Chapter 2
Preliminaries

2.1 Mathematical Notation

Standard notation is used throughout this book. Let $\mathbb{R}, \mathbb{R}_{\geq 0}, \mathbb{R}_{>0}, \mathbb{Z}, \mathbb{Z}_{\geq 0}, \mathbb{Z}_{>0}$ denote the sets of real numbers, nonnegative real numbers, positive real numbers, integers, nonnegative integers, and positive integers, respectively.

Let $\mathbb{E}$, Var, Corr, Cov denote the expectation, variance, correlation, and the covariance operators, respectively.

Let $\mathbf{A}^T \in \mathbb{R}^{M\times N}$ be the transpose of a matrix $\mathbf{A} \in \mathbb{R}^{N\times M}$. Let $\mathrm{tr}(\mathbf{A})$ and $\det(\mathbf{A})$ denote the trace and the determinant of a matrix $\mathbf{A} \in \mathbb{R}^{N\times N}$, respectively. Let $\mathrm{row}_i(\mathbf{A}) \in \mathbb{R}^M$ and $\mathrm{col}_j(\mathbf{A}) \in \mathbb{R}^N$ denote the ith row and the jth column of a matrix $\mathbf{A} \in \mathbb{R}^{N\times M}$, respectively.

The positive definiteness and the positive semi-definiteness of a square matrix $\mathbf{A}$ are denoted by $\mathbf{A} \succ 0$ and $\mathbf{A} \succeq 0$, respectively.

Let $|x|$ denote the absolute value of a scalar x. Let $\|\mathbf{x}\|$ denote the standard Euclidean norm (2-norm) of a vector $\mathbf{x}$. The induced 2-norm of a matrix $\mathbf{A}$ is denoted by $\|\mathbf{A}\|$. Let $\|\mathbf{x}\|_\infty$ denote the infinity norm of a vector $\mathbf{x}$.

Let $\mathbf{1}$ denote the vector with all elements equal to one and $\mathbf{I}$ denote the identity matrix with an appropriate size. Let $\mathbf{e}_i$ be the standard basis vector of appropriate size with 1 as its ith element and 0 on all other elements.

The symbol $\otimes$ denotes the Kronecker product. The symbol $\circ$ denotes the Hadamard product (also known as the entry-wise product and the Schur product).

A random vector $\mathbf{x}$, which is distributed by a normal distribution of mean $\boldsymbol{\mu}$ and covariance matrix $\mathbf{C}$, is denoted by $\mathbf{x} \sim \mathbb{N}(\boldsymbol{\mu}, \mathbf{C})$. The corresponding probability density function is denoted by $\mathbb{N}(\mathbf{x}; \boldsymbol{\mu}, \mathbf{C})$.

The relative complement of a set $\mathcal{A}$ in a set $\mathcal{B}$ is denoted by $\mathcal{B} \setminus \mathcal{A} := \mathcal{B} \cap \mathcal{A}^c$, where $\mathcal{A}^c$ is the complement of $\mathcal{A}$. For a set $\mathcal{A} \in \mathcal{I}$, we define $z_\mathcal{A} = \{z_i \mid i \in \mathcal{A}\}$. Let $-\mathcal{A}$ denote the set $\mathcal{I} \setminus \mathcal{A}$.

Y. Xu et al., *Bayesian Prediction and Adaptive Sampling Algorithms for Mobile Sensor Networks*, SpringerBriefs in Control, Automation and Robotics, DOI 10.1007/978-3-319-21921-9_2

An undirected graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ is a tuple consisting of a set of vertices $\mathcal{V} := \{1, \cdots, n\}$ and a set of edges $\mathcal{E} \subset \mathcal{V} \times \mathcal{V}$. The neighbors of $i \in \mathcal{V}$ in $\mathcal{G}$ are denoted by $\mathcal{N}_i := \{j \in \mathcal{V} \mid \{i, j\} \in \mathcal{E}\}$.

Other notations will be explained in due course.

2.2 Physical Process Model

In this section, we review important notions for the Gaussian process which will be used to model the physical phenomenon. In particular, we introduce a class of spatiotemporal Gaussian process model with anisotropic covariance functions. The properties of Gaussian Markov random fields (GMRF) are also briefly reviewed.

