol{c},\boldsymbol{h}\,|\,\boldsymbol{g},\boldsymbol{X}\,) = \prod_{j=1}^{l}\mathcal{N}\bigl(h_j\big|\,g_j,\tau^2\bigr)\prod_{i=1}^{k}\Phi\,(c_i\,g_i)\;, \tag{9}$$

where we used (4) and a Gaussian regression model for $\boldsymbol{h}$ with noise variance τ^2. Employing the Gaussian prior for all $g_i = g(\boldsymbol{x}_i)$ and the Laplace approximation,

we get the Gaussian posterior approximation $q(\boldsymbol{g} \mid \boldsymbol{c}, \boldsymbol{h}, \boldsymbol{X})$ with respect to the exact posterior

$$p(\boldsymbol{g} \mid \boldsymbol{c}, \boldsymbol{h}, \boldsymbol{X}) \approx q(\boldsymbol{g} \mid \boldsymbol{c}, \boldsymbol{h}, \boldsymbol{X}) = \mathcal{N}(\boldsymbol{g} \mid \boldsymbol{\mu}, \boldsymbol{\Sigma}) \,, \tag{10}$$

where $\boldsymbol{\mu} = \arg\max_{\boldsymbol{g}} \big(p(\boldsymbol{g} \mid \boldsymbol{c}, \boldsymbol{h}, \boldsymbol{X})\big)$ and $\boldsymbol{\Sigma} = \left(\boldsymbol{W} + \boldsymbol{K}^{-1}\right)^{-1}$. Here, the diagonal Matrix $\boldsymbol{W}$ is given by

$$\boldsymbol{W} = -\frac{\partial^2}{\partial \boldsymbol{g} \partial \boldsymbol{g}^T} \log\left(p(\boldsymbol{c}, \boldsymbol{h} \mid \boldsymbol{g}, \boldsymbol{X})\right)\Big|_{\boldsymbol{g}=\boldsymbol{\mu}} .$$

For the predictive distribution of test cases, we then obtain

$$q(g_* \mid \boldsymbol{x}_*, \boldsymbol{c}, \boldsymbol{h}, \boldsymbol{X}) = \mathcal{N}\big(g_* \big| \mu_{g_*}, \sigma^2_{g_*}\big) \tag{11}$$

through the efficient Laplace approximation by Nickisch and Rasmussen [2008], where $\mu_{g_*} = \boldsymbol{k}_*^T \boldsymbol{K}^{-1} \boldsymbol{\mu}$, $\sigma^2_{g_*} = k_{**} - \boldsymbol{k}_*^T \boldsymbol{W}^{\frac{1}{2}} \boldsymbol{B}^{-1} \boldsymbol{W}^{\frac{1}{2}} \boldsymbol{k}_*$ and $\boldsymbol{B} = \boldsymbol{I} + \boldsymbol{W}^{\frac{1}{2}} \boldsymbol{K} \boldsymbol{W}^{\frac{1}{2}}$. The safety constraint in our AL approach should ensure that the probability of making a failure is small, e.g. less than $1 - p$ for some $p \in (0, 1)$. Formally, our safety constraint is thus given by

$$\Pr[\, g_* \geq 0 \,|\, \boldsymbol{x}_*, \boldsymbol{c}, \boldsymbol{h}, \boldsymbol{X}] \geq p \,. \tag{12}$$

For a successful safe exploration, the function h must fulfill some conditions, which will be presented in Section 4.

