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

**Andreas Krause** KRAUSEA@CS.CMU.EDU
**Carlos Guestrin** GUESTRIN@CS.CMU.EDU
Carnegie Mellon University, Pittsburgh, PA, USA

## 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. We also present extensive empirical evaluation on several real-world problems.

## 1. Introduction

When monitoring spatial phenomena, such as the ecological condition of a river as in Figure 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 observed values, and hence nothing is lost by committing to sampling locations in advance. This is logistically simpler, since optimization can be carried without waiting for observations. 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.

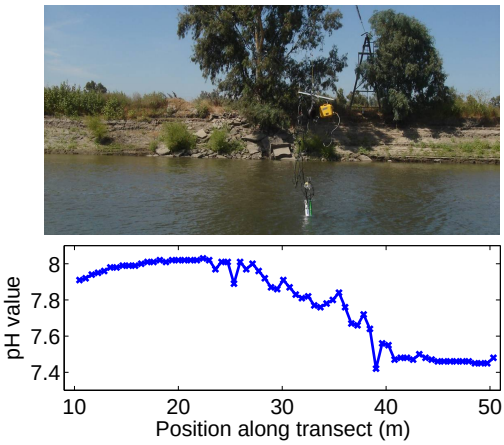



*Figure 1.* Top: Active sampling using the NIMS sensor (Harmon et al., 2006) deployed at Merced River, CA. The sensor can perform horizontal and vertical traversal. Bottom: Samples of pH acquired along horizontal transect.

Our main theoretical result is a bound which quantifies the performance difference between sequential and

Appearing in *Proceedings of the 24$^{th}$ International Conference on Machine Learning*, Corvallis, OR, 2007. 

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, each of which has unique advantages. 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 isotropic 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.

## 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* $\mathcal{K}(\cdot,\cdot)$. For each random variable $\mathcal{X}_u$ with index $u \in \mathcal{V}$, its mean $\mu_u$ is given by $\mathcal{M}(u)$, and for each pair of indices $u, v \in \mathcal{V}$, their covariance $\sigma_{uv}$ is given by $\mathcal{K}(u, v)$. 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:

$$\mu_{y|\mathcal{A}} = \mu_y + \Sigma_{y\mathcal{A}}\Sigma_{\mathcal{A}\mathcal{A}}^{-1}(\mathbf{x}_\mathcal{A} - \mu_\mathcal{A}), \quad (2.1)$$

$$\sigma^2_{y|\mathcal{A}} = \mathcal{K}(y, y) - \Sigma_{y\mathcal{A}}\Sigma_{\mathcal{A}\mathcal{A}}^{-1}\Sigma_{\mathcal{A}y}, \quad (2.2)$$

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, or even *isotropic*, which means that the covariance only depends on the distance between locations, i.e., $\mathcal{K}(u, v) = \mathcal{K}_\theta(||u - v||_2)$, where $\theta$ is a set of parameters. 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 $\theta$. We adopt a hierarchical Bayesian approach and assign a prior $P(\theta)$ to the parameters $\theta$, 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, we also assume that the prior mean function $\mathcal{M}(\cdot)$ is zero. This assumption can be relaxed, e.g., by assigning a normal prior to the mean function.

## 3. Observation Selection Policies

**Entropy.** The *entropy* criterion has been frequently used to select observations in GPs (*c.f.*, Seo et al. (2000); Shewry and Wynn (1987)). Here, we select 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 $i$ the location $y_{i+1} = \operatorname{argmax}_{y \in \mathcal{V} \setminus \mathcal{A}_i} H(\mathcal{X}_{y_{i+1}} \mid \mathcal{X}_{\mathcal{A}_i})$ is added to the set of already selected locations $\mathcal{A}_i$.