2.2.1 Gaussian Process

A Gaussian process can be thought of a generalization of a Gaussian distribution over a finite vector space to function space of infinite dimension. It is formally defined as follows [53, 91]:

Definition 2.1 A Gaussian process (GP) is a collection of random variables, any finite number of which have a consistent[1] joint Gaussian distribution.

A Gaussian process, denoted by

$$z(\mathbf{x}) \sim \mathcal{GP}\left(\mu(\mathbf{x}), C(\mathbf{x}, \mathbf{x}'; \boldsymbol{\theta})\right) \tag{2.1}$$

is completely specified by its mean function $\mu(\mathbf{x})$ and covariance function $C(\mathbf{x}, \mathbf{x}'; \boldsymbol{\theta})$ which are defined as

$$\begin{aligned} \mu(\mathbf{x}) &= \mathbb{E}\left[z(\mathbf{x})\right], \\ C(\mathbf{x}, \mathbf{x}'; \boldsymbol{\theta}) &= \mathbb{E}\left[\left(z(\mathbf{x}) - \mu(\mathbf{x})\right)\left(z(\mathbf{x}') - \mu(\mathbf{x}')\right)|\boldsymbol{\theta}\right]. \end{aligned}$$

Although not needed to be done, we take the mean function to be zero for notational simplicity,[2] i.e., $\mu(\mathbf{x}) = \mathbf{0}$. If the covariance function $C(\mathbf{x}, \mathbf{x}'; \boldsymbol{\theta})$ is invariant with respect to translations in the input space, i.e., $C(\mathbf{x}, \mathbf{x}'; \boldsymbol{\theta}) = C(\mathbf{x} - \mathbf{x}'; \boldsymbol{\theta})$, we call it stationary. Furthermore, if the covariance function is a function of only the distance between the inputs, i.e., $C(\mathbf{x}, \mathbf{x}'; \boldsymbol{\theta}) = C(\left\|\mathbf{x} - \mathbf{x}'\right\|; \boldsymbol{\theta})$, then it is called isotropic.

[1] It is also known as the marginalization property. It means simply that the random variables obey the usual rules of marginalization, etc.

[2] This is not a drastic limitation since the mean of the posterior process is not confined to zero [53].

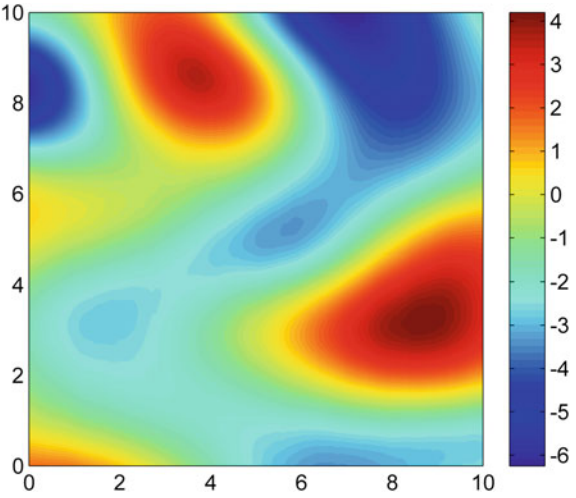

Fig. 2.1 Realization of a two-dimensional ($D = 2$) Gaussian process with $\sigma_f^2 = 5$, $\sigma_1 = 2.5$, and $\sigma_2 = 1.5$.

In practice, a parametric family of functions is used instead of fixing the covariance function [84]. One common choice of a stationary covariance function is

$$C(\mathbf{x}, \mathbf{x}'; \boldsymbol{\theta}) = \sigma_f^2 \exp\left\{ -\sum_{\ell=1}^{D} \frac{(x_\ell - x'_\ell)^2}{2\sigma_\ell^2} \right\}, \tag{2.2}$$

where x_ℓ is the ℓth element of $\mathbf{x} \in \mathbb{R}^D$. From (2.2), it can be easily seen that the correlation between two inputs decreases as the distance between them increases. This decreasing rate depends on the choice of the length scales $\{\sigma_\ell\}$. A very large length scale means that the predictions would have little bearing on the corresponding input which is then said to be insignificant. σ_f^2 gives the overall vertical scale relative to the mean of the Gaussian process in the output space. These parameters play the role of hyperparameters since they correspond to the hyperparameters in neural networks and in the standard parametric model. Therefore, we define $\boldsymbol{\theta} = (\sigma_f^2, \sigma_1, \cdots, \sigma_D)^T \in \mathbb{R}^{D+1}$ as the hyperparameter vector. A realization of a Gaussian process that is numerically generated is shown in Fig. 2.1.

