Nonmyopic Active Learning of Gaussian Processes: An Exploration–Exploitation Approach

May 2007
CMU-ML-07-105

School of Computer Science
Carnegie Mellon University
Pittsburgh, PA 15213

Abstract

When monitoring spatial phenomena, such as the ecological condition of a river, deciding where to make observations is a challenging task. In these settings, a fundamental question is when an active learning, or sequential design, strategy, where locations are selected based on previous measurements, will perform significantly better than sensing at an a priori specified set of locations. For Gaussian Processes (GPs), which often accurately model spatial phenomena, we present an analysis and efficient algorithms that address this question. Central to our analysis is a theoretical bound which quantifies the performance difference between active and a priori design strategies. We consider GPs with unknown kernel parameters and present a nonmyopic approach for trading off exploration, i.e., decreasing uncertainty about the model parameters, and exploitation, i.e., near-optimally selecting observations when the parameters are (approximately) known. We discuss several exploration strategies, and present logarithmic sample complexity bounds for the exploration phase. We then extend our algorithm to handle nonstationary GPs exploiting local structure in the model. A variational approach allows us to perform efficient inference in this class of nonstationary models. We also present extensive empirical evaluation on several real-world problems.

Keywords: Active Learning, Gaussian Processes, Sequential design, Nonstationarity

1 Introduction

When monitoring spatial phenomena, such as the ecological condition of a river as in Fig. 1, it is of fundamental importance to decide on the most informative locations to make observations. However, to find locations which predict the phenomena best, one needs a model of the spatial phenomenon itself. Gaussian processes (GPs) have been shown to be effective models for this purpose (Cressie, 1991; Rasmussen & Williams, 2006).

Most previous work on observation selection in GPs has considered the *a priori design* problem, in which the locations are selected in advance prior to making observations (*c.f.*, Guestrin et al. (2005); Seo et al. (2000); Zhu and Stein (2006)). Indeed, if the GP model parameters are completely known, the predictive variances do *not* depend on actual observed values, and hence nothing is lost by committing to sampling locations in advance. In the case of unknown parameters however, this independence is no longer true. Key questions we strive to understand in this paper are *how much* better a *sequential* algorithm, taking into account previous observations, can perform compared to a priori design when the parameters are *unknown*, and how can this understanding lead to better observation selection methods.

Our main theoretical result is a bound which quantifies the performance difference between sequential and a priori strategies in terms of the parameter entropy of the prior over kernels. The lower the uncertainty about the parameters, the less we can potentially gain by using an active learning (sequential) strategy. This relationship bears a striking resemblance to the exploration–exploitation tradeoff in Reinforcement Learning. If the model parameters are known, we can *exploit* the model by finding a near-optimal policy for sampling using the mutual information criterion (Caselton & Zidek, 1984; Guestrin et al., 2005). If the parameters are unknown, we present several *exploration* strategies for efficiently decreasing the uncertainty about the model. Most approaches for active sampling of GPs have been *myopic* in nature, in each step selecting observations which, e.g., most decrease the predictive variance. Our approach however is *nonmyopic* in nature: we prove logarithmic sample complexity bounds on the duration of the exploration phase, and near optimal performance in the exploitation phase.

Often, e.g., in spatial interpolation (Rasmussen & Williams, 2006), GP models are assumed to be isotropic, where the covariance of two locations depends only on their distance, and some (unknown) parameters. Many phenomena of interest however are nonstationary (Paciorek, 2003; Nott & Dunsmuir, 2002). In our river example (*c.f.*, Figure 1), the pH values are strongly correlated along the border, but weakly in the turbulent inner region. Our approach is applicable to both stationary and nonstationary processes. However, nonstationary processes are often defined by a much larger number of parameters. To address this issue, we extend our algorithm to handle nonstationary GPs with local structure, providing efficient exploration strategies and computational techniques that handle high dimensional parameter vectors. In summary, our contributions are:

- A theoretical and empirical investigation of the performance difference between sequential and a priori strategies for sampling in GPs;
- An exploration–exploitation analysis and sample complexity bounds for sequential design;
- An efficient, nonmyopic, sequential algorithm for observation selection in isotropic GPs;
- Extension of our method to nonstationary GPs;
- Empirical evaluation on several real-world spatial monitoring problems.

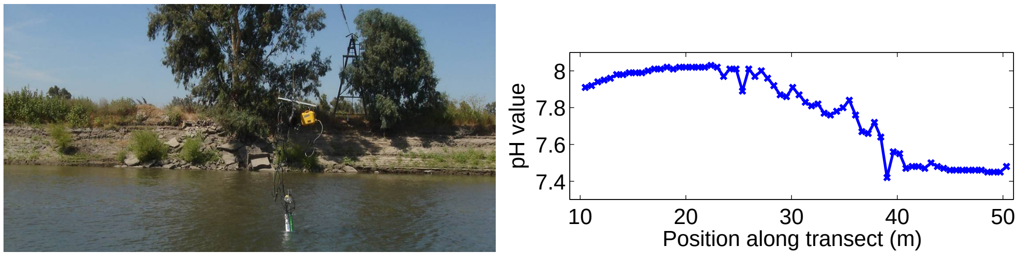


Figure 1: Left: Active sampling using the Networked Infomechanical System (NIMS) sensor (Harmon et al., 2006), deployed at the Merced River. The sensor is attached to a wire, which enables horizontal traversal of the transect. On fixed horizontal position, it can vertically lower or raise the sensing unit. Right: Samples of pH acquired along horizontal transect near the confluence of the San Joaquin and Merced rivers.

2 Gaussian Processes

Consider, for example, the task of monitoring the ecological state of a river using a robotic sensor, such as the one shown in Figure 1. We can model the pH values as a random process $\mathcal{X}_{\mathcal{V}}$ over the locations $\mathcal{V}$, e.g., $\mathcal{V} \subset \mathbb{R}^2$. Hereby, the pH value at every location $y \in \mathcal{V}$ is a random variable $\mathcal{X}_y$. Measurements $\mathbf{x}_{\mathcal{A}}$ at sensor locations $\mathcal{A} \subset \mathcal{V}$ then allow us to predict the pH value at uninstrumented locations y, by conditioning on the observations, i.e., predicting $\mathbb{E}[\mathcal{X}_y \mid \mathcal{X}_{\mathcal{A}} = \mathbf{x}_{\mathcal{A}}]$.

It has been shown, that pH values, temperatures and many other spatial phenomena, can be effectively modeled using Gaussian processes (GPs) (*c.f.*, Shewry and Wynn (1987); Cressie (1991)). A GP (*c.f.*, Rasmussen and Williams (2006)) is a random process $\mathcal{X}_{\mathcal{V}}$, such that every finite subset of variables $\mathcal{X}_{\mathcal{A}} \subseteq \mathcal{X}_{\mathcal{V}}$ has a (consistent) multivariate normal distribution: $P(\mathcal{X}_{\mathcal{A}} = \mathbf{x}_{\mathcal{A}}) = \frac{1}{(2\pi)^{n/2}|\Sigma_{\mathcal{A}\mathcal{A}}|} e^{-\frac{1}{2}(\mathbf{x}_{\mathcal{A}}-\mu_{\mathcal{A}})^T \Sigma_{\mathcal{A}\mathcal{A}}^{-1}(\mathbf{x}_{\mathcal{A}}-\mu_{\mathcal{A}})}$, where $\mu_{\mathcal{A}}$ is the mean vector and $\Sigma_{\mathcal{A}\mathcal{A}}$ is the covariance matrix. A GP is fully specified by a *mean function* $\mathcal{M}(\cdot)$, and a symmetric positive-definite *kernel function* $\mathcal{K}(\cdot, \cdot)$, often called the covariance function. For each random variable $\mathcal{X}_u$ with index $u \in \mathcal{V}$, its mean μ_u is given by $\mathcal{M}(u)$, and for each pair of indices $u, v \in \mathcal{V}$, their covariance σ_{uv} is given by $\mathcal{K}(u, v)$. For simplicity of notation, we denote the mean vector of a set of variables $\mathcal{X}_{\mathcal{A}}$ by $\mu_{\mathcal{A}}$, where the entry for element u of $\mu_{\mathcal{A}}$ is $\mathcal{M}(u)$. Similarly, we denote their covariance matrix by $\Sigma_{\mathcal{A}\mathcal{A}}$, where the entry for u, v is $\mathcal{K}(u, v)$. The GP representation allows us to efficiently compute predictive distributions, $P(\mathcal{X}_y \mid x_{\mathcal{A}})$, which, e.g., correspond to the predicted temperature at location y after observing sensor measurements $\mathcal{X}_{\mathcal{A}} = \mathbf{x}_{\mathcal{A}}$. The distribution of $\mathcal{X}_y$ given these observations is a Gaussian whose conditional mean $\mu_{y|\mathcal{A}}$ and variance $\sigma^2_{y|\mathcal{A}}$ are:

$$\begin{aligned} \mu_{y|\mathcal{A}} &= \mu_y + \Sigma_{y\mathcal{A}} \Sigma_{\mathcal{A}\mathcal{A}}^{-1} (\mathbf{x}_{\mathcal{A}} - \mu_{\mathcal{A}}), & (2.1) \\ \sigma^2_{y|\mathcal{A}} &= \mathcal{K}(y, y) - \Sigma_{y\mathcal{A}} \Sigma_{\mathcal{A}\mathcal{A}}^{-1} \Sigma_{\mathcal{A}y}, & (2.2) \end{aligned}$$

where $\Sigma_{y\mathcal{A}}$ is a covariance vector with one entry for each $u \in \mathcal{A}$ with value $\mathcal{K}(y, u)$, and $\Sigma_{\mathcal{A}y} = \Sigma_{y\mathcal{A}}^T$. An important property of GPs is that the posterior variance (2.2) does *not* depend on the observed values $\mathbf{x}_{\mathcal{A}}$.

In order to compute predictive distributions using (2.1) and (2.2), the mean and kernel functions have to be known. The mean function can usually be estimated using regression techniques.