This a priori greedy rule can be readily turned into a *sequential* algorithm, selecting $y_{i+1} = \operatorname{argmax}_{y \in \mathcal{V} \setminus \mathcal{A}_i} H(\mathcal{X}_{y_{i+1}} \mid \mathcal{X}_{\mathcal{A}_i} = \mathbf{x}_{\mathcal{A}_i})$. Now, the selected location $y_{i+1}$ depends on the observations $\mathbf{x}_{\mathcal{A}_i}$. 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 $\pi$ can select a *different* set of observations $\pi(\mathbf{x}_\mathcal{V}) \subseteq \mathcal{V}$. Hereby, the $i$-th element, $\pi_i$, 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 $\pi_H$ to our river example, $\pi_{H,i}$ 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 $\pi$ 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 $\pi$ 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 $\pi$ 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 $\pi$. The greedy policy $\pi_{IE}$ for mutual information, after some algebraic manipulation, is given by:

$$\pi_{i+1} = \operatorname{argmax}_y H(\mathcal{X}_y \mid \mathcal{X}_{\pi_{1:i}} = \mathbf{x}_{\pi_{1:i}}) - H(\mathcal{X}_y \mid \mathcal{X}_{\mathcal{V}\setminus\{y,\pi_{1:i}\}}), \tag{3.2}$$

where $\pi_{1:i} \equiv \pi_{1:i}(\mathbf{x}_\mathcal{V})$ are the first $i$ locations selected by $\pi$ when $\mathcal{X}_\mathcal{V} = \mathbf{x}_\mathcal{V}$. Hence, $\pi_{IE}$ selects the location $\pi_{i+1}$ which is uncertain given the previous observations ($H(\mathcal{X}_y | \mathcal{X}_{\pi_{1:i}} = \mathbf{x}_{\pi_{1:i}})$ is large) and relevant to the unobserved locations ($H(\mathcal{X}_y \mid \mathcal{X}_{\mathcal{V}\setminus\{y,\pi_{1:i}\}})$ is small).

## 4. Bounds on the Advantage of Active Learning Strategies

A key question in active learning is to determine the potential for 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 $\theta$ 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 $\sigma^2_{\mathcal{X}_y | \mathcal{X}_\mathcal{A}}$ is given by Eq. (2.2). Thus, the entropy of a set of variables does not depend on the 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, such as mutual information, cannot benefit from sequential strategies. Note that for non-Gaussian models, sequential strategies can strictly outperform a priori designs, even with known parameters.

In the case of GPs 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. The following central result formalizes this intuition, by bounding $H(\mathcal{X}_\pi)$ (and similarly for mutual information) of the optimal *policy* $\pi$ by a mixture of entropies of *sets* $H(\mathcal{X}_\mathcal{A} \mid \theta)$, which are chosen optimally for each fixed parameter (and can thus be selected a priori):

**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 can be found in (Krause & Guestrin, 2007). Theorem 1 bounds the advantage of sequential designs by two components: The expected advantage of 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 $\theta$, 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. Near-optimal Exploitation

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. They showed, that, if the parameter $\Theta = \theta$ is known, the a priori greedy algorithm, which starts with the empty set $\mathcal{A} = \emptyset$ and iteratively adds the element $s = \operatorname{argmax} \mathrm{MI}(\mathcal{X}_\mathcal{A} \cup \{\mathcal{X}_s\})$ until $k$ elements have been selected, finds a near-optimal set. Their result uses the concept of *submodularity*, an intuitive diminishing returns property: a new observation decreases our uncertainty more if we know less. Due to space limitations, we refer the reader to (Guestrin et al., 2005) for details. Combining their Theorem 6 with our Theorem 1, we have the following result about exploitation using mutual information:

**Corollary 2.** *Under sufficiently fine discretization $\mathcal{V}$:*

$$\max_{|\pi|=k} \mathrm{MI}(\mathcal{X}_\pi) \leq \frac{e}{e-1} \sum_\theta P(\theta)\, \mathrm{MI}(\mathcal{X}^{(\theta)}_{\mathcal{A}_G} \mid \theta) + k\varepsilon + H(\Theta),$$

*where $\mathcal{A}^{(\theta)}_G$ is the greedy set for $\mathrm{MI}(\mathcal{X}_\mathcal{A} \mid \theta)$.*

Here, $\varepsilon$ depends polynomially on the discretization of $\mathcal{V}$. 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 2 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 $\eta$, i.e., we stop if