2.2.2 Spatiotemporal Gaussian Process

In this section, spatiotemporal Gaussian processes are of particular interest. Spatiotemporal Gaussian processes are obtained as a special case of (2.1) by setting $\mathbf{x} \subset \mathbb{R}^D \times \mathbb{R}_{\geq 0}$, where $\mathbb{R}^D$ is for spatial locations and $\mathbb{R}_{\geq 0}$ is the temporal domain. A spatiotemporal Gaussian process can be written as

$$z(\mathbf{s}, t) \sim \mathcal{GP}(\mu(\mathbf{s}, t), C(\mathbf{s}, t, \mathbf{s}', t'; \boldsymbol{\theta})),$$

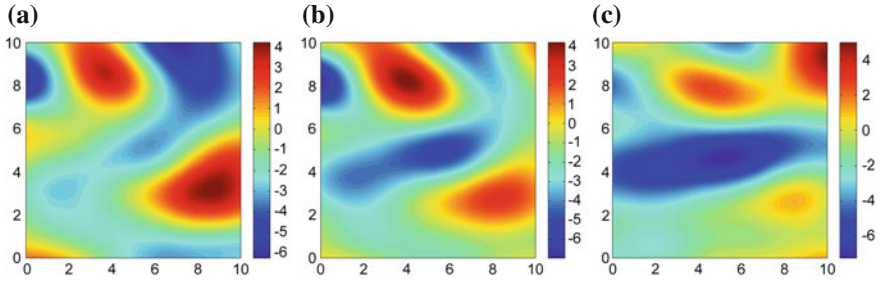

Fig. 2.2 Realization of a spatiotemporal ($D = 2$) Gaussian process with $\sigma_f^2 = 5$, $\sigma_1 = 2.5$, $\sigma_2 = 1.5$, and $\sigma_t = 8$ at **a** $t = 1$, **b** $t = 5$, and **c** $t = 10$.

where $\mathbf{x} = (\mathbf{s}^T, t)^T \in \mathbb{R}^D \times \mathbb{R}_{\geq 0}$. We consider the following generalized anisotropic covariance function $C(\mathbf{x}, \mathbf{x}'; \boldsymbol{\theta})$ with a hyperparameter vector $\boldsymbol{\theta} := (\sigma_f^2, \sigma_1, \cdots, \sigma_D, \sigma_t)^T \in \mathbb{R}^{D+2}$:

$$C(\mathbf{x}, \mathbf{x}'; \boldsymbol{\theta}) = \sigma_f^2 \exp\left(-\sum_{\ell=1}^{D} \frac{(s_\ell - s'_\ell)^2}{2\sigma_\ell^2}\right) \exp\left(-\frac{(t - t')^2}{2\sigma_t^2}\right), \tag{2.3}$$

where $\mathbf{s}, \mathbf{s}' \in \mathcal{Q} \subset \mathbb{R}^D$, $t, t' \in \mathbb{R}_{\geq 0}$. $\{\sigma_1, \cdots, \sigma_D\}$ and σ_t are kernel bandwidths for space and time, respectively. (2.3) shows that points close in the measurement space and time indices are strongly correlated and produce similar values. In reality, the larger temporal distance two measurements are taken with, the less correlated they become, which strongly supports our generalized covariance function in (2.3). This may also justify the truncation (or windowing) of the observed time series data to limit the size of the covariance matrix for reducing the computational cost. A spatially isotropic version of the covariance function in (2.3) has been used in [36]. A realization of a spatiotemporal Gaussian process that is numerically generated is shown in Fig. 2.2.

2.2.3 Gaussian Markov Random Field

The Gaussian Markov random field is formally defined as follows [92]:

Definition 2.2 (*GMRF, [92, Definition 2.1]*) A random vector $\mathbf{z} = (z_1, \cdots, z_N)^T \in \mathbb{R}^N$ is called a GMRF with respect to a graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ with mean $\boldsymbol{\mu}$ and precision matrix $\mathbf{Q} \succ 0$, if and only if its density has the form

$$\pi(\mathbf{z}) = \frac{|\mathbf{Q}|^{1/2}}{(2\pi)^{N/2}} \exp\left(-\frac{1}{2}(\mathbf{z} - \boldsymbol{\mu})^T \mathbf{Q}(\mathbf{z} - \boldsymbol{\mu})\right),$$

and $(\mathbf{Q})_{ij} \neq 0 \Leftrightarrow \{i, j\} \in \mathcal{E}$ for all $i \neq j$, where the precision matrix (or information matrix) $\mathbf{Q} = \mathbf{C}^{-1}$ is the inverse of the covariance matrix $\mathbf{C}$, and $|\mathbf{Q}|$ denotes the determinant of $\mathbf{Q}$.