Estimating kernel functions is difficult, and usually, strongly limiting assumptions are made. For example, it is commonly assumed that the kernel $\mathcal{K}(u,v)$ is *stationary*, depending only on the difference between the locations, i.e., $\mathcal{K}(u,v) = \mathcal{K}_\theta(u-v)$, where θ is a set of parameters. Very often, the kernel is even assumed to be *isotropic*, which means that the covariance only depends on the distance between locations, i.e., $\mathcal{K}(u,v) = \mathcal{K}_\theta(||u-v||_2)$. A common choice for an isotropic kernel is the exponential kernel, $\mathcal{K}_\theta(\delta) = \exp(-\frac{|\delta|}{\theta})$, or the Gaussian kernel, $\mathcal{K}_\theta(\delta) = \exp(-\frac{\delta^2}{\theta^2})$. Many other parametric forms are possible.

In Section 3, we address a general form (not necessarily isotropic), where the kernel function is specified by a set of parameters θ. We adopt a hierarchical Bayesian approach and assign a prior $P(\theta)$ to the parameters θ, which we assume to be discretized in our analysis. Hence, $P(\mathcal{X}_y \mid \mathcal{X}_\mathcal{A}) = \sum_\theta P(\mathcal{X}_y \mid \mathcal{X}_\mathcal{A}, \theta) P(\theta \mid \mathcal{X}_\mathcal{A})$. For clarity of presentation, we also assume that the prior mean function $\mathcal{M}(\cdot)$ is zero. This assumption can be relaxed, for example by assigning a normal prior to the mean function.

3 Observation selection policies

Entropy. In order to select informative observations, the *entropy* criterion has been frequently used (*c.f.*, Seo et al. (2000); Shewry and Wynn (1987); Gramacy (2005)). This criterion selects observations $\mathcal{A}^* \subseteq \mathcal{V}$ with highest entropy,

$$\mathcal{A}^* = \operatorname{argmax}_{\mathcal{A}\subseteq\mathcal{V}} H(\mathcal{X}_\mathcal{A}), \tag{3.1}$$

where $H(\mathcal{X}_\mathcal{A}) = -\int p(\mathbf{x}_\mathcal{A}) \log p(\mathbf{x}_\mathcal{A}) d\mathbf{x}_\mathcal{A}$ is the joint (differential) entropy of the random variables $\mathcal{X}_\mathcal{A}$. We call (3.1) an *a priori* design criterion, as it does not depend on the actual observed values, and can be optimized in advance. Maximizing (3.1) is NP-hard (Ko et al., 1995), so usually, a myopic (greedy) algorithm is used. Starting with the empty set, $\mathcal{A}^{(0)}$, at each step t it adds the location $y_i = \operatorname{argmax}_{y\in\mathcal{V}\setminus\mathcal{A}_{i-1}} H(\mathcal{X}_{y_i} \mid \mathcal{X}_{\mathcal{A}_{i-1}})$ to the set of already selected locations $\mathcal{A}_{i-1}$.

This a priori greedy rule is readily turned into a *sequential* algorithm, selecting

$$y_i = \operatorname*{argmax}_{y\in\mathcal{V}\setminus\mathcal{A}_{i-1}} H(\mathcal{X}_{y_i} \mid \mathcal{X}_{\mathcal{A}_{i-1}} = \mathbf{x}_{\mathcal{A}_{i-1}}).$$

In this sequential setting, the selected location y_i depends on the observations $\mathbf{x}_{\mathcal{A}_{i-1}}$. More generally, we define a *policy* for selecting variables, which *does not* need to be greedy: For each instantiation of the process $\mathcal{X}_\mathcal{V} = \mathbf{x}_\mathcal{V}$, such a sequential policy π can select a *different* set of observations $\pi(\mathbf{x}_\mathcal{V}) \subseteq \mathcal{V}$. Hereby, the i-th element, π_i, deterministically depends on the observations made in the first $i-1$ steps, i.e., on $\mathbf{x}_{\pi_{1:i-1}}$. Hence, a policy can be considered a decision tree, where after each observation, we decide on the next observation to make. If we apply the greedy policy π_{GH} to our river example, $\pi_{GH,i}$ would select the location which has highest entropy for predicting pH, conditioned on the measurements we have made so far. We write $|\pi| = k$ to indicate that π selects sets $\mathcal{X}_\pi$ of k elements. In analogy to the definition of $H(\mathcal{X}_\mathcal{A})$, we can define the joint entropy of any sequential policy π as $H(\mathcal{X}_\pi) \equiv -\int p(\mathbf{x}_\mathcal{V}) \log p(\mathbf{x}_\pi) d\mathbf{x}_\mathcal{V}$, whereby $\pi = \pi(\mathbf{x}_\mathcal{V})$ denotes the set of observations selected by the policy in the event $\mathcal{X}_\mathcal{V} = \mathbf{x}_\mathcal{V}$. $H(\mathcal{X}_\mathcal{A})$ is the entropy of a fixed set of variables $\mathcal{A}$. Since π will typically select different observations in different realizations $\mathcal{X}_\mathcal{V} = \mathbf{x}_\mathcal{V}$, $H(\mathcal{X}_\pi)$ will measure the "entropy" of different variables in each realization $\mathbf{x}_\mathcal{V}$.

Mutual information. Caselton and Zidek (1984) proposed the *mutual information* criterion for observation selection, $\mathrm{MI}(\mathcal{X}_\mathcal{A}) = H(\mathcal{X}_{\mathcal{V}\setminus\mathcal{A}}) - H(\mathcal{X}_{\mathcal{V}\setminus\mathcal{A}} \mid \mathcal{X}_\mathcal{A})$. Guestrin et al. (2005) showed that this criterion selects locations which most effectively reduce the uncertainty at the unobserved locations, hence it often leads to better predictions compared to the entropy criterion. A natural generalization of mutual information to the sequential setting is

$$\begin{aligned}\mathrm{MI}(\mathcal{X}_\pi) &= H(\mathcal{X}_{\mathcal{V}\setminus\pi}) - H(\mathcal{X}_{\mathcal{V}\setminus\pi} \mid \mathcal{X}_\pi) \\ &= -\int p(\mathbf{x}_\mathcal{V})[\log p(\mathbf{x}_{\mathcal{V}\setminus\pi}) - \log p(\mathbf{x}_{\mathcal{V}\setminus\pi} \mid \mathbf{x}_\pi)]d\mathbf{x}_\mathcal{V}.\end{aligned}$$

Hereby, for each realization $\mathcal{X}_\mathcal{V} = \mathbf{x}_\mathcal{V}$, $\mathcal{V}\setminus\pi = \mathcal{V}\setminus\pi(\mathbf{x}_\mathcal{V})$ is the set of locations not picked by the policy π. The greedy policy π_{GMI} for mutual information, after some algebraic manipulation, is given by:

$$\pi_i {=} \mathrm{argmax}_y\, H(\mathcal{X}_y|\mathcal{X}_{\pi_{1:i-1}} {=} \mathbf{x}_{\pi_{1:i-1}}) {-} H(\mathcal{X}_y \,|\, \mathcal{X}_{\bar{\pi}_{1:i-1}}), \tag{3.2}$$

where $\pi_i \equiv \pi_i(\mathbf{x}_{\pi_{1:i-1}})$, and $\bar{\pi} \equiv \mathcal{V} \setminus \{y, \pi(\mathbf{x}_\mathcal{V})\}$ is the set of "unsensed" locations if $\mathcal{X}_\mathcal{V} = \mathbf{x}_\mathcal{V}$, excluding y.

4 Bounds on the advantage of active learning strategies

A key question in active learning is to determine the potential of improvement of sequential strategies over a priori designs, e.g., how much greater $\max_{|\pi|=k} H(\mathcal{X}_\pi)$ is than $\max_{|\mathcal{A}|=k} H(\mathcal{X}_\mathcal{A})$. If the GP parameters θ are known, it holds that

$$H(\mathcal{X}_y|\mathcal{X}_\mathcal{A} {=} \mathbf{x}_\mathcal{A}, \theta) {=} \frac{1}{2}\log 2\pi e \sigma^2_{\mathcal{X}_y|\mathcal{X}_\mathcal{A}} {=} H(\mathcal{X}_y|\mathcal{X}_\mathcal{A}, \theta), \tag{4.1}$$

where $\sigma^2_{\mathcal{X}_y|\mathcal{X}_\mathcal{A}}$, as given by Equation (2.2). Thus, the entropy of a set of variables does not depend on the actual observed values $\mathbf{x}_\mathcal{A}$. Hence, perhaps surprisingly, in this case, $\max_{|\pi|=k} H(\mathcal{X}_\pi) = \max_{|\mathcal{A}|=k} H(\mathcal{X}_\mathcal{A})$. More generally, any objective function depending only on the predictive variances, cannot benefit from sequential strategies. Note that for non-Gaussian models, sequential strategies can strictly outperform a priori designs, even with known parameters.

With unknown parameters, $H(\mathcal{X}_\mathcal{A}) = -\sum_\theta \int P(\mathbf{x}_\mathcal{A}, \theta) \log \left(\sum_{\theta'} \int P(\mathbf{x}_\mathcal{A}, \theta')\right) d\mathbf{x}_\mathcal{A}$ is the entropy of a mixture of GPs. Since observed values affect the posterior over the parameters $P(\Theta|\mathcal{X}_\mathcal{A} {=} \mathbf{x}_\mathcal{A})$, the predictive distributions now depend on these values. Intuitively, if we have low uncertainty about our parameters, the predictive distributions should be *almost* independent of the observed values, and there should be *almost* no benefit from sequential strategies. We will now theoretically formalize this intuition.

The following central result achieves this goal, by bounding $H(\mathcal{X}_\pi)$ (and similarly for mutual information) of the optimal *policy* π by a mixture of entropies of *sets* $H(\mathcal{X}_{\mathcal{A}_\theta} \mid \theta)$, whereby the sets are $\mathcal{A}_\theta$ are chosen optimally for each fixed parameter θ (and can thus be selected a priori, without a sequential policy):

Theorem 1.

$$\max_{|\pi|=k} H(\mathcal{X}_\pi) \leq \sum_\theta P(\theta) \max_{|A|=k} H(\mathcal{X}_\mathcal{A} \mid \theta) + H(\Theta);$$

$$\max_{|\pi|=k} \mathrm{MI}(\mathcal{X}_\pi) \leq \sum_\theta P(\theta) \max_{|A|=k} \mathrm{MI}(\mathcal{X}_\mathcal{A} \mid \theta) + H(\Theta).$$

The proofs of all theorems can be found in the Appendix.