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

where $\mathcal{A}_G$ is the greedy set for $\mathrm{MI}(\mathcal{X}_\mathcal{A} \mid \Theta)$. Hereby, $\mathrm{MI}(\mathcal{X}_\mathcal{A} \mid \Theta) = \sum_\theta P(\theta)\, \mathrm{MI}(\mathcal{X}_\mathcal{A} \mid \theta)$. We can then 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 2 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 2 as a stopping criterion for exploration, we have to, for each parameter $\theta$, 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 3.** *Let $M = \max_A \max_{\theta_1,\theta_2} \frac{\mathrm{MI}(\mathcal{X}_A|\theta_1)}{\mathrm{MI}(\mathcal{X}_A|\theta_2)} < \infty$. Let $K = \max_\theta \max_\mathcal{A} \mathrm{MI}(\mathcal{X}_\mathcal{A} \mid \theta)$, $H(\Theta) < 1$. Then*

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

*where $\mathcal{A}^* = \operatorname{argmax}_\mathcal{A} \mathrm{MI}(\mathcal{X}_\mathcal{A} \mid \Theta)$ and $\pi^* = \operatorname{argmax}_\pi \mathrm{MI}(\mathcal{X}_\pi)$, 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 3 hence provides the following computational advantage over Corollary 2: once the parameter entropy is small enough, we do not need to compute the term $\sum_\theta P(\theta)\,\mathrm{MI}(\mathcal{X}_{\mathcal{A}_G}^{(\theta)} \mid \theta)$ determining the stopping point. Hence, in the following, we concentrate on directly decreasing the parameter uncertainty, $H(\Theta)$, as required by Theorem 3. 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 $\delta$ decreases exponentially with their distance $\delta$. Hence, there is an *exponentially large gap* between the correlation for bandwidths $\theta_i$ and $\theta_{i+1}$: There will be a distance $\hat{\delta}$, for which two random variables within this distance will appear dependent if the true bandwidth $\theta$ is at least $\theta \geq \theta_{i+1}$, and (roughly) independent if $\theta \leq \theta_i$. Our goal is to exploit this gap to efficiently determine the correct parameter.

Note that if we can separate $\theta_i$ from $\theta_{i+1}$, we effectively distinguish any $\theta_j$, for $j \leq i$, from $\theta_l$, for $l \geq i+1$, since the bandwidths scale the kernels. Let $I_i$ be a function of $\Theta$, such that $(I_i \mid \Theta) = 0$ if $\Theta \leq \theta_i$, and $(I_i \mid \Theta) = 1$ if $\Theta \geq \theta_{i+1}$. Assume we have tests $T_i$, using $\hat{N}$ samples, such that $P(T_i \neq I_i \mid \theta) \leq \alpha$ for all $\theta$. We can then perform *binary search* to find 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 4.** *Under the assumptions of Corollary 2 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 $\theta$: $P(T_i \neq I_i \mid \theta) \leq \alpha/(\lceil \log m \rceil^2 (\max_\theta |\mathrm{MI}(\mathcal{X}_{\pi_{G \circ ITE}} \mid \Theta) - \mathrm{MI}(\mathcal{X}_{\mathcal{A}_G^{(\theta)}} \mid \theta)|))$, then it holds that*

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

In order to make use of Theorem 4, we need to find tests $T_i$ such that $P(T_i \neq I_i \mid \theta)$ is sufficiently small for all $\theta$. 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 (1999)), 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 $\alpha$, 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 $\theta_i$ from those at least $\theta_{i+1}$. $\hat{\delta}_i$ 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 5.** *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 $\theta$.*

In the case of kernels with non-compact support, such as the Gaussian or Exponential kernel, we cannot generate such independent samples, since distant points will have some (exponentially small) correlation. However, it can be shown that these almost independent samples still suffice to get logarithmic sample complexity bounds (Krause & Guestrin, 2007).

Note that while this discussion 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; guaranteeing that with a small number of samples we can decrease the parameter uncertainty enough such that, using Theorem 3 as stopping criterion, we can switch to exploitation.