The Markov property of a GMRF can be shown by the following theorem.

Theorem 2.1 ([92, Theorem 2.4]) *Let* $\mathbf{z}$ *be a GMRF with respect to* $\mathcal{G} = (\mathcal{V}, \mathcal{E})$. *Then the followings are equivalent.*

1. *The pairwise Markov property:*

$$z_i \perp z_j \mid z_{-ij} \quad \text{if } \{i, j\} \notin \mathcal{E} \text{ and } i \neq j,$$

   where $\perp$ *denotes conditional independence and* $z_{-ij} := z_{-\{i,j\}} = z_{\mathcal{I}\setminus\{i,j\}}$. *This implies that* z_i *and* z_j *are conditionally independent given observations at all other vertices except* $\{i, j\}$ *if* i *and* j *are not neighbors.*
2. *The local Markov property:*

$$z_i \perp z_{-\{i,\mathcal{N}_i\}} \mid z_{\mathcal{N}_i} \quad \text{for every } i \in \mathcal{I}.$$

3. *The global Markov property:*

$$z_{\mathcal{A}} \perp z_{\mathcal{B}} \mid z_{\mathcal{C}}$$

   for disjoint sets $\mathcal{A}$, $\mathcal{B}$, *and* $\mathcal{C}$ *where* $\mathcal{C}$ *separates* $\mathcal{A}$ *and* $\mathcal{B}$, *and* $\mathcal{A}$ *and* $\mathcal{B}$ *are nonempty.*

If a graph $\mathcal{G}$ has small cardinalities of the neighbor sets, its precision matrix $\mathbf{Q}$ becomes sparse with many zeros in its entries. This plays a key role in computation efficiency of a GMRF which can be greatly exploited by the resource-constrained mobile sensor network. For instance, some of the statistical inference can be obtained directly from the precision matrix $\mathbf{Q}$ with conditional interpretations.

Theorem 2.2 ([92, Theorem 2.3]) *Let* $\mathbf{z}$ *be a GMRF with respect to* $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ *with mean* $\boldsymbol{\mu}$ *and precision matrix* $\mathbf{Q} \succ 0$, *then we have*

$$\begin{aligned}
\mathbb{E}(z_i \mid z_{-i}) &= \mu_i - \frac{1}{(\mathbf{Q})_{ii}} \sum_{j \in \mathcal{N}_i} (\mathbf{Q})_{ij}(z_j - \mu_j), \\
\mathrm{Var}(z_i \mid z_{-i}) &= \frac{1}{(\mathbf{Q})_{ii}}, \\
\mathrm{Corr}(z_i, z_j \mid z_{-ij}) &= -\frac{(\mathbf{Q})_{ij}}{\sqrt{(\mathbf{Q})_{ii}(\mathbf{Q})_{jj}}}, \quad \forall i \neq j.
\end{aligned}$$

2.3 Mobile Sensor Network

In this section, we explain the sensor network formed by multiple mobile sensing agents and present the measurement model used throughout the thesis.

Let N be the number of sensing agents distributed over the surveillance region $\mathcal{Q} \in \mathbb{R}^D$. The identity of each agent is indexed by $\mathcal{I} := \{1, 2, \cdots, N\}$. Assume that all agents are equipped with identical sensors and take noisy observations at time $t \in \mathbb{Z}_{>0}$. At time t, the sensing agent i takes a noise-corrupted measurement $y_i(t)$ at its current location $\mathbf{q}_i(t) \in \mathcal{Q}$, i.e.,

$$y_i(t) = z(\mathbf{q}_i(t), t) + \epsilon_i, \quad \epsilon_i \overset{i.i.d.}{\sim} \mathbb{N}(0, \sigma_w^2),$$

where the sensor noise ϵ_i is considered to be an independent and identically distributed Gaussian random variable. $\sigma_w^2 > 0$ is the noise level and we define the signal-to-noise ratio as

$$\gamma = \frac{\sigma_f^2}{\sigma_w^2}.$$

Notice that when a static field is considered, we have $z(\mathbf{s}, t) = z(\mathbf{s})$.