Theorem 1 bounds the advantage of sequential designs by two components: The expected advantage by optimizing sets for known parameters, i.e., $\sum_\theta P(\theta) \max_{|A|=k} \mathrm{MI}(\mathcal{X}_\mathcal{A} \mid \theta)$, and the parameter entropy, $H(\Theta)$. This result implies, that if we are able to (approximately) find the best set of observations $\mathcal{A}_\theta$ for a GP with known parameters θ, we can bound the advantage of using a sequential design. If this advantage is small, we select the set of observations ahead of time, without having to wait for the measurements.

5 Exploration–Exploitation Approach towards Learning GPs

Theorem 1 allows two conclusions: Firstly, if the parameter distribution $P(\Theta)$ is very peaked, we cannot expect active learning strategies to drastically outperform a priori designs. More importantly however, it motivates an exploration–exploitation approach towards active learning of GPs: If the bound provided by Theorem 1 is close to our current mutual information, we can exploit our current model, and optimize the sampling without having to wait for further measurements. If the bound is very loose, we explore, by making observations to improve the bound from Theorem 1. We can compute the bound while running the algorithm to decide when to stop exploring.

5.1 Exploitation using Submodularity

Theorem 1 shows that in order to bound the value of the optimal policy, it suffices to bound the value of the optimal set. Guestrin et al. (2005) derived such a bound for mutual information, using the concept of *submodularity.* A set function F on $\mathcal{V}$ is called submodular if it satisfies the following diminishing returns property: for all $\mathcal{A} \subseteq \mathcal{B} \subseteq \mathcal{V}$ and all $x \notin \mathcal{B}$ it must hold that $F(\mathcal{A} \cup \{x\}) - F(\mathcal{A}) \geq F(\mathcal{B} \cup \{x\}) - F(\mathcal{B})$. Intuitively, this diminishing returns property makes sense for selecting observations: a new observation decreases our uncertainty more if we know less. A set function is called *nondecreasing* if for all $\mathcal{A} \subseteq \mathcal{B} \subseteq \mathcal{V}$ it holds that $F(\mathcal{A}) \leq F(\mathcal{B})$. A fundamental result about nondecreasing submodular functions is the guarantee that the greedy algorithm, which greedily adds the element x to $\mathcal{A}$ such that $F(\mathcal{A} \cup \{x\}) - F(\mathcal{A})$ is largest, selects a set $\mathcal{A}_G$ of k elements which is at most a constant factor $(1 - 1/e)$ worse than the set of k elements of maximal value, i.e., $F(\mathcal{A}_G) \geq (1 - 1/e) \max_{|\mathcal{A}|=k} F(\mathcal{A})$ (Nemhauser et al., 1978). Guestrin et al. (2005) showed that mutual information is submodular and approximately non-decreasing. More specifically:

Theorem 2 (Guestrin et al. (2005))**.** *Let $\mathcal{X}_\mathcal{V}$ be a Gaussian process. Under sufficiently fine discretization $\mathcal{V}$, the greedy algorithm for mutual information is guaranteed to select a set $\mathcal{A}_G$ of k sensors for which* $\mathrm{MI}(\mathcal{X}_{\mathcal{A}_G}) \geq (1 - 1/e)(\mathrm{OPT} - k\varepsilon)$, *where* OPT *is the mutual information achieved by the optimal placement, and ε depends polynomially on the discretization.*

Hence, we have the following result about exploitation using the mutual information criterion:

Corollary 3. *Choose the discretization of the GP such that Theorem 2 holds for all discrete values of Θ. Then* $\mathrm{MI}(\mathcal{X}_{\mathcal{A}_G} \mid \Theta) \leq \max_{|\pi|=k} \mathrm{MI}(\mathcal{X}_\pi) \leq (1 - 1/e)^{-1} \sum_\theta P(\theta) \mathrm{MI}(\mathcal{X}_{\mathcal{A}_G}^{(\theta)} \mid \theta) + k\varepsilon + H(\Theta)$, *where $\mathcal{A}_G$ is the greedy set for* $\mathrm{MI}(\mathcal{X}_\mathcal{A} \mid \Theta) = \sum_\theta P(\theta) \mathrm{MI}(\mathcal{X}_\mathcal{A} \mid \theta)$, *and $\mathcal{A}_G^{(\theta)}$ is the greedy set for* $\mathrm{MI}(\mathcal{X}_\mathcal{A} \mid \theta)$.

This result allows us to *efficiently compute* online bounds on how much can be gained by following a sequential active learning strategy. Intuitively, it states that if this bound is close to our current mutual information, we can stop exploring, and exploit our current knowledge about the model by near-optimally finding the best set of observations. We can also use Corollary 3 as a *stopping criterion*: We can use exploration techniques (as described in the next section) until the bound on the advantage of the sequential strategy drops below a specified threshold η, i.e., we stop if

$$\frac{(1-1/e)^{-1}\sum_\theta P(\theta)\,\mathrm{MI}(\mathcal{X}^{(\theta)}_{\mathcal{A}_G}\mid\theta)+k\varepsilon+H(\Theta)-\mathrm{MI}(\mathcal{X}_{\mathcal{A}_G}\mid\Theta)}{\mathrm{MI}(\mathcal{X}_{\mathcal{A}_G}\mid\Theta)}\le\eta.$$

In this case, we can use the greedy a priori design to achieve near-optimal mutual information, and obtain performance comparable to the optimal sequential policy. This a priori design is logistically simpler and easier to analyze. Hence, the stopping criterion interpretation of Corollary 3 has strong practical value, and we are not aware of any other approach for actively learning GPs which allow to compute such a stopping criterion.

5.2 Implicit and Explicit Exploration

In order to practically use Corollary 3 as a stopping criterion for exploration, we have to, for each parameter θ, solve the optimization problem $\max_{\mathcal{A}} H(\mathcal{X}_{\mathcal{A}}\mid\theta)$. The following theorem shows, that if the parameter entropy is small enough, the contribution of the term $\sum_\theta P(\theta)\max_{|A|=k}\mathrm{MI}(\mathcal{X}_{\mathcal{A}}\mid\theta)$ to the bound diminishes quickly, and hence, we should concentrate solely on minimizing the parameter entropy $H(\Theta)$.

Theorem 4. *Let* $M=\max_A\max_{\theta_1,\theta_2}\frac{\mathrm{MI}(\mathcal{X}_{\mathcal{A}}|\theta_1)}{\mathrm{MI}(\mathcal{X}_{\mathcal{A}}|\theta_2)}<\infty$. *Let* $K=\max_\theta\max_{\mathcal{A}}\mathrm{MI}(\mathcal{X}_{\mathcal{A}}\mid\theta)$, $H(\Theta)<1$. *Then*

$$\mathrm{MI}(\mathcal{X}_{\mathcal{A}^*}\mid\Theta)-H(\Theta)\le\mathrm{MI}(\mathcal{X}_{\pi^*})\le\mathrm{MI}(\mathcal{X}_{\mathcal{A}^*}\mid\Theta)+CH(\Theta),$$

where $\mathcal{A}^*=\mathrm{argmax}_{\mathcal{A}}\,\mathrm{MI}(\mathcal{X}_{\mathcal{A}}\mid\Theta)$ *and* $\pi^*=\mathrm{argmax}_\pi\,\mathrm{MI}(\mathcal{X}_\pi)$, *and* $C=\left(1+\frac{MK}{\log_2\frac{1}{H(\Theta)}}\right)$.

As a function of $H(\Theta)$, C converges to 1 very quickly as $H(\Theta)$ decreases. Theorem 4 hence provides the computational advantage, that, once the parameter entropy is small enough, we do not need to recompute the term $\sum_\theta P(\theta)\max_{|A|=k}\mathrm{MI}(\mathcal{X}_{\mathcal{A}}\mid\theta)$ when using Theorem 1 as a criterion for stopping exploration. Hence, in the following, we concentrate on directly decreasing the parameter uncertainty. We describe three natural strategies for this goal. As we show in Section 7, none of these strategies dominates the other; whichever is more appropriate depends on the particular application.

Explicit Exploration via Independence Tests (ITE). In many cases, the unknown parameter of an isotropic GP is the bandwidth of the kernel, effectively scaling the kernel over space. Let $\theta_1<\cdots<\theta_m$ be the possible bandwidths. In the exponential kernel, $\mathcal{K}_\theta(\delta)=\exp(-\frac{|\delta|}{\theta})$, or the Gaussian kernel, $\mathcal{K}_\theta(\delta)=\exp(-\frac{\delta^2}{\theta^2})$, the correlation between two variables at distance δ decreases exponentially with their distance δ. Hence, there is an *exponentially large gap* between the correlation for bandwidths θ_i and θ_{i+1}: There will be a distance $\hat{\delta}$, for which two random variables within this distance will appear dependent if the true bandwidth θ is at least $\theta\ge\theta_{i+1}$, and (roughly) independent if $\theta\le\theta_i$. Our goal is to exploit this gap to efficiently determine the correct parameter.

First note that if we can separate θ_i from θ_{i+1}, we effectively distinguish any θ_j, for $j\le i$, from θ_l, for $l\ge i+1$, since the bandwidths scale the kernels. Let I_i be a function of Θ, such that $(I_i\mid\Theta)=0$ if $\Theta\le\theta_i$, and $(I_i\mid\Theta)=1$ if $\Theta\ge\theta_{i+1}$. Assume we have tests T_i, using $\hat{N}$ samples, such that

$P(T_i \neq I_i \mid \theta) \leq \alpha$ for all θ. We can now use a *binary search* procedure to identify the true bandwidth with high probability using at most $\hat{N}\lceil \log_2 m \rceil$ samples. Let $\pi_{G\circ ITE}$ be the policy, where we first explore using ITE, and then greedily select the set $\mathcal{A}_G$ maximizing $\mathrm{MI}(\mathcal{X}_{\mathcal{A}_G} \mid \Theta, \mathbf{x}_{\pi_{ITE}})$. Let $\mathbf{x}_{\pi_{ITE}}$ be the observations made by ITE, and let $\mathcal{A}_G^{(\theta)}$ be the solution of the greedy algorithm for optimizing $\mathrm{MI}(\mathcal{X}_{\mathcal{A}} \mid \theta)$.