3.4 Safe Active Learning: The Algorithm

In this subsection, we present our entropy based active learning framework extended by a constraint inducing safety, cf. Algorithm 1. Since the entropy from the greedy selection rule (7) is a monotonic function in the posterior variance, we can reduce the query strategy and search for points where our regression model is maximally uncertain. We can also simplify the safety constraint (12) by defining the confidence parameter $\nu = \Phi^{-1}(p) \in \mathbb{R}$. Thus, we obtain for our agent the optimization problem

$$\begin{aligned} \boldsymbol{x}_{i+1} &= \arg\max_{\boldsymbol{x}_* \in \mathbb{X}} \operatorname{Var}\big[\, y_* \,\big|\, \boldsymbol{x}_*, \mathcal{D}_i\big] \\ \text{s.t.:} \quad & \mu_{g_*} - \nu\, \sigma_{g_*} \geq 0 \,, \end{aligned} \tag{13}$$

where the moments μ_{g_*} and $\sigma^2_{g_*}$ are defined as in (11). This optimization task is solved via second order optimization techniques. For notational simplicity, let $\mathcal{D}_i$ contain all data for both GP's. In addition to the hyperparameters given previously, we assume that at least $m_0 \geq 1$ safe starting points, i.e. points with positive label c_i, are given at the beginning of the exploration. As stopping criterion for the above problem, we used the maximal number of queries n. It is also possible to replace the latter by a bound for the current model accuracy or the number of feasible points m.

Algorithm 1 Safe Active Learning with GP's

Require: $\mathcal{D}_{m_0}$, ν, $\boldsymbol{\theta}_f$, σ^2, $\boldsymbol{\theta}_g$, τ^2
1: $i = m_0$
2: train model and discriminative GP function on $\mathcal{D}_{m_0}$
3: **while** $i < n$ **do**
4: $\quad i = i + 1$
5: $\quad$ get $\boldsymbol{x}_i$ from solving (13)
6: $\quad$ sample y_i, c_i or h_i and add them to $\mathcal{D}_i$ with query $\boldsymbol{x}_i$
7: $\quad$ train model and discriminative GP function on $\mathcal{D}_i$
8: **end while**

4 Theoretical Analysis

The goal of this section is to investigate our proposed safe AL algorithm. It should be noted that the main objective of our exploration strategy is to avoid samples from unsafe regions $\mathbb{X}_-$ as much as possible. However, a desirable property of an exploration scheme is that it induces a nearly space-filling, i.e. uniform, distribution of the queries in the whole input space $\mathbb{X}$. More precisely, an exploration strategy is called space-filling, if it yields a low-discrepancy input design $\boldsymbol{X}$ of size n such that the discrepancy

$$D(\boldsymbol{X}) = \sup_{B \in \mathcal{B}(\mathbb{X})} \left| \frac{1}{n} \left| \{ \boldsymbol{x}_i \mid \boldsymbol{x}_i \in B \} \right| - \mu_d(B) \right| \leq \gamma \frac{\log^d(n)}{n}$$

for some positive constant γ independent of n and the normalized Lebesgue measure $\mu_d(B)$ for any B which is contained in the Borel algebra $\mathcal{B}(\mathbb{X})$. An example strategy which yields a low-discrepancy design is the Sobol sequence, see Sobol [1976]. Intuitively, it is clear that a space-filling exploration scheme will cover the input space $\mathbb{X}$ and, thus, query in dangerous regions $\mathbb{X}_-$. This supposition is confirmed by the following theorem, where the proof by contradiction is given in the appendix.

Theorem 1. *For every space-filling exploration strategy on $\mathbb{X}$ with a Lebesgue measurable subset $B \subset \mathbb{X}_- \subset \mathbb{X}$ such that $\mu_d(B) > \epsilon$ for some $\epsilon > 0$, there exists a query $\boldsymbol{x}_n \in B$ for an adequate n possibly depending on ϵ.*

For the initialization of our scheme, we assumed that we have at least one safe starting point with positive label. However, depending on the hyperparameter $\boldsymbol{\theta}_g$ and especially the confidence parameter ν, the optimization problem (13) may be empty, i.e. there may not exist any $\boldsymbol{x}_* \in \mathbb{X}$ fulfilling the safety constraint. This behavior is due to the classification part for learning the discriminative function g. Namely, this problem occurs if the user wants a very safe exploration which yields a high confidence parameter ν, but the small GP classification model, i.e. consisting only of a few data points, is too uncertain to enable exploration under such a strong safety constraint. To solve this problem and to obtain a more confident discrimination model, the user has to define an initialization point set