For notational simplicity, we denote the collection of positions of all N agents at time t as $\mathbf{q}(t)$, i.e.,

$$\mathbf{q}(t) := \left(\mathbf{q}_1(t)^T, \cdots, \mathbf{q}_N(t)^T\right)^T \in \mathcal{Q}^N.$$

The collective measurements from all N mobile sensors at time t are denoted by

$$\mathbf{y}_t := (y_1(t), \cdots, y_N(t))^T \in \mathbb{R}^N.$$

The cumulative measurements from time $t \in \mathbb{Z}_{>0}$ to time $t' \in \mathbb{Z}_{>0}$ are denoted by

$$\mathbf{y}_{t:t'} := \left(\mathbf{y}_t^T, \cdots, \mathbf{y}_{t'}^T\right)^T \in \mathbb{R}^{N(t'-t+1)}.$$

The communication network of mobile agents can be represented by an undirected graph. Let $\mathcal{G}(t) := (\mathcal{I}, \mathcal{E}(t))$ be an undirected communication graph such that an edge $(i, j) \in \mathcal{E}(t)$ if and only if agent i can communicate with agent $j \neq i$ at time t. We define the neighborhood of agent i at time t by $\mathcal{N}_i(t) := \{j \in \mathcal{I} \mid (i, j) \in \mathcal{E}(t)\}$. Similarly, let $\mathbf{q}^{[i]}(t)$ denote the vector form of the collection of positions in $\left\{\mathbf{q}_j(t) \mid j \in \{i\} \cup \mathcal{N}_i(t)\right\}$. Let $\mathbf{y}_t^{[i]}$ denote vector form of the collection of observations in $\left\{y(\mathbf{q}_j(t), t) \mid j \in \{i\} \cup \mathcal{N}_i(t)\right\}$. The cumulative measurements of agent i from time t to time t' are denoted as $\mathbf{y}_{t:t'}^{[i]}$.

2.4 Gaussian Processes for Regression

Suppose we have a dataset $\mathcal{D} = \left\{(\mathbf{x}^{(i)}, y^{(i)}) \mid i = 1, \cdots, n\right\}$ collected by mobile sensing agents where $\mathbf{x}^{(i)}$ denotes an input vector of dimension D and $y^{(i)}$ denotes a scalar value of the noise-corrupted output. The objective of probabilistic regression is to compute the predictive distribution of the function values $z_* := z(\mathbf{x}_*)$ at some test input $\mathbf{x}_*$.

For notational simplicity, we define the design matrix $\mathbf{X}$ of dimension $n \times D$ as the aggregation of n input vectors (i.e., $\text{row}_i(\mathbf{X}) := (\mathbf{x}^{(i)})^T$), and the outputs are collected in a vector $\mathbf{y} := (y^{(1)}, \cdots, y^{(n)})^T$. The corresponding vector of noise-free outputs is defined as $\mathbf{z} := (z(\mathbf{x}^{(1)}), \cdots, z(\mathbf{x}^{(n)}))^T$.

The advantage of the Gaussian process formulation is that the combination of the prior and noise models can be carried out exactly via matrix operations [93]. The idea of Gaussian process regression is to place a GP prior directly on the space of functions without parameterizing the function $z(\cdot)$, i.e.,

$$\pi(\mathbf{z}|\boldsymbol{\theta}) = \mathbb{N}(\mathbf{z}; \boldsymbol{\mu}, \mathbf{K}),$$

where $\boldsymbol{\mu} \in \mathbb{R}^n$ is the mean vector obtained by $(\boldsymbol{\mu})_i = \mu(\mathbf{x}^{(i)})$, and $\mathbf{K} := \text{Cov}(\mathbf{z}, \mathbf{z}|\boldsymbol{\theta}) \in \mathbb{R}^{n \times n}$ is the covariance matrix obtained by $(\mathbf{K})_{ij} = C(\mathbf{x}^{(i)}, \mathbf{x}^{(j)}; \boldsymbol{\theta})$. Notice that the GP model and all expressions are always conditional on the corresponding inputs. In the following, we will always neglect the explicit conditioning on the input matrix $\mathbf{X}$.