Theorem 5. *Under the assumptions of Corollary 3 for sets of sizes up to $k + \hat{N}\lceil \log m \rceil$, if we have tests T_i using at most $\hat{N}$ samples, such that for all θ: $P(T_i \neq I_i \mid \theta) \leq \alpha/(\lceil \log m \rceil^2 (\max_\theta |\mathrm{MI}(\mathcal{X}_{\pi_{G\circ ITE}} \mid \Theta) - \mathrm{MI}(\mathcal{X}_{\mathcal{A}_G^{(\theta)}} \mid \theta)|))$:*

$$\mathbb{E}_T[\mathrm{MI}(\mathcal{X}_{\pi_{G\circ ITE}} \mid \Theta)] \geq (1 - 1/e) \max_{|\pi|=k} \mathrm{MI}(\mathcal{X}_\pi) - k\varepsilon - \alpha.$$

In order to make use of Theorem 5, we need to find tests T_i such that $P(T_i \neq I_i \mid \theta)$ is sufficiently small for all θ. If only the bandwidth is unknown, we can for example use a test based on Pearson's correlation coefficient. Since this test requires independent samples, let us first assume, that the kernel function has bounded support (*c.f.*, Storkey (99)), and that the domain of the GP is sufficiently large, such that we can get independent samples by sampling pairs of variables outside the support of the "widest" kernel. The number of samples will depend on the error probability α, and the difference $\hat{\rho}$ between the correlations depending on whether $\Theta \leq \theta_i$ or $\Theta \geq \theta_{i+1}$. This difference will in turn depend on the distance between the two samples. Let

$$\hat{\rho}_i = \max_{\delta} \min_{j \leq i, l \geq i+1} \left| \mathcal{K}_{\theta_j}(\delta) - \mathcal{K}_{\theta_l}(\delta) \right|, \text{ and}$$

$$\hat{\delta}_i = \operatorname*{argmax}_{\delta} \min_{j \leq i, l \geq i+1} \left| \mathcal{K}_{\theta_j}(\delta) - \mathcal{K}_{\theta_l}(\delta) \right|.$$

$\hat{\rho}_i$ is the maximum "gap" achievable for separating bandwidths at most θ_i from those at least θ_{i+1}. $\hat{\delta}_i$ is the distance at which two samples should be taken to achieve this gap in correlation. If several feasible pairs of locations are avaible, we choose the one which maximizes mutual information.

Theorem 6. *We need $\hat{N}_i = \mathcal{O}\left(\frac{1}{\hat{\rho_i}^2} \log^2 \frac{1}{\alpha}\right)$ independent pairs of samples at distance $\hat{\delta}_i$ to decide between $\theta \leq \theta_i$ or $\theta \geq \theta_{i+1}$ with $P(T_i \neq I_i \mid \theta) \leq \alpha$ for all θ.*

In the case of kernels with non-compact support, such as the Gaussian or Exponential kernel[1], we cannot generate such independent samples, since distant points will have some (exponentially small) correlation. However, these almost independent samples suffice:

Corollary 7. *Let $\mathcal{X}$ have variance σ^2, measurement noise σ_n^2 at each location, $\hat{\rho} = \min_i \hat{\rho}_i$, and $\xi < \hat{\rho}$. We can obtain a test T_i with $P(T_i \neq I_i \mid \theta) \leq \alpha$ using $\hat{N} = \mathcal{O}\left(\frac{1}{(\hat{\rho}-\xi)^2} \log^2 \frac{1}{\alpha}\right)$ pairs of samples $\mathcal{X}_s = (\mathcal{X}_{s_1}, \mathcal{X}_{s_2})$ at distance $\hat{\delta}_i$, if, for every $\mathcal{X}_s$ and $\mathcal{X}_t$ in our sample set, $\mathrm{Cor}(\mathcal{X}_{s_i}, \mathcal{X}_{t_j}) \leq \sqrt{\frac{\xi \sigma_n^2}{4\sigma^2 \hat{N} \lceil \log_2 m \rceil}}$, for $i, j \in \{1, 2\}$.*

Hence, since most kernel functions decay exponentially fast, only a small spatial distance has to be guaranteed between the pairs of samples of the independence tests. Note that while this discussion

[1] For the Gaussian and the Exponential kernel for example, we can compute $\hat{\rho}_i$ analytically.

focused on detecting bandwidths, the technique is general, and can be used to distinguish other parameters, e.g., variance, as well, as long as appropriate tests are available.

This hypothesis testing exploration strategy gives us sample complexity bounds. It guarantees that with a small number of samples we can decrease the parameter uncertainty enough such that, using Theorem 4 as stopping criterion, we can switch to exploitation.

Explicit Exploration based on Information Gain (IGE). As the bound in Theorem 4 directly depends on $H(\Theta)$, another natural exploration strategy is to select samples which have highest information gain about the *parameters*, $H(\Theta)$. More formally, this strategy, after observing samples $\mathcal{X}_{\pi_{1:i}} = \mathbf{x}_{\pi_{1:i}}$, selects the location π_{i+1} such that $\pi_{i+1} = \operatorname{argmax}_y H(\Theta \mid \mathbf{x}_{\pi_{1:i}}) - H(\Theta \mid \mathcal{X}_y, \mathbf{x}_{\pi_{1:i}})$.

Implicit Exploration (IE). The following generalization of the "information never hurts" principle (Cover & Thomas, 1991) to policies shows that *any* exploration strategy will, in expectation, decrease $H(\Theta)$.

Proposition 8. *Let $\mathcal{X}_\mathcal{V}$ be a GP with kernel parameters Θ. Let π be a policy for selecting observations. Then $H(\Theta \mid \mathcal{X}_\pi) \leq H(\Theta)$.*

Considering the near-optimal performance of the greedy heuristic in the a priori case, a natural implicit exploration strategy is the sequential greedy algorithm. Using Eq. (3.2), IE considers the previous observations, when deciding on the next observation, and, using Proposition 8, *implicitly* decreases $H(\Theta)$.

6 Actively learning nonstationary GPs

Many spatial phenomena are nonstationary, being strongly correlated in some areas of the space and very weakly correlated in others. In our river example, we consider the pH values in the region just below the confluence of the San Joaquin and Merced rivers. The former was dominated by agricultural and wetland drainage, whereas, in contrast, the latter was less saline. The data (*c.f.*, Figure 2(a)) is very nonstationary. There is very high correlation and low variance in the outer regions. The turbulent confluence region however exhibits high variance and low correlation.

Modeling nonstationarity has to trade off richness of the model and computational and statistical tractability. Even though the covariance function is a an infinite dimensional object, often a parametric form is chosen. For example, Nott and Dunsmuir (2002) suggest to model nonstationarity by a spatially varying linear combination of isotropic processes. In any such a parametric setting, Corollary 3 holds without additional assumptions; the major difference is that $H(\Theta)$ can be much larger, increasing the potential for improvement of the active strategy over the a priori design.

6.1 Nonstationary model

Motivated by the river monitoring problem, we partition the space into disjoint regions $\mathcal{V}^{(1)}, \dots, \mathcal{V}^{(m)}$, which are specified by the user. With each region $\mathcal{V}^{(i)}$, we associate an isotropic process $\mathcal{X}_\mathcal{V}^{(i)}$, with parameters $\Theta^{(i)}$, which are assumed to have independent priors. We define our GP prior for the full space $\mathcal{V}$ as a linear combination of the local GPs: $\mathcal{X}_s = \sum_i \lambda_i(s)\mathcal{X}_s^{(i)}$. Note that such a linear combination is still a valid GP. How should we choose the weights $\lambda_i(s)$? We want a model which behaves similar to process $\mathcal{X}_{\mathcal{V}}^{(i)}$ within region i, and interpolates smoothly between regions. In order to achieve that, we associate a weighting function $\nu_i(s)$ with each region. This function should achieve its maximum value in region i and decrease with distance to region i. In our river example, we set the weighting functions as indicated in Figure 2(a). We can then set $\lambda_i(s) = \sqrt{\frac{\nu_i(s)}{\sum_{i'} \nu_{i'}(s)}}$, which ensures that the variance at location s is a convex combination of the variances of the local GPs, with contribution proportional to $\nu_i(s)$. If each $\mathcal{X}_{\mathcal{V}}^{(i)}$ has zero mean, and kernel $\mathcal{K}_i(s,t)$, then the new, nonstationary GP $\mathcal{X}_{\mathcal{V}}$ has the kernel $\sum_i \lambda_i(s)\lambda_i(t)\,\mathcal{K}_i(s,t)$. By adding a deterministic function $\mathcal{M}(s)$, one can also modify the prior mean of the GP. While the decomposition into prespecified regions might appear restrictive, in many applications, as in the river monitoring setting, a good decomposition can be provided by an expert. Furthermore, one can control the amount of smoothing by a bandwidth parameter, which can be part of the model. By this approach, the data itself can decide whether two adjacent regions should be joined (high smoothing bandwidth) or almost independent (low smoothing bandwidth).

6.2 Efficient Nonstationary Active Learning

Now, in principle we could apply Corollary 3 to this model to determine when to switch from exploration (e.g., using information gain) to exploitation. However, even if each $\Theta^{(i)}$ is discretized so that the distribution over $\Theta^{(i)}$ can be exactly maintained, the joint distribution over $\Theta = (\Theta^{(1)}, \dots, \Theta^{(m)})$ is exponentially large in m. In order to address this problem, let us first consider the special case where each $\nu_{(i)}$ is positive only within region i. In this case, an observation made in region i only affects the prediction and parameter estimation in region i. The joint distribution over Θ will always stay fully factorized, and efficient inference is possible. We effectively monitor a collection of independent GPs, and our active learning algorithm attempts to optimally allocate the samples to the independent GPs.

Now let us consider the general case, where the weights ν_i take positive values outside region i. In this case, an observation s made with positive weights $\nu_i(s) > 0$ and $\nu_j(s) > 0$ for two regions i and j effectively couples the parameters $\Theta^{(i)}$ and $\Theta^{(j)}$. Eventually, all parameters become dependent, and we need to maintain the full, exponentially large joint distribution. In order to cope with this complexity, we apply a variational approach: After making an observation, we find a fully factorized[2] approximate posterior distribution, which is closest in KL divergence. More formally, given a prior $P(\Theta)$ over the parameters and a set of locations $\mathcal{A} \subseteq \mathcal{V}$ and their values $\mathcal{X}_{\mathcal{A}} = \mathbf{x}_{\mathcal{A}}$, we seek the distribution

$$\widehat{P}(\Theta) = \underset{P' \text{ factorized}}{\operatorname{argmin}} \; KL(P(\Theta \mid \mathcal{X}_{\mathcal{A}} = \mathbf{x}_{\mathcal{A}}) \mid\mid P'(\Theta)).$$

For the multinomial distribution, the solution $\widehat{P}$ minimizing the KL divergence can be obtained by matching the marginals of the exact posterior (Koller & Friedman, 2007). The following proposition shows that this procedure does not invalidate our stopping criterion.