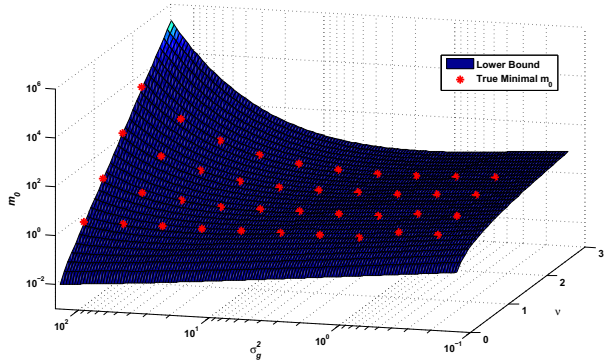

Fig. 2. Lower bound for the number of necessary initialization points m_0 given by Theorem 2 to ensure a non-empty optimization problem (13). The red stars indicate some true necessary set sizes, calculated according to the explanations in the proof.

where potentially all points lie close to each other. The next theorem provides a lower bound for the size m_0 of the initial point set. This result is especially relevant, when employing the proposed safe AL in practice. The proof is moved to the supplemental material.

Theorem 2. *To ensure a non-empty safety constraint in the optimization problem* (13) *for our GP setting, we need at least an initial point set of size*

$$m_0 \geq \left(2\mathcal{N}\left(\tfrac{1}{2}(\sqrt{1+4\nu\sigma_g}-1)\right)\right)^{-1} \min\left(\frac{\nu}{\sqrt{3}\sigma_g}, \sqrt{\frac{\nu}{\sqrt{3}\sigma_g^3}}\right).$$

In Figure 2, the bound from Theorem 2 is illustrated and compared to some true necessary set sizes m_0. As it is hard to restrict the explicit expressions for μ_{g_*} and $\sigma^2_{g_*}$ depending on m_0, the bound of the theorem is adequately tight. Nevertheless, the magnitude and the asymptotic behavior are captured by the lower bound of the theorem. Note that the safety constraint of the optimization problem (13) is always satisfied, if ν is non-positive, i.e. $p \leq 0.5$, which follows from the centered GP prior.

Finally, we will bound the probability of failure for our active exploration scheme to ensure a high level of safety. Having already queried $i-1$ points and if our prior GP assumptions are correct, the probability of failure when sampling $\boldsymbol{x}_* \in \mathbb{X}$ without considering the safety constraint (12) is given by

$$\Pr\left[g_* \leq 0 \,|\, \boldsymbol{x}_*, \mathcal{D}_{i-1}\right] = 1 - \Phi\left(\frac{\mu_{g_*}}{\sigma_{g_*}}\right), \tag{14}$$

where the moments of g_* are explained by the exact posterior distribution followed from (9). In other words, the discriminative function is not positive in the case of failure. We are interested in an upper bound for the probability of making at least one failure, when querying n data points with our safe active learning scheme summarized in Algorithm 1. Our result is stated in the following theorem, where the sketch of the proof is subsequently given.

Theorem 3. *Let $\mathbb{X} \subset \mathbb{R}^d$ be compact and non-empty, pick $\delta \in (0,1)$ and define $\nu = \Phi^{-1}\left(1 - \frac{\delta}{n-m_0}\right)$. Ensure that the discriminative prior and posterior GP is correct. Suppose also that at least an initialization point set of size $m_0 < n$ with respect to Theorem 2 is given. Selecting n possible queries satisfying the safety constraint, our active learning scheme presented in Algorithm 1 is unsafe with probability δ, i.e.*

$$\Pr\left[\bigvee_{i=m_0+1}^{n} \left(g_i \leq 0 \,\middle|\, \boldsymbol{x}_i : \mu_{g_i} - \nu\sigma_{g_i} \geq 0, \mathcal{D}_{i-1}\right)\right] \leq \delta\,.$$