The inference in the Gaussian process model is as follows. First, we assume a joint GP prior $\pi(\mathbf{z}, z_*|\boldsymbol{\theta})$ over functions, i.e.,

$$\pi(\mathbf{z}, z_*|\boldsymbol{\theta}) = \mathbb{N}\left(\begin{bmatrix} \boldsymbol{\mu} \\ \mu(\mathbf{x}_*) \end{bmatrix}, \begin{bmatrix} \mathbf{K} & \mathbf{k} \\ \mathbf{k}^T & C(\mathbf{x}_*, \mathbf{x}_*; \boldsymbol{\theta}) \end{bmatrix}\right), \tag{2.4}$$

where $\mathbf{k} := \text{Cov}(\mathbf{z}, z_*|\boldsymbol{\theta}) \in \mathbb{R}^n$ is the covariance between $\mathbf{z}$ and z_* obtained by $(\mathbf{k})_i = C(\mathbf{x}^{(i)}, \mathbf{x}_*; \boldsymbol{\theta})$. Then, the joint posterior is obtained using Bayes rule, i.e.,

$$\pi(\mathbf{z}, z_*|\boldsymbol{\theta}, \mathbf{y}) = \frac{\pi(\mathbf{y}|\mathbf{z})\pi(\mathbf{z}, z_*|\boldsymbol{\theta})}{\pi(\mathbf{y}|\boldsymbol{\theta})},$$

where we have used $\pi(\mathbf{y}|\mathbf{z}, z_*) = \pi(\mathbf{y}|\mathbf{z})$. Finally, the desired predictive distribution $\pi(z_*|\boldsymbol{\theta}, \mathbf{y})$ is obtained by marginalizing out the latent variables in $\mathbf{z}$, i.e.,

$$\begin{aligned} \pi(z_*|\boldsymbol{\theta}, \mathbf{y}) &= \int \pi(\mathbf{z}, z_*|\theta, \mathbf{y}) d\mathbf{z} \\ &= \frac{1}{\pi(\mathbf{y}|\boldsymbol{\theta})} \int \pi(\mathbf{y}|\mathbf{z})\pi(\mathbf{z}, z_*|\boldsymbol{\theta}, \mathbf{y}) d\mathbf{z}. \end{aligned} \tag{2.5}$$

Since we have the joint Gaussian prior given in (2.4) and

$$\mathbf{y}|\mathbf{z} \sim \mathbb{N}\left(\mathbf{z}, \sigma_w^2\mathbf{I}\right),$$

the integral in (2.5) can be evaluated in closed-form and the predictive distribution turns out to be Gaussian, i.e.,

$$z_*|\boldsymbol{\theta}, \mathbf{y} \sim \mathbb{N}\left(\mu_{z_*|\boldsymbol{\theta},\mathbf{y}}, \sigma^2_{z_*|\boldsymbol{\theta},\mathbf{y}}\right), \tag{2.6}$$

where

$$\mu_{z_*|\boldsymbol{\theta},\mathbf{y}} = \mu(\mathbf{x}_*) + \mathbf{k}^T(\mathbf{K} + \sigma_w^2\mathbf{I})^{-1}(\mathbf{y} - \boldsymbol{\mu}), \tag{2.7}$$

and

$$\sigma^2_{z_*|\boldsymbol{\theta},\mathbf{y}} = C(\mathbf{x}_*, \mathbf{x}_*; \boldsymbol{\theta}) - \mathbf{k}^T(\mathbf{K} + \sigma_w^2\mathbf{I})^{-1}\mathbf{k}. \tag{2.8}$$

For notational simplicity, we define the covariance matrix of the noisy observations as $\mathbf{C} := \mathrm{Cov}(\mathbf{y}, \mathbf{y}|\boldsymbol{\theta}) = \mathbf{K} + \sigma_w^2\mathbf{I}$.

Springer

http://www.springer.com/978-3-319-21920-2

Bayesian Prediction and Adaptive Sampling Algorithms for Mobile Sensor Networks
Online Environmental Field Reconstruction in Space and Time

Xu, Y.; Choi, J.; Dass, S.; Maiti, T.

2016, XII, 115 p. 43 illus., 2 illus. in color., Softcover
ISBN: 978-3-319-21920-2