Proposition 9. $H(\widehat{P}(\Theta)) \geq H(P(\Theta \mid \mathcal{X}_{\mathcal{A}} = \mathbf{x}_{\mathcal{A}}))$. *Hence, using Theorem 4, our variational approach never stops exploring too early.*

In order to use this nonstationary model for active learning, we need to condition on observations and compute mutual information efficiently.

[2] More complex distributions, which still allow efficient inference, such as trees, can be used as well.

Computing conditional distributions. We assume we have a fully factorized distribution $\widehat{P}(\Theta)$, which already incorporates previous observations $\mathcal{X}_\mathcal{A} = \mathbf{x}_\mathcal{A}$, and we want to incorporate a new observation $\mathcal{X}_s = \mathbf{x}_s$ at location s. We first find the *relevant regions* $\mathcal{V}^{(i_1)}, \dots, \mathcal{V}^{(i_m)}$. A region is relevant[3] to location s if $\nu_j(s) > 0$. For each joint instantiation of the relevant parameters $\bar{\theta} = (\theta_{i_1}, \dots, \theta_{i_m})$, we compute the likelihood of the observation $P(\mathcal{X}_s = \mathbf{x}_s \mid \bar{\theta}, \mathbf{x}_{\mathcal{A}'})$, where $\mathbf{x}_{\mathcal{A}'}$ are the previous observations made within the *relevant* regions. Using Bayes' rule, $P(\bar{\theta} \mid \mathbf{x}_s, \mathbf{x}_\mathcal{A}) \propto \widehat{P}(\bar{\theta}) P(\mathbf{x}_s \mid \mathbf{x}_\mathcal{A}, \bar{\theta})$, we can compute the exact parameter posterior. Remembering all observed data, we can always compute $P(\bar{\theta} \mid \mathbf{x}_s, \mathbf{x}_\mathcal{A})$ using GP regression. Now that we have the exact parameter posterior, we find the KL-minimizing fully factorized approximation to $P(\bar{\theta} \mid \mathbf{x}_s, \mathbf{x}_\mathcal{A})$ by marginalisation.

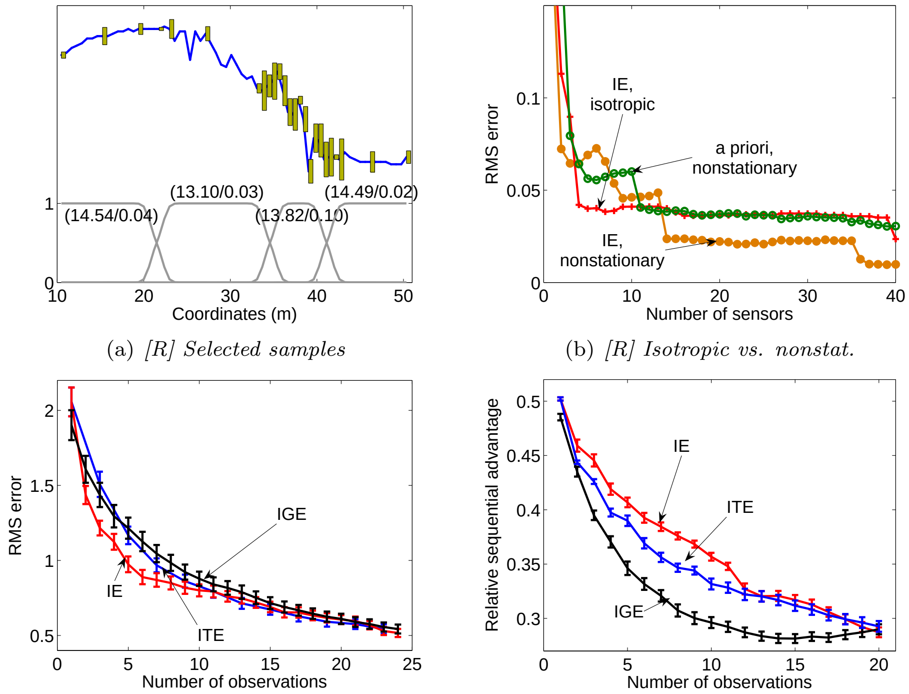
Computing entropy and mutual information. In order to implement the greedy policy for mutual information π_{GMI} or entropy π_{GH}, we need to be able to compute $H(\mathcal{X}_s \mid \mathcal{X}_\mathcal{A}, \theta)$ for the location s under consideration, and a set of observations $\mathcal{A}$ (or $\mathcal{V} \setminus (\mathcal{A} \cup \{s\})$ for mutual information). We can compute this quantity very similarly to the procedure described above. We first find the regions relevant to s, $\mathcal{V}^{(i_1)}, \dots, \mathcal{V}^{(i_m)}$, and set $\mathcal{A}' = \mathcal{V}' \cap \mathcal{A}$, where $\mathcal{V}' = \mathcal{V}^{(i_1)} \cup \dots \cup \mathcal{V}^{(i_m)}$. As above, for every joint instantiation of the relevant parameters $\bar{\theta}$, we compute the conditional entropy on the GP $\mathcal{X}'_\mathcal{V}$, which we can do efficiently in closed form given the parameters $\bar{\theta}$. We can then compute $H(\mathcal{X}_s \mid \mathcal{X}_\mathcal{A} = \mathbf{x}_\mathcal{A}, \Theta) = \sum_{\bar{\theta}} \widehat{P}(\bar{\theta}) H(\mathcal{X}_s \mid \mathcal{X}_\mathcal{A} = \mathbf{x}_\mathcal{A}, \bar{\theta})$.

In summary, our active learning strategy for nonstationary GPs is similar to the isotropic case: We explore until Corollary 3 proves that the advantage of the sequential strategy is small enough, then switch to exploitation. The difference is that we use a variational approach to leverage the structure of the nonstationary GP as a linear combination of locally supported isotropic GPs.

7 Experiments

River Monitoring. We first describe results on our river monitoring application. We consider one high-resolution spatial scan of pH measurements from the NIMS sensor deployed just below the confluence of the San Joaquin and the Merced rivers in California (denoted by [R]) (Harmon et al., 2006). We partition the transect into four regions, with smoothing weights indicated in Figure 2(a), and we use 2 bandwidth and 5 noise variance levels. Figure 2(a) illustrates the samples chosen by implicit exploration (IE) using the entropy criterion. The bars indicate the sequence of observations, and larger bars correspond to later observations (i.e., based on more knowledge about the model). We can observe that while the initial samples are roughly uniformly distributed, the later samples are mostly chosen in the weakly correlated, high variance turbulent confluence region. In parentheses, we display the estimated bandwidths and noise standard deviations. Figure 2(b) presents the results from our algorithms. The sequential algorithm leads to a quicker decrease in Root Mean Squared (RMS) error than the a priori design. Initially, the isotropic model with two parameters provides a better fit than the nonstationary model with 8 parameters, but, after about 15 samples, the situation is inverted, and the nonstationary model drastically outperforms the isotropic model after 28 samples, providing more than 50% lower error.

[3] We assume here that the ν_i are supported in a small number of regions. If this is not the case, we can use truncation arguments similar to those by Guestrin et al. (2005).



(a) *[R] Selected samples*

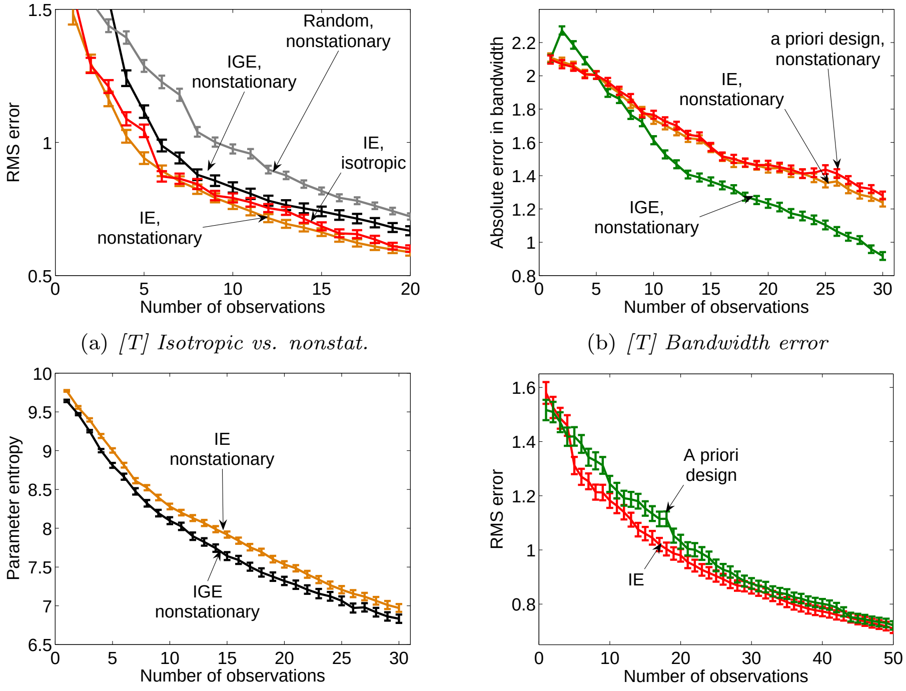
(b) *[R] Isotropic vs. nonstat.*

(c) *[T] Exploration strategies*

(d) *[T] Sequential advantage*

Figure 2: Results on pH [R] and temperature [T] data. (a) Top: sampling locations chosen by active learning algorithm. Higher bars indicate later (i.e., more informed) choice. Bottom: Smoothing functions used for spatial partitioning. (b) Comparison of prediction error for pH data. Note that the sequential algorithm on the nonstationary model eventually reduces the error incurred by the a priori design and isotropic model by more than 50%. (c) Comparison of exploration strategies, isotropic model. (d) Bounds on the potential advantage of the sequential algorithm using Theorem 3 (Stopping criterion). Information gain leads to quickest drop of bound, but worse spatial prediction.