Proof (Theorem 3). Firstly, we need to bound the probability of failure (14) for every possible query $\boldsymbol{x}_*$ fulfilling the safety constraint. For an arbitrary but firmly selected input point $\boldsymbol{x}_i \in \mathbb{X}$

$$\begin{aligned}
&\Pr\left[g_i \leq 0 \,|\, \boldsymbol{x}_i : \mu_{g_i} - \nu\sigma_{g_i} \geq 0, \mathcal{D}_{i-1}\right] \\
&\leq \Pr\left[g_i \leq \mu_{g_i} - \nu\sigma_{g_i} \,|\, \boldsymbol{x}_i : \mu_{g_i} - \nu\sigma_{g_i} \geq 0, \mathcal{D}_{i-1}\right] \\
&\quad = 1 - \Phi(\nu)\ ,
\end{aligned}$$

holds true under our condition. That is, the safety constraint (12) induces a probability of failure which is less than $1-p$ in each iteration, remembering the relationship $p = \Phi(\nu)$. Furthermore, we use the union bound to obtain

$$\begin{aligned}
&\Pr\left[\bigvee_{i=m_0+1}^{n} \left(g_i \leq 0 \,\middle|\, \boldsymbol{x}_i : \mu_{g_i} - \nu\sigma_{g_i} \geq 0, \mathcal{D}_{i-1}\right)\right] \\
&\leq \sum_{i=m_0+1}^{n} \Pr\left[\left(g_i \leq 0 \,\middle|\, \boldsymbol{x}_i : \mu_{g_i} - \nu\sigma_{g_i} \geq 0, \mathcal{D}_{i-1}\right)\right] \\
&\quad = (n - m_0)\left(1 - \Phi(\nu)\right) = \delta\,.
\end{aligned}$$

Note that the m_0 initialization points are feasible with probability 1 under the assumptions of the theorem. □

The lower bound of Theorem 3 provides a safety level of Algorithm 1 greater than or equal to $1-\delta$. The user is then able to choose a sufficiently small δ, when carrying out our algorithm. After determining δ, we calculate ν and, if necessary, p as explained in the theorem. It is clear that, for fixed safety level, ν has to increase, if n increases. In this case, the number of necessary initialization points also increases, see Theorem 2.

To get a more detailed illustration of the bound in Theorem 3, we assume that each query is selected independently of all others. In contrast to the proof of the safety bound, where we did not assume independence, we then have $1-p^{n-m_0} \leq \delta$. Intuitively, we anticipate an upper bound for the expected number of failures when sampling n points. Examples for this bound are shown in the next section.

Moreover, it is necessary for our AL scheme that the function h is a lower bound of the true discriminative function of the system. In this case, we may

loose information when exploring the system. However, the validity of the safety level compared to Theorem 3 is the main requirement for us. The function h must also satisfy the conditions for the specified discriminative GP prior, e.g. continuity and mean square differentiability.

5 Evaluations

In this section, we verify the presented Algorithm 1 on a 1-dimensional toy example and, subsequently, evaluate our safe exploration scheme on an inverse pendulum policy search problem.

Firstly, we learn to approximate the cardinal sine function $f(x) = 10\,\mathrm{sinc}(x-10)$ with additive Gaussian noise $\varepsilon \sim \mathcal{N}(0, 0.0625)$ on the interval $\mathbb{X} = [-10, 15]$. We define a safe region $\mathbb{X}_+ = [-5, 11]$, and, consequently, unsafe regions $\mathbb{X} \setminus \mathbb{X}_+$ for which we always sample negative labels. To recognize when exploring gets dangerous we define $h(x) = \frac{1}{2}(x+5)^2$ for $-5 \leq x < -4$ and $h(x) = \frac{1}{2}(x-11)^2$ for $10 < x \leq 11$. The observation noise for the function h is set to $\mathcal{N}(0, 0.01)$. Otherwise, i.e. within $[-4, 10]$, we sample positive class labels. Hyperparameters of the target and discriminative GP are set to $\sigma_f^2 = 4$, $\lambda_f = 3$, and $\sigma_g^2 = 1$, $\lambda_g = 10$, respectively. The m_0 starting points with respect to Theorem 2 are uniformly sampled in the range $[-0.5, 0.5]$. Finally, we wish to select $n = 40$ input points for different probability levels δ. Table 1 shows the selected safety levels with corresponding confidence parameter ν and the necessary number of starting points m_0 derived in the proposed theorems. The results in the presented table, i.e. the final differential entropy (6) and the number (#) of failures are averaged over ten runs. The maximum possible differential entropy value with 40 points is 19.50, which is almost reached in all cases. The expected number of failures provides only an upper bound for the true bound, since independence is assumed in its calculation over $(n - m_0)\,p$. They are additionally compared in Figure 3. Due to this strong assumption, the upper bound for the expected number of failures will not be very tight. The normalized entropy ratios (8) for all cases and averaged over ten runs are illustrated in Figure 4. Note that for the calculation of the differential entropy H only the safe queries are taken into account, analogously to Table 1. Therefore, the curves for the normalized