**Explicit Exploration based on Information Gain (IGE).** As the bound in Theorem 3 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 $\pi_{i+1}$ 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).** 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. Using an argument presented in (Krause & Guestrin, 2007), it can be shown that, in expectation, IE *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. Hence, often a parametric form is chosen. In this case, Corollary 2 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. An example of a parametric form for nonstationary is given by Nott and Dunsmuir (2002), who suggest to model nonstationarity by a spatially varying linear combination of isotropic processes.

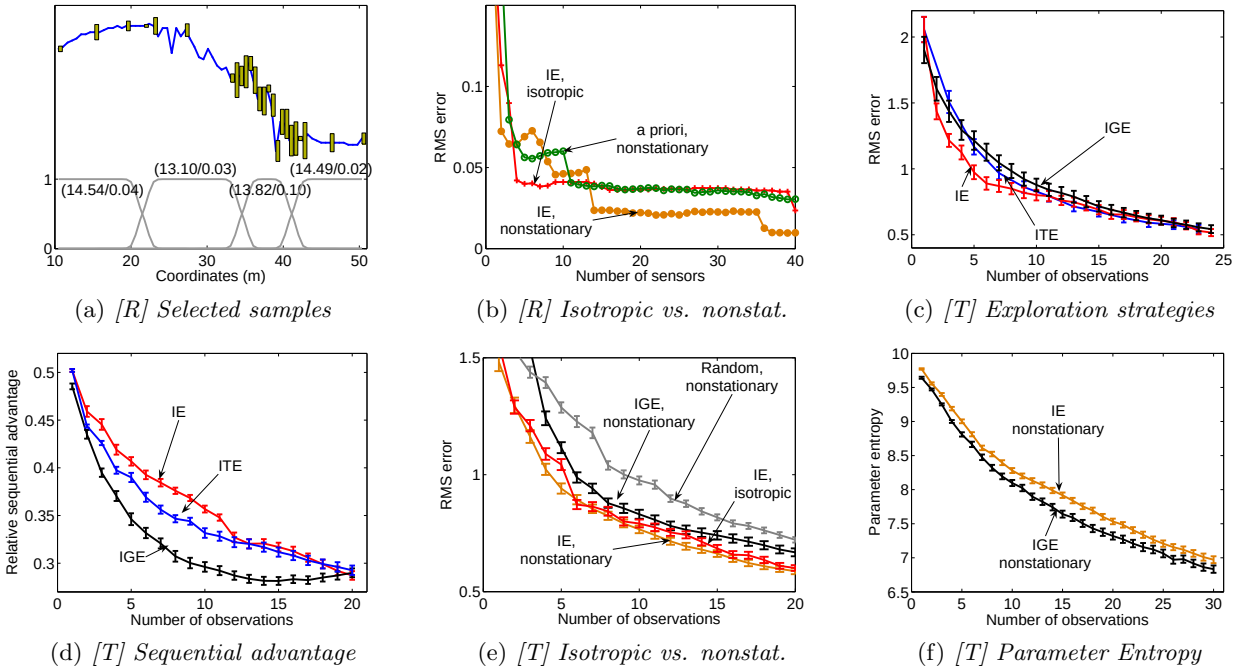
Motivated by the river monitoring problem, we partition the space into disjoint regions $\mathcal{V}^{(1)}, \ldots, \mathcal{V}^{(m)}$, specified by the user. With each region $\mathcal{V}^{(i)}$, we associate a stationary process $\mathcal{X}_{\mathcal{V}}^{(i)}$, with parameters $\Theta^{(i)}$, which 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. We want to choose the weights $\lambda_i(s)$ such that the model behaves similar to process $\mathcal{X}_{\mathcal{V}}^{(i)}$ within region $i$, and interpolates smoothly between regions. Hence, we associate a locally supported weighting function $\nu_i(s)$ with each region, which achieves its maximum value in region $i$ and decreases 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 $\sum_i \lambda_i(s)\lambda_i(t)\,\mathcal{K}_i(s,t)$. Note that, contrary to the GP Mixture of Experts approach (*c.f.*, Tresp (2000)), where the marginal distributions are mixtures of Gaussians, our construction uses a *linear combination* of GPs, hence, for fixed parameters, our nonstationary model is still a 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.