Temperature Data. We consider temperature data [T] from a sensor network deployment with 54 sensors at Intel Research Berkeley. Our 145 samples consist of measurements taken every hour by the sensors over 5 days. We modeled the data as an isotropic process with unknown variance and an Exponential kernel with unknown bandwidth. We discretized the variance in $\sigma^2 \in \{1^2, 2^2, 3^2, 4^2, 5^2\}$, and the bandwidth in $\{3, 5, 7, 9, 11, 13, 15\}$ meters based on expert knowledge. We compared the performance of the active learning strategies, each using a different exploration strategy. Figure 2(c) shows the RMS prediction error, and Figure 2(d) presents the potential relative advantage obtained by Theorem 3 (our stopping criterion). While IE leads to the best prediction, followed by the independence test exploration (ITE), information gain exploration (IGE) tightens the bound on the sequential advantage the fastest. For example., if we decide to stop exploring once the sequential advantage drops below $\eta = 35\%$, 5 samples suffice for IGE, 8 for ITE and 12 for IE. This analysis (which is also supported by other data sets) indicates that none of the exploration strategies dominates each other, their differences can be well-characterized, and the choice of



(a) *[T] Isotropic vs. nonstat.*

(b) *[T] Bandwidth error*

(c) *[T] Parameter Entropy*

(d) *[P] Sequential vs. a priori*

Figure 3: Results on temperature [T] and precipitation [P] data. (a) Comparison of isotropic, nonstationary model, using random and sequential selection. Information gain achieves worst prediction, but reduces error in bandwidth (b) and parameter entropy (c) fastest. (d) Sequential design outperforms a priori design on rain data.

strategy depends on the needs of each application. Hence, if the goal is to switch to a priori design as quickly as possible, IGE might be the right choice, whereas if we can afford to always perform the logistically more complex sequential design, IE would decrease the predictive RMS error the fastest. ITE performs well w.r.t. both criteria, and has theoretical sample complexity guarantees.

We also modeled the temperature using a nonstationary GP, with the space partitioned into four regions, each modeled as an isotropic GP. We adopted a softmax function with smoothing bandwidth 8 meters to spatially average over the local isotropic GPs. The results in Figure 3(a) show that the nonstationary model leads to reduced prediction error compared to the isotropic model. All active learning models drastically outperform random selection. Since the parameter uncertainty is still very high after 20 samples, IGE leads to worse prediction accuracy than IE. However, IGE decreases the parameter error Figure 3(b) (compared to the estimates when given all observations) and parameter entropy $H(\Theta)$ Figure 3(c) the fastest. These results indicate (along with higher log-likelihood), that even though we are estimating its 8 parameters from only up to 20 data points, the nonstationary model provides a better fit to the data.

Precipitation Data. In another experiment, we considered precipitation data [P] from 167 detector stations in the Pacific Northwest. We followed the preprocessing suggested by Guestrin et al. (2005). Figure 3(d) shows the RMS error for 110 samples, spaced roughly three months apart, using an isotropic GP with 5 bandwidth and 3 variance parameter levels. Here, IE, ITE, IGE all outperform the a priori design.

8 Conclusions

In this paper, we presented a nonmyopic analysis for active learning of Gaussian Processes. We proved bounds on how much better a sequential algorithm can perform than an a priori design when optimizing observation locations under unknown parameters. Our bounds show that key potential for improvement is in the parameter entropy, motivating an exploration–exploitation approach to active learning, and provide insight into when to switch between the two phases. Using submodularity of our objective function, we provided bounds on the quality of our exploitation strategy. We proposed several natural exploration strategies for decreasing parameter uncertainty, and proved logarithmic sample complexity results for exploration phase using hypothesis testing. We extended our algorithm to handle nonstationary GP, exploiting local structure in the model. Here, we used a variational approach to address the combinatorial growth of the parameter space. In addition to our theoretical analyses, we evaluated our algorithms on several real-world problems, including data from a real deployment for monitoring the ecological condition of a river. We believe that our results provide significant new insights on the potential of sequential active learning strategies for monitoring spatial phenomena using GPs.

References

Balcan, N., Beygelzimer, A., & Langford, J. (2006). Agnostic active learning. *ICML*.

Caselton, W., & Zidek, J. (1984). Optimal monitoring network designs. *Statist. Prob. Lett.*, *2*, 223–227.

Castro, R., Willett, R., & Nowak, R. (2005). Faster rates in regression via active learning. *NIPS*.

Cover, T. M., & Thomas, J. A. (1991). *Elements of information theory*. Wiley Interscience.

Cressie, N. A. (1991). *Statistics for spatial data*. Wiley.

Dasgupta, S. (2005). Coarse sample complexity bounds for active learning. *NIPS*.

Gramacy, R. B. (2005). *Bayesian treed Gaussian process models*. Doctoral dissertation, University of California.

Gretton, A., Borgwardt, K., Rasch, M., Schlkopf, B., & Smola, A. (2006). A kernel method for the two-sample-problem. *NIPS*.

Guestrin, C., Krause, A., & Singh, A. (2005). Near-optimal sensor placements in gaussian processes. *ICML*.

Harmon, T. C., Ambrose, R. F., Gilbert, R. M., Fisher, J. C., Stealey, M., & Kaiser, W. J. (2006). *High resolution river hydraulic and water quality characterization using rapidly deployable networked infomechanical systems (nims rd)* (Technical Report 60). CENS.

Ko, C., Lee, J., & Queyranne, M. (1995). An exact algorithm for maximum entropy sampling. *Ops Res*, *43*.

Koller, D., & Friedman, N. (2007). *Structured probabilistic models*. Electronic Preprint.

Nemhauser, G., Wolsey, L., & Fisher, M. (1978). An analysis of the approximations for maximizing submodular set functions. *Mathematical Programming*, *14*, 265–294.

Nott, D. J., & Dunsmuir, W. T. M. (2002). Estimation of nonstationary spatial covariance structure. *Biomet.*, *89*.

Ong, C., Smola, A., & Williamson, R. (2005). Learning the kernel with hyperkernels. *JMLR*, *6*, 1043–1071.

Paciorek, C. (2003). *Nonstationary gaussian processes for regression and spatial mod.* Doctoral dissertation, CMU.

Rasmussen, C. E., & Williams, C. K. (2006). *Gaussian process for machine learning*. MIT Press.

Seo, S., Wallat, M., Graepel, T., & Obermayer, K. (2000). Gaussian process regression: Active data selection and test point rejection. *IJCNN* (pp. 241–246).

Shewry, M., & Wynn, H. (1987). Maximum entropy sampling. *Journal of Applied Statistics*, *14*, 165–170.

Storkey, A. J. (99). Truncated covariance matrices and toeplitz methods in gaussian processes. *ICANN*.

Zhu, Z., & Stein, M. L. (2006). Spatial sampling design for prediction with estimated parameters. *J Agric., Biol. Env. Statist.*, *11*, 24–49.

A Proofs

Lemma 10. *Let π be a policy. Then*

$$H(\mathcal{X}_\pi) \leq H(\mathcal{X}_\pi \mid \Theta) + H(\Theta),$$

$$\mathrm{MI}(\mathcal{X}_\pi) \leq \mathrm{MI}(\mathcal{X}_\pi \mid \Theta) + H(\Theta).$$

Proof. From definition, $\mathrm{MI}(\mathcal{X}_\pi) = H(\mathcal{X}_{\mathcal{V}\setminus\pi}) - H(\mathcal{X}_{\mathcal{V}\setminus\pi} \mid \mathcal{X}_\pi)$. Using the chain rule, we find $\mathrm{MI}(\mathcal{X}_\pi) = H(\mathcal{X}_\pi) - H(\mathcal{X}_\pi \mid \mathcal{X}_{\mathcal{V}\setminus\pi})$. Consider

$$\begin{aligned}
H(\mathcal{X}_\pi, \Theta) &= -\sum_\theta \int p(\mathbf{x}_\mathcal{V}, \theta) \log p(\mathbf{x}_\pi, \theta) d\mathbf{x}_\mathcal{V} \\
&= -\sum_\theta \int p(\mathbf{x}_\mathcal{V}, \theta)[\log p(\mathbf{x}_\pi) + \log p(\theta \mid \mathbf{x}_\pi)] d\mathbf{x}_\mathcal{V} \\
&= H(\mathcal{X}_\pi) + \int p(\mathbf{x}_\mathcal{V}) H(\Theta \mid \mathbf{x}_\pi) d\mathbf{x}_\mathcal{V} \\
&\geq H(\mathcal{X}_\pi).
\end{aligned}$$

Also,

$$\begin{aligned}
H(\mathcal{X}_\pi, \Theta) &= -\sum_\theta \int p(\mathbf{x}_\mathcal{V}, \theta)[\log p(\mathbf{x}_\pi \mid \theta) + \log p(\theta)] d\mathbf{x}_\mathcal{V} \\
&= H(\Theta) - \sum_\theta \int p(\mathbf{x}_\mathcal{V}) \log p(\mathbf{x}_\pi \mid \theta) d\mathbf{x}_\mathcal{V} \\
&= H(\Theta) + H(\mathcal{X}_\pi \mid \Theta).
\end{aligned}$$

Similarly, $H(\mathcal{X}_{\bar\pi}) \leq H(\mathcal{X}_{\bar\pi} \mid \Theta) + H(\Theta)$ (just replace π by $\bar\pi$ in the above proof).