Table 1. Averaged results over ten runs of the 1-dimensional toy example. The entropy and the number of failures is slightly decreasing for smaller δ. Otherwise, the parameter ν and m_0 are growing with decreasing δ as explained in the text.

δ	ν	m_0	H	# failures	# expected failures
0.05	2.99	4	17.39	0.0	1.8
0.10	2.77	4	18.54	0.2	3.4
0.20	2.54	4	18.73	0.6	6.5
0.30	2.40	3	18.69	1.1	9.6
0.40	2.30	3	18.88	1.6	12.3
0.50	2.21	3	18.87	2.3	14.6

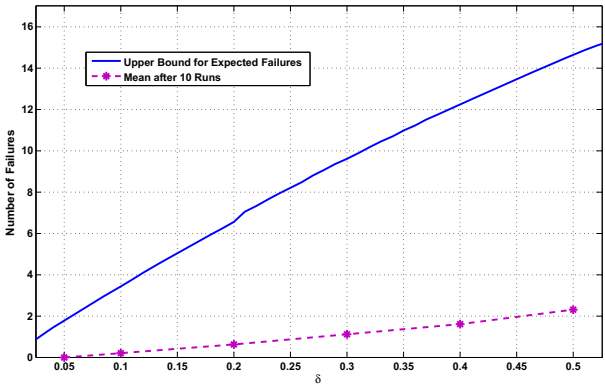

Fig. 3. Expected number of failures calculated under the independence assumption (upper bound) for the toy example compared with the failures obtained by averaging over ten runs of the safe learning algorithm.

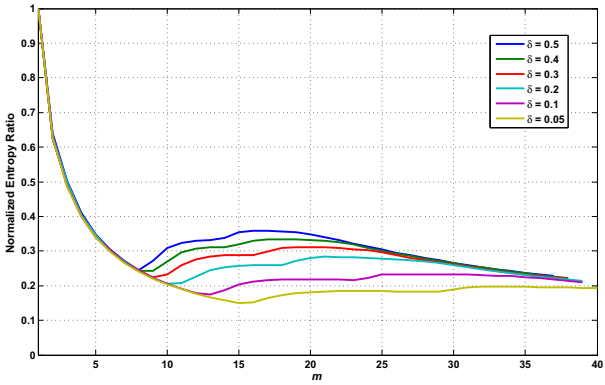

Fig. 4. Normalized entropy ratios for different safety levels and averaged over ten runs. In the beginning phase the ratios fall until the safety constraint is satisfied, since we sample only around zero. After that the gain in entropy increases until the safe region is explored. This behavior and the value of the gain depends on δ and the confidence parameter ν, respectively.

entropy ratios end before the final number of queries is reached. The plot also shows the effects of the decreasing failure level δ for the number of initialization points and the gain in entropy, cf. the theorems in Section 4, where we explore faster and a greater input region with higher δ. In Figure 5 we present the final result for one run of the safe active learning Algorithm 1 after selecting 40 queries. The cardinal sine function is learned accurately over the safe input region. Furthermore, the plot shows that the small definition regions for the function h around the decision boundary are sufficient for safe exploration.

