Entropy **1999**, *1*, 69-98

entropy
ISSN 1099-4300
www.mdpi.org/entropy/

A Bayesian Reflection on Surfaces

David R. Wolf

PO 8308, Austin, TX 78713-8308, USA
E-mail: drwolf@realtime.net

Received: 20 September 1999 / Accepted 20 October 1999 / Published: 30 October 1999

Abstract: The topic of this paper is a novel Bayesian continuous-basis field representation and inference framework. Within this paper several problems are solved: The maximally informative inference of continuous-basis fields, that is where the basis for the field is itself a continuous object and not representable in a finite manner; the tradeoff between accuracy of representation in terms of information learned, and memory or storage capacity in bits; the approximation of probability distributions so that a maximal amount of information about the object being inferred is preserved; an information theoretic justification for multigrid methodology. The maximally informative field inference framework is described in full generality and denoted the Generalized Kalman Filter. The Generalized Kalman Filter allows the update of field knowledge from previous knowledge at any scale, and new data, to new knowledge at any other scale. An application example instance, the inference of continuous surfaces from measurements (for example, camera image data), is presented.

Keywords: Bayesian inference; Generalized Kalman filter; Kalman filter; Kullback-Leibler distance; Maximally informative statistical inference; Knowledge representation; MinimumDescription Length; Sufficient statistics; Multigrid methods; Adaptive scale inference; Adaptive grid inference; Mutual information.

© 1999 by the author. Reproduction for noncommercial purposes permitted.

1 Overview

The paper begins by reviewing traditional approaches to surface representation and inference. Then the new field representation and inference paradigm is introduced within the context of maximally informative (MI) inference [5], early ideas appearing in [4]. The knowledge representation distribution is introduced and discussed in the context of MI inference. Then, using the MI inference approach, the here-named Generalized Kalman Filter (GKF) equations are derived for a specific example instance of inferring a surface height field. The GKF equations motivate a location-dependent adaptive scale or multigrid approach to the MI inference of continuous-basis fields.

2 Introduction: Surface representation

2.1 Traditional methods

Many methods for representing surfaces have been utilized previously, however these methods involve representing the surface by a *discrete* basis field, perhaps with a deterministic interpolation defined (bi-linear, tensor B-splines, etc.) to provide a definition for the surface at points intermediate to the discrete field. Probability distributions or densities of these discrete fields then often take the form of normalized exponentials of sums of clique energy functions, and produce a construct commonly known as a Markov Random Field. (See Geman [2], for an often cited example.) There are several immediate observations on these approaches:

- The surface remains unspecified at points intermediate to the discrete field, except by the often undefined notion of interpolation.
- When interpolation is *not* defined, the discrete field probability distribution says nothing about the probability distribution of surface at points intermediate to the discrete field points.
- When interpolation is defined then, given a value of the discrete field, there is *no uncertainty* in the surface intermediate to the discrete field points. There is a deterministic mapping from any given discrete field to the corresponding continuous surface. In particular, when the discrete field basis covers a fixed grid on the (x, y) plane with z heights at each grid point, known here as a height field, all sampling of the surface intermediate to the fixed grid is determined at the scale of the fixed grid. This is generally not physical, see next.
- The surface distribution is not an intrinsic property of any physical surface, rather a post-hoc imposition of the analyst attempting a useful regularization. For instance, necessary

scaling properties are ignored: Moving a camera closer to the surface, for example, so that the density of sample points on the physical surface increases, is not properly represented in the fixed basis of the discrete field distribution; there is no consistency imposed that requires a subsampled set of points to have the same probability density that one would find by marginalizing the surface distribution over the sample points not in the subsampling.

2.2 Scaling consistency

The consistency condition mentioned in the last section, which must be imposed on probability distributions for continuous fields is:

Scaling of sample points consistency: For $S \subset A$ indices of discrete field variables,

$$P(X_S) = \int P(X_A)\, dX_{A\setminus S} \tag{1}$$

Note that equation 1 is a condition which must be imposed on the distributions which any modelling system learns where it is sensible to supersample or subsample the field arbitrarily, as in the continuous field basis case.

2.3 Elements of the paradigm

The rest of this paper discusses an approach to continuous field inference which corrects the deficiencies, including the intermediate value and scaling problems, of traditional discrete-basis approaches to the inference of discrete height fields, for example. The new approach is here named the *Generalized Kalman Filter*.

There are four central objects of importance within the inference approach described in this paper, one of which is a new object to Bayesian inference:

- The **prior** distribution for field. The prior holds all information about fields before any data is observed.
- The **likelihood** distribution. The likelihood is predictive for data, given the field. It incorporates all of the physics of the measurement process.
- The **posterior** distribution. The posterior distribution summarizes everything knowable about the field given assumptions of likelihood form, the prior knowledge, and all data.
- The **knowledge-representation** (KR) distribution. Within the usual Bayesian point of view, the KR distribution is the new mathematical object. In the paradigm described in this paper the KR distribution is the object updated when new data arrives. The KR

distribution is parameterized by *maximally informative statistics* (see [5]) for the *learned* field knowledge. Note that because the KR distribution has a finite number-of-values limitation, the KR distribution is *not necessarily* able to represent what *could have been learned* from data about the (continuous) field. Generally, the prior distribution and the KR distribution determine an approximation (possibly exact) to the field posterior distribution. It should be noted that modern computer architecture (memory and space-time) constraints appear to be the fundamental physical drivers for the utilization of the KR distribution, simply because storing the exact posterior generally requires an infinite amount of memory.

In the height field inference application discussed later the KR distribution is parameterized by heights at a set of discrete basis points, but holds knowledge about a continuous basis height field. However, generally, the KR distribution may use an arbitrary set of basis functions.

One advance of the GKF is that the KR distribution is naturally adaptive in both dimension and scale, allowing the learning of continuous-basis field information at the appropriate scale, where appropriate.

Benefits of the approach described in this paper are that it has these information theoretically optimal features: 1. A location-dependent adaptive and scalable multigrid-like algorithm, so that only the bytes necessary to represent the learned information are stored, leading to a style of maximally sparse representation of surface knowledge; 2. A recursive updating algorithm. It will become clear that the Bayesian GKF field inference paradigm also has these properties:

- It is the *information learned* about the field, (the KR distribution), which takes the form of a distribution over discrete values. In the surface inference example these discrete values are heights at discrete basis points.

- The prior distribution for fields, in conjunction with the learned knowledge of the field held within the KR distribution determine a well-defined posterior distribution over continuous fields.

- The field posterior distribution is always a well defined quantity everywhere. In the surface inference example discussed later, this continuity is at points intermediate to the discrete height field basis points of the KR distribution.

- The scaling condition equation 1 is automatically imposed because the posterior distribution is a distribution over *fields*.

As an example consider the inference of continuous surfaces: While it may seem obvious, in the case of continuous surface inference, that what one is actually representing with a discrete set of

values in memory is only a part of the information which helps to determine the surface posterior distribution, it is unusual to *not* be discussing the height field as the primary representation of surface. It is the inherently discrete nature of the storage of information in machines which forces us into this stance - *generally* it is impossible to represent an arbitrary *continuous* field with a finite set of discrete values - one must also have another object from which to compute the intermediate