Now consider

$$
\begin{aligned}
H(\mathcal{X}_{\bar{\pi}} \mid \mathcal{X}_{\pi}) &= -\int p(\mathbf{x}_{\mathcal{V}}) \log p(\mathbf{x}_{\bar{\pi}} \mid \mathbf{x}_{\pi}) d\mathbf{x}_{\mathcal{V}} \\
&= -\int p(\mathbf{x}_{\mathcal{V}}) \sum_{|A|=k} [\pi(\mathbf{x}_{\mathcal{V}}) = A] \log p(\mathbf{x}_{\bar{\mathcal{A}}} \mid \mathbf{x}_{\mathcal{A}}) d\mathbf{x}_{\mathcal{V}} \\
&= -\sum_{A} \int p(\mathbf{x}_{\mathcal{A}}) [\pi(\mathbf{x}_{\mathcal{A}}) = A] \int p(\mathbf{x}_{\bar{\mathcal{A}}} \mid \mathbf{x}_{\mathcal{A}}) \log p(\mathbf{x}_{\bar{\mathcal{A}}} \mid \mathbf{x}_{\mathcal{A}}) d\mathbf{x}_{\mathcal{V}} \\
&= \sum_{A} \int p(\mathbf{x}_{\mathcal{A}}) [\pi(\mathbf{x}_{\mathcal{A}}) = A] H(\mathcal{X}_{\bar{\mathcal{A}}} \mid \mathbf{x}_{\mathcal{A}}) d\mathbf{x}_{\mathcal{A}} \\
&\geq \sum_{A} \int p(\mathbf{x}_{\mathcal{A}}) [\pi(\mathbf{x}_{\mathcal{A}}) = A] H(\mathcal{X}_{\bar{\mathcal{A}}} \mid \mathbf{x}_{\mathcal{A}}, \Theta) d\mathbf{x}_{\mathcal{A}} \\
&= \sum_{A} \int p(\mathbf{x}_{\mathcal{A}}) [\pi = A] \sum_{\theta} P(\theta \mid \mathbf{x}_{\mathcal{A}}) H(\mathcal{X}_{\bar{\mathcal{A}}} \mid \mathbf{x}_{\mathcal{A}}, \theta) d\mathbf{x}_{\mathcal{A}} \\
&= -\sum_{A,\theta} \int p(\mathbf{x}_{\mathcal{A}}, \theta) [\pi = A] p(\mathbf{x}_{\bar{\mathcal{A}}} \mid \mathbf{x}_{\mathcal{A}}, \theta) \log p(\mathbf{x}_{\bar{\mathcal{A}}} \mid \mathbf{x}_{\mathcal{A}}, \theta) d\mathbf{x}_{\mathcal{V}} \\
&= -\sum_{A,\theta} \int p(\mathbf{x}_{\mathcal{V}}, \theta) [\pi = A] \log p(\mathbf{x}_{\bar{\mathcal{A}}} \mid \mathbf{x}_{\mathcal{A}}, \theta) d\mathbf{x}_{\mathcal{V}} \\
&= -\sum_{\theta} p(\theta) \int p(\mathbf{x}_{\mathcal{V}} \mid \theta) \log p(\mathbf{x}_{\bar{\pi}} \mid \mathbf{x}_{\pi}, \theta) d\mathbf{x}_{\mathcal{V}} \\
&= \sum_{\theta} P(\theta) H(\mathcal{X}_{\bar{\pi}} \mid \mathcal{X}_{\pi}, \theta) \\
&= H(\mathcal{X}_{\bar{\pi}} \mid \mathcal{X}_{\pi}, \Theta)
\end{aligned}
$$

Hence,

$$
\begin{aligned}
\mathrm{MI}(\mathcal{X}_{\pi}) &= H(\mathcal{X}_{\bar{\pi}}) - H(\mathcal{X}_{\bar{\pi}} \mid \mathcal{X}_{\pi}) \\
&\leq H(\Theta) + H(\mathcal{X}_{\bar{\pi}} \mid \Theta) - H(\mathcal{X}_{\bar{\pi}} \mid \mathcal{X}_{\pi}, \Theta) \\
&= \mathrm{MI}(\mathcal{X}_{\pi} \mid \Theta) + H(\Theta).
\end{aligned}
$$

□

Proof of Theorem 1. Using Lemma 10, it suffices to show that

$$\max_{|\pi|=k} H(\mathcal{X}_{\pi} \mid \Theta) \leq \sum_{\theta} P(\theta) \max_{|A|=k} H(\mathcal{X}_{\mathcal{A}} \mid \theta),$$

and

$$\max_{|\pi|=k} \mathrm{MI}(\mathcal{X}_{\pi} \mid \Theta) = \sum_{\theta} P(\theta) \max_{|A|=k} \mathrm{MI}(\mathcal{X}_{\mathcal{A}} \mid \theta).$$

This follows from the fact that $H(\mathcal{X}_\pi \mid \Theta) = \sum_\theta P(\theta) H(\mathcal{X}_\pi \mid \theta) \leq \sum_\theta P(\theta) \max_{|\mathcal{A}|=k} H(\mathcal{X}_\mathcal{A} \mid \theta)$. Similarly, $\mathrm{MI}(\mathcal{X}_\pi \mid \Theta) = \sum_\theta P(\theta) \, \mathrm{MI}(\mathcal{X}_\pi \mid \theta) \leq \sum_\theta P(\theta) \max_{|\mathcal{A}|=k} \mathrm{MI}(\mathcal{X}_\mathcal{A} \mid \theta)$.

□

Proof of Corollary 3. We first observe that submodularity is closed under taking expectations, and $\mathrm{MI}(\mathcal{X}_{\mathcal{A} \cup y} \mid \Theta) - \mathrm{MI}(\mathcal{X}_\mathcal{A} \mid \Theta) \geq \sum P(\theta)[-\varepsilon] = -\varepsilon$ shows that $\mathrm{MI}(\cdot \mid \Theta)$ is ε-nondecreasing. The result follows from combining the statements of Theorem 2 and Theorem 1. □

Proof of Proposition 8. Consider the sequence $H_i = H(\Theta \mid \mathcal{X}_{\pi_{1:i}})$. The information never hurts principle (Cover & Thomas, 1991) shows that $\mathbb{E}[H_{i+1} \mid \mathcal{X}_{\pi_{1:i}}] \leq H_i$ with probability 1 over the observations $\mathcal{X}_{\pi_{1:i}}$. Hence $(H_i)_i$ is a super-martingale, which proves that $H(\Theta \mid \mathcal{X}_{\pi_{1:i}}) = \mathbb{E}[H_i] \leq \mathbb{E}[H_0] = H(\Theta)$. □

Lemma 11. *Let $\hat{\Theta}$ be the parameter identified by binary search procedure. Let $M = \lceil \log m \rceil$. If $P(T_i \neq I_i \mid \theta) \leq \alpha$, then for all θ, $P(\hat{\Theta} = \theta \mid \Theta = \theta) \geq 1 - M\alpha$.*

Proof of Lemma 11.

$$\begin{aligned} P(\hat{\Theta} = \theta \mid \Theta = \theta) &= P(T^{(1)} = I^{(1)} \wedge . . \wedge T^{(M)} = I^{(M)} \mid \theta) \\ &= 1 - P(T^{(1)} \neq I^{(1)} \vee \cdots \vee T^{(M)} \neq I^{(M)} \mid \theta) \\ &\geq 1 - \sum_{i=1}^{M} P(T^{(i)} \neq I^{(i)} \mid \theta) \geq 1 - M\alpha, \end{aligned}$$

□

Proof of Theorem 5. Let C be a statistic of the T_i and I_i, such that $C = 1$ if $T_i = I_i$ for all i. From Lemma 11, we have that $P(C = 1 \mid \theta) \geq 1 - \alpha / \lceil \log_2 m \rceil$. Hence also unconditionally $P(C = 1) \geq 1 - \alpha / \lceil \log_2 m \rceil$. From Corollary 3 we have that if we know the correct parameter, $\mathrm{MI}(\mathcal{X}_{\mathcal{A}_G \cup \pi_H} \mid \Theta)] \geq (1 - 1/e) \max_{|\pi|=k} \mathrm{MI}(\mathcal{X}_\pi) - k\varepsilon$. This event happens with probability $P(C = 1)$. If $C = 0$, we have identified the wrong parameter. In this event, we have that $\mathrm{MI}(\mathcal{X}_{\mathcal{A}_G \cup \pi_H} \mid \Theta)] \geq (1 - 1/e) \sum_\theta P(\theta) \, \mathrm{MI}(\mathcal{X}_{\mathcal{A}^\theta} \mid \theta) - k\varepsilon + H(\Theta \mid \mathcal{X}_{\pi_H})$. But $H(\Theta \mid \mathcal{X}_{\pi_H}) \leq \lceil \log_2 m \rceil$. Hence

$$\begin{aligned} \mathbb{E}_T[\, \mathrm{MI}(\mathcal{X}_{\mathcal{A}_G \cup \pi_H} \mid \Theta)] &\geq (1 - 1/e) \max_{|\pi|=k} \mathrm{MI}(\mathcal{X}_\pi) - k\varepsilon - \\ &- \left(0 \cdot P(C = 1) + \lceil \log_2 m \rceil P(C = 0) (\max_\theta |\, \mathrm{MI}(\mathcal{X}_{\mathcal{A}_G \cup \pi_H} \mid \Theta) - \mathrm{MI}(\mathcal{X}_{\mathcal{A}^\theta} \mid \theta)|) \right). \end{aligned}$$

□

Proof of Theorem 6. This asymptotic bound follows directly from confidence intervals obtained after applying Fisher's transform to the sample correlation coefficient. □

Lemma 12. *Let $\mathcal{X}$ be an isotropic GP with variance σ^2, $\mathcal{X}_s, \mathcal{X}_t$ and $\mathcal{X}_\mathcal{A}$ be such that $\mathrm{Cor}(\mathcal{X}_s, \mathcal{X}_z) \leq \varepsilon$ and $\mathrm{Cor}(\mathcal{X}_t, \mathcal{X}_z) \leq \varepsilon$ for all $z \in \mathcal{A}$, and assume every location has independent measurement noise (nugget) σ_n^2. Then $|\mathrm{Cor}(\mathcal{X}_s, \mathcal{X}_t) - \mathrm{Cor}(\mathcal{X}_s, \mathcal{X}_t \mid \mathcal{X}_\mathcal{A})| \leq 2 \, |\mathcal{A}| \, \varepsilon^2 \frac{\sigma^2}{\sigma_n^2}$.*

Proof. By induction on $|\mathcal{A}|$. First divide by σ, such that the process has unit variance. This does not change the correlation. Let $\mathcal{X}_z \in \mathcal{A}$. Then

$$\Sigma_{xy,xy|Z} = \Sigma_{xy,xy} - \Sigma_{xy,z}\sigma_{z,z}^{-1}\Sigma_{z,xy}.$$

Hence

$$||\Sigma_{xy,xy|Z} - \Sigma_{xy,xy}||_\infty = ||\Sigma_{xy,z}\sigma_{z,z}^{-1}\Sigma_{z,xy}||_\infty \leq 2\varepsilon^2 \frac{\sigma^2}{\sigma_n^2}.$$

The induction step uses the independence of the measurement noise, such that $\sigma^2_{z|\mathcal{A}'} \geq \sigma^2_n$. □

Proof of Corollary 7. This is a direct consequence of Theorem 6 and Lemma 12, by observing that by ignoring n ε-correlated samples, the correlation does not change by more than $2n\varepsilon^2\frac{\sigma^2}{\sigma_n^2}$. Since the correlation for θ_i cannot increase by and the correlation for θ_{i+1} cannot decrease by more than that amount, the gap in correlation decreases by at most $4n\varepsilon^2\frac{\sigma^2}{\sigma_n^2}$. Requiring this to be less than ξ and solving for ε proves the claim. □

Proof of Theorem 4. Let $M = \max_A \max_{\theta_1,\theta_2} \frac{\mathrm{MI}(A|\theta_1)}{\mathrm{MI}(A|\theta_2)} < \infty$. Also let $K = \max_\theta \mathrm{MI}(A_\theta \mid \theta)$, where $A_\theta = \mathrm{argmax}_A \, \mathrm{MI}(A \mid \theta)$. Let $A^* = \mathrm{argmax}_A \, \mathrm{MI}(A \mid \Theta)$. Let θ' be the most likely parameter, and assume $H(\Theta) < 1$, which implies $\delta < \frac{1}{2}$.