As a second test scenario, we consider exploration of control parameters for the inverse pendulum hold up problem. Here, we wish to learn the mapping between parameters of a linear controller and performance on keeping the pole

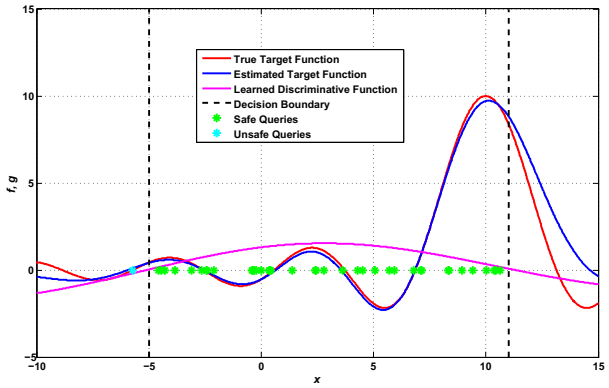

Fig. 5. Final result for safe active learning of a generalized cardinal sine function in the secure interval $[-5, 11]$ with 40 queries and $\delta = 0.30$. Only one selected query fails, i.e. falls below the lower decision boundary. All other chosen data points cover the safe input space $\mathbb{X}_+$ well. The final discriminative GP separates the safe and unsafe regions of the whole input space adequately, even if we selected no query above the upper border.

upwards. The parameters should be explored while avoiding that the pendulum falls over. The simulated system is an inverse pendulum mounted on a cart, see Deisenroth et al. [2015]. The system state $\boldsymbol{s}_t \in \mathbb{S}$ is 4-dimensional (cart position z_t and velocity, pendulum angle ϑ_t and angular velocity), which results in 5 open parameters $\boldsymbol{x}, x_0$ of the linear controller $\pi(\boldsymbol{s}_t) = \boldsymbol{x}^T \boldsymbol{s}_t + x_0$. The desired goal state is defined by cart position $0\,cm$ and the pendulum pointing upwards with $0°$. The controller is applied for 10 seconds with a control frequency of 10 Hz starting from a state sampled around this goal state. We evaluate the performance of a given controller by first measuring the distance of the current state $\boldsymbol{s}_t$ to the goal state, i.e. $\|\boldsymbol{s}_t - \mathbf{0}\|$, for each time step t. These distances are used to compute a saturating cost $r_t = 1 - \exp\left(-\|\boldsymbol{s}_t\|^2\right)$. Additionally, we average over all time steps and ten different starting states to yield a meaningful interpretation of the cost. To that we add Gaussian noise $\varepsilon \sim \mathcal{N}(0, 0.001)$. The hyperparameters of the target and discriminative GP are previously learned over a uniformly distributed point set of size 100 over the entire input space, i.e. policy parameter space

Table 2. Median with respect to the number of failures over ten runs for the policy exploration task from the cart pole. Here the entropy decreases as δ increases, except for the lower values of δ, since the number of unsafe queries increases strongly.

δ	ν	m_0	H	# failures	# expected failures
0.01	4.26	8	393.3	0	9.9
0.05	3.89	7	397.7	6	48.4
0.10	3.72	6	412.8	29	94.6
0.20	3.54	6	402.7	57	180.2
0.30	3.43	5	395.6	80	257.9
0.40	3.35	5	389.3	101	328.1
0.50	3.29	5	380.7	127	391.6

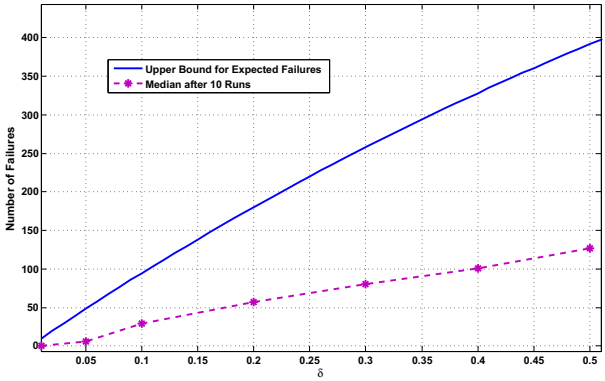

Fig. 6. Expected number of failures calculated under the independence assumption (upper bound) for the policy search task with the failures obtained by the median of ten runs of the safe learning algorithm.