Note, that if the number of regions $m$ is large, the (discretized) joint distribution $\Theta$ requires exponentially many parameters. In (Krause & Guestrin, 2007), we describe a variational, KL-divergence minimizing approach which allows efficient approximate inference in this model, by finding a factorized approximation to the posterior $P(\Theta \mid \mathcal{X}_{\mathcal{A}} = \mathbf{x}_{\mathcal{A}})$ after observations $\mathcal{X}_{\mathcal{A}} = \mathbf{x}_{\mathcal{A}}$ have been made. There it is also shown that even under this variational approach, Corollary 2 remains valid, and hence exploration never stops early.

We can apply Corollary 2 to this nonstationary model in order to determine when to switch from exploration to exploitation. While it is not clear how to generalize the hypothesis testing (ITE) approach to the nonstationary setting, the information gain exploration (IGE) can be readily applied. Hence, our active learning strategy for nonstationary GPs is similar to the stationary case: We explore until Corollary 2 proves that the advantage of the sequential strategy is small enough, and then we switch to exploitation.

## 7. Experiments

**River Monitoring.** We consider one high-resolution spatial scan of pH measurements from the NIMS sensor (Figure 1) 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




(a) *[R] Selected samples* (b) *[R] Isotropic vs. nonstat.* (c) *[T] Exploration strategies*

(d) *[T] Sequential advantage* (e) *[T] Isotropic vs. nonstat.* (f) *[T] Parameter Entropy*

*Figure 2.* Results on pH [R] and temperature [T] data. (a) Top: sampling locations chosen by active learning algorithm. Larger bars indicate later, more informed, choice. Bottom: Smoothing functions for spatial partitioning. (b) IE on the nonstationary model eventually achieves 50% lower prediction error. (c,e) IE decreases prediction error fastest, but the bound (d) on the potential advantage of the sequential algorithm (Corollary 2) and parameter entropy (f) slowest.

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, providing more than 50% lower error.

**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 Corollary 2 (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 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 with respect to both criteria, and has theoretical sample complexity guarantees.

We also modeled the temperature data 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 2(e) 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 parameter entropy $H(\Theta)$ (Figure 2(f)) the fastest, which is consistent with the isotropic case.

## 8. Related Work

Previous work on active learning in GPs mostly focused on the case where the model is completely specified (Seo et al., 2000; Guestrin et al., 2005). Gramacy (2005) presents a nonstationary, hierarchical Bayesian GP approach based on spatial partitioning, where the spatial decomposition is integrated out using MCMC methods, which however does not provide any theoretical performance guarantee. Zhu and Stein (2006) present an approach for a priori design for spatial prediction in the case of unknown parameters. While their criterion accomodates parameter uncertainty, they do not consider the benefit of sequential over a priori designs. The potential of active learning for improving sample complexity has been studied (*c.f.*, Balcan et al. (2006)), however these approaches usually make assumption which do not apply to GPs, e.g., the availability of i.i.d. samples. Castro et al. (2005) provide a near-optimal algorithm for learning Hoelder-smooth functions; their method however does not apply in the case of learning a GP. There is also a large body of work on kernel learning (*c.f.*, Ong et al. (2005)). We are however unaware of any results on nonmyopic sample complexity guarantees in this area. Gretton et al. (2006) proposed a kernel based hypothesis testing approach which could potentially be used for exploration in the nonstationary setting.

## 9. 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 mutual information, 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 the 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.

### Acknowledgements

This work was supported by NSF grant CNS-0509383 and a gift from Intel Corporation. Carlos Guestrin was supported in part by an IBM Faculty Fellowship, and an Alfred P. Sloan Fellowship. We would like to thank Amarjeet Singh and Prof. William Kaiser for providing data and an image of the NIMS system.

## 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.

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.

Krause, A., & Guestrin, C. (2007). *Nonmyopic active learning of gaussian processes: An exploration–exploitation approach* (Techn. Report CMU-ML-07-105).

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*.

Tresp, V. (2000). Mixtures of gaussian processes. *NIPS* (pp. 654–660).

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