Let us show the lower bound first. Clearly, $\mathrm{MI}(\mathcal{X}_{\pi^*}) \geq \max_{\mathcal{A}} \mathrm{MI}(\mathcal{X}_{\mathcal{A}})$. Now, $\mathrm{MI}(A) = H(A) - H(A \mid V \setminus A) \geq H(A \mid \Theta) - H(A, \Theta \mid V \setminus A)$, since Θ is discrete. But $H(A, \Theta \mid V \setminus A) = H(A \mid V \setminus A, \Theta) + H(\Theta \mid V \setminus A) \leq H(A \mid V \setminus A, \Theta) + H(\Theta)$, hence the lower bound follows. Note that if $\mathcal{A}$ is "small" compared to $\mathcal{V} \setminus \mathcal{A}$, then we can expect $H(\Theta \mid V \setminus A) \approx 0$.

$$\begin{aligned}
\mathrm{MI}(A^* \mid \Theta) &= (1-\delta)\,\mathrm{MI}(A^* \mid \theta') + \delta\,\mathrm{MI}(A^* \mid \neg\theta') \\
&\leq (1-\delta)\,\mathrm{MI}(A^* \mid \theta') + M\delta\,\mathrm{MI}(A^* \mid \theta') \\
&= (1 + (M-1)\delta)\,\mathrm{MI}(A^* \mid \theta'),
\end{aligned}$$

where

$$\mathrm{MI}(A^* \mid \neg\theta') = \frac{1}{\delta}\sum_{\Theta \neq \theta'} P(\theta)\,\mathrm{MI}(A^* \mid \theta).$$

We have that $\mathrm{MI}(A^* \mid \Theta) \geq \mathrm{MI}(A_{\theta'} \mid \Theta)$. Hence

$$\begin{aligned}
\mathrm{MI}(A^* \mid \theta') &\geq \frac{1}{1 + (M-1)\delta}\,\mathrm{MI}(A_{\theta'} \mid \Theta) \\
&\geq \frac{1-\delta}{1 + (M-1)\delta}\,\mathrm{MI}(A_{\theta'} \mid \theta'),
\end{aligned}$$

where we use that $\mathrm{MI}(A_{\theta'} \mid \neg\theta') \geq 0$.

Now define the loss

$$L(A) = \sum_\theta P(\theta)\left[\max_{A'} \mathrm{MI}(A' \mid \theta) - \mathrm{MI}(A \mid \theta)\right].$$

This loss is minimized by $A^* = \operatorname{argmax}_A \mathrm{MI}(A \mid \Theta)$. Now,

$$\begin{aligned}
L(A^*) \leq & (1-\delta)\left[\mathrm{MI}(A_{\theta'} \mid \theta') - \mathrm{MI}(A^* \mid \theta')\right] \\
& + \delta(K - \mathrm{MI}(A^* \mid \neg\theta')) \\
\leq & (1-\delta)\left[\mathrm{MI}(A_{\theta'} \mid \theta') - \frac{1-\delta}{1+(M-1)\delta}\mathrm{MI}(A_{\theta'} \mid \theta')\right] \\
& + \delta(K - \mathrm{MI}(A^* \mid \neg\theta')) \\
= & \frac{(1-\delta)(M-1)\delta}{1+(M-1)\delta}\mathrm{MI}(A_{\theta'} \mid \theta') + \delta(K - \mathrm{MI}(A^* \mid \neg\theta')) \\
= & \delta\left[\frac{(1-\delta)(M-1)}{1+(M-1)\delta}\mathrm{MI}(A_{\theta'} \mid \theta') + K - \mathrm{MI}(A^* \mid \neg\theta')\right] \\
\leq & \delta\left[(M-1)\,\mathrm{MI}(A_{\theta'} \mid \theta') + K\right] \\
\leq & \delta M K,
\end{aligned}$$

where we use that $\mathrm{MI}(A^* \mid \neg\theta') \geq 0$. Hence $L(A^*)$ is $O(\delta)$. If $\delta < \frac{1}{2}$, then $\delta \leq \frac{H(\Theta)}{-\log_2 H(\Theta)}$.

We know

$$\mathrm{MI}(A^* \mid \Theta) - H(\Theta) \leq \mathrm{MI}(X_\pi) \leq \mathrm{MI}(A^* \mid \Theta) + H(\Theta) + L(A^*).$$

If we approximate $\mathrm{MI}(\mathcal{X}_{\pi^*})$ by $\mathrm{MI}(\mathcal{X}_{\mathcal{A}^*}|\Theta)$, the absolute error is hence bounded by

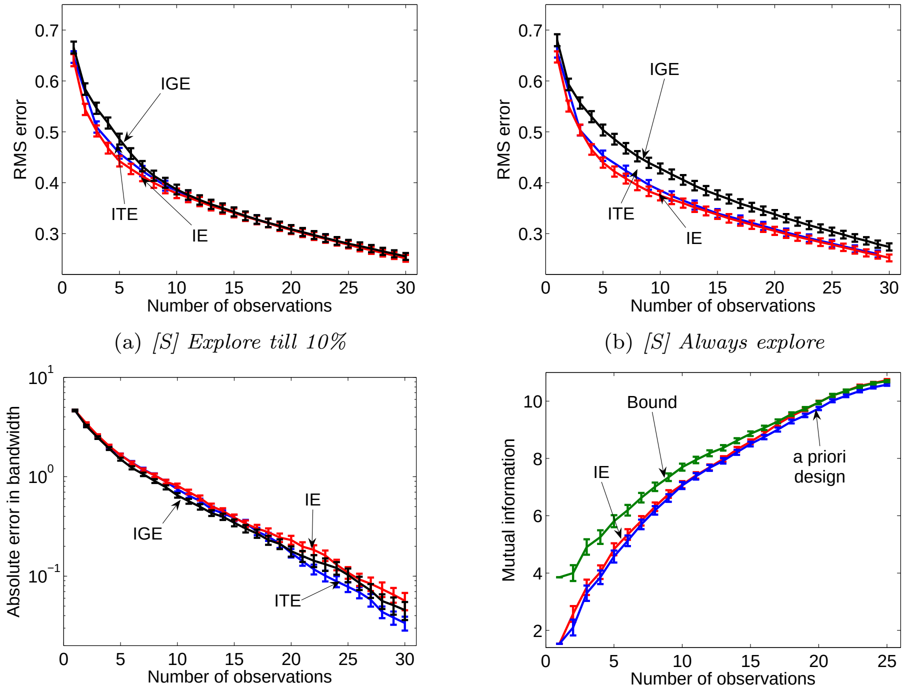
$$H(\Theta)\left(1 + \frac{MK}{\log_2 \frac{1}{H(\Theta)}}\right).$$

□

Proof of Proposition 9. This proposition follows from the fact that the fully factorized distribution is the maximum entropy distribution with a specified set of marginal distributions. □

B Results on synthetic data.

In order to study our method in more detail, we created a synthetic data set [S] out of 1000 samples from an isotropic GP with unit variance, on a 8×8 grid. We uniformly selected bandwidths for the Exponential kernel from $\{1, 2, 4, 8, 16\}$. For each sample, we ran our exploration-exploitation algorithm with all three exploration strategies (ITE, IGE and IE), as well as the a priori design based on $\mathrm{MI}(\cdot \mid \Theta)$. Figure 4(a) shows the average Root Mean Squares (RMS) prediction error with increasing number of observations. Hereby, the exploration stopped after, based on Corollary 3, the sequential design could only perform at most 10% better than the a priori design. Information gain exploration (IGE) decreases the RMS error slightly more slowly. In this experiment, the a priori design achieved prediction accuracy only insignificantly worse than the sequential designs. When we never stop exploring (*c.f.*, Figure 4(b)), then IGE performs significantly worse. The independence testing (ITE) performs almost as well as the implicit exploration (IE), since it prefers tests with high mutual information. Figure 4(c) shows the error in estimating the bandwidth parameter with an increasing number of samples. All three strategies decrease the parameter error exponentially, empirically demonstrating the logarithmic bound. Initially, IGE decreases the parameter error significantly more quickly. Both explicit exploration strategies, IGE and ITE, lead to a lower bandwidth error after all



(a) *[S] Explore till 10%*

(b) *[S] Always explore*

(c) *[S] Bandwidth error*

(d) *[S] Mutual information*

Figure 4: Results on synthetic data [S]. (a) If we stop exploring when the potential advantage goes below 10%, all 3 exploration strategies achieve approximately the same prediction. (b) If we never stop exploring, IGE leads to worse prediction than ITE. (c) Both IGE and ITE decrease parameter error more quickly than IE. (d) bound on optimal sequential performance quickly becomes tight, indicating that one can stop exploring early.

30 samples are observed. Figure 4(d) compares the mutual information achieved by the a priori and active strategies, along with the bound from Corollary 3 on the best sequential strategy (without the factor $(1-1/e)^{-1}$, since the greedy sets tend to be very close to optimal in practice (Guestrin et al., 2005)). Experimentally we observed that the contribution by the term $\sum_\theta P(\theta) \max_{|A|=k} \mathrm{MI}(\mathcal{X}_\mathcal{A} \mid \theta)$ to the stopping criterion was negligible compared to the parameter entropy $H(\Theta)$.