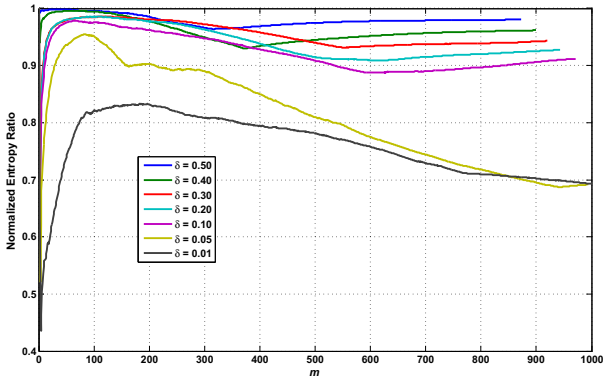

Fig. 7. Median of normalized entropy ratios for different safety levels over ten runs of the policy search task. In the beginning phase the ratios are nearly maximal, i.e. close to one, since we obtain many very safe queries with positive class labels which yield to a strong exploration behavior. After the beginning phase, the safe region is roughly explored until a slight valley is reached. Then the most queries are sampled from the inner region.

$\mathbb{X} \subset \mathbb{R}^5$. The function h is defined over the deviation from the unstable goal state for each time step, i.e. $2 - |z_t| \left[|z_t| > 1\,cm \right] - |\vartheta_t| \left[|\vartheta_t| > 1^\circ \right]$, where $[\cdot]$ is the indicator function. These local errors are averaged over all time steps of the roll-out and all starting states to get the value of h. If $h \geq 0.95$ we get a positive class label and a negative one if $h \leq -1$ or the pendulum falls down. In this case we set $n = 1000$ for various values of δ. Table 2 summarizes the resulting values for ν and m_0 given by the theorems in Section 4. The value of the differential entropy H increases until $\delta = 0.10$, where we yield a good tradeoff between a low number of failures and a fast exploration. In Figure 6 we show that the median of the number of failures after ten runs of our safe exploration scheme is much

lower than the upper bound of the expected number. This behavior clarifies the effectiveness of the Theorem 3 for our safe active learning scheme, analogously to the results of the toy example. Also the nearly heuristic definition of the function h works very well on this exploring task for five different input dimensions. A reason for that is the almost always favorable modeling accuracy when using GP's. The medians of the normalized entropy ratios after ten runs of the inverse pendulum control task are presented in Figure 7. The trends show the effect of the different δ after the beginning phase, where more exploration yields a higher curve. The order of the curves results from the chooses safety level $1 - \delta$ too.

6 Conclusion and Outlook

In this paper, we propose a novel and provable safe exploration strategy in the active learning (AL) context. This approach is especially relevant for real-world AL applications, where critical and unsafe measurements need to be avoided. Empirical evaluations on a toy example and on a policy search task for the inverse pendulum control confirm the safety bounds provided by the approach. Moreover, the experiments show the effectiveness of the presented algorithm, when selecting a near optimal input design, even under the induced safety constraint. The next steps will include evaluations on physical systems and – on the theoretical side – the error bounding of the resulting regression model, which is generally a hard problem.

References

Auer, P.: Using Confidence Bounds for Exploitation-Exploration Trade-Offs. Journal of Machine Learning Research **3**, 397–422 (2002)

Cover, T.M., Thomas, J.A.: Elements of Information Theory. John Wiley & Sons (2006)

Deisenroth, M.P., Fox, D., Rasmussen, C.E.: Gaussian Processes for Data-Efficient Learning in Robotics and Control. Transactions on Pattern Analysis and Machine Intelligence **37**, 408–423 (2015)

Fedorov, V.V.: Theory of Optimal Experiments. Academic Press (1972)

Galichet, N., Sebag, M., Teytaud, O.: Exploration vs exploitation vs safety: risk-aware multi-armed bandits. In: Ong, C.S., Ho, T.B. (eds.) Proceedings of the 5th Asian Conference on Machine Learning, JMLR: W&CP, vol. 29, pp. 245–260 (2013)

Geibel, P.: Reinforcement learning with bounded risk. In: Brodley, C.E., Danyluk, A.P. (eds.) Proceedings of the 18th International Conference on Machine Learning, pp. 162–169 (2001)

Gillula, J.H., Tomlin, C.J.: Guaranteed safe online learning of a bounded system. In: Amato, N.M. (ed.) Proceedings of the International Conference on Intelligent Robots and Systems, pp. 2979–2984 (2011)

Guestrin, C., Krause, A., Singh, A.: Near-Optimal sensor placements in gaussian processes. In: De Raedt, L., Wrobel, S. (eds.) Proceedings of the 22nd International Conference on Machine Learning, pp. 265–275 (2005)

Hans, A., Schneegaß, D., Schäfer, AM., Udluft, S.: Safe Exploration for reinforcement learning. In: Verleysen, M. (ed.) Proceedings of the European Symposium on Artificial Neural Networks, pp. 143–148 (2008)

Ko, C., Lee, J., Queyranne, M.: An Exact Algorithm for Maximum Entropy Sampling. Operations Research **43**, 684–691 (1995)

Krause, A., Guestrin, C.: Nonmyopic active learning of gaussian processes: an exploration–exploitation approach. In: Ghahramani, Z. (ed.) Proceedings of the 24th International Conference on Machine Learning, pp. 449–456 (2007)

Lang, K.J., Baum, E.B.: Query learning can work poorly when a human oracle is used. In: Proceedings of the International Joint Conference on Neural Networks, pp. 335–340 (1992)

Moldovan, T.M., Abbeel, P.: Safe exploration in markov decision processes. In: Langford, J., Pineau, J. (eds.) Proceedings of the 29th International Conference on Machine Learning, pp. 1711–1718 (2012)

Nemhauser, G.L., Wolsey, L.A., Fisher, M.L.: An Analysis of the Approximations for Maximizing Submodular Set Functions. Mathematical Programming **14**, 265–294 (1978)

Nickisch, H., Rasmussen, C.E.: Approximations for Binary Gaussian Process Classification. Journal of Machine Learning Research **9**, 2035–2078 (2008)

Polo, F.J.G., Rebollo, F.F.: Safe reinforcement learning in high-risk tasks through policy improvement. In: Proceedings of the Symposium on Adaptive Dynamic Programming and Reinforcement Learning, pp. 76–83 (2011)

Ramakrishnan, N., Bailey-Kellogg, C., Tadepalli, S., Pandey, V.N.: Gaussian processes for active data mining of spatial aggregates. In: Kargupta, H., Kamath, C., Srivastava, J., Goodman, A. (eds.) Proceedings of the 5th SIAM International Conference on Data Mining, pp. 427–438 (2005)

Rasmussen, C.E., Williams, C.K.I.: Gaussian Processes for Machine Learning. The MIT Press (2006)

Seo, S., Wallat, M., Graepel, T., Obermayer, K.: Gaussian process regression: active data selection and test point rejection. In: Proceedings of the International Joint Conference on Neural Networks vol. 3, pp. 241–246 (2000)

Settles, B.: Active Learning Literature Survey. In: Computer Sciences Technical Report University of Wisconsin, Madison (2010)

Sobol, I.M.: Uniformly Distributed Sequences with an Additional Uniform Property. USSR Computational Mathematics and Mathematical Physics **16**, 236–242 (1976)

Srinivas, N., Krause, A., Kakade, S.M., Seeger, M.W.: Information-Theoretic Regret Bounds for Gaussian Process Optimization in the Bandit Setting. Transactions on Information Theory **58**, 3250–3265 (2012)

Valiant, L.G.: A Theory of the Learnable. Communications of the ACM **27**, 1134–1142 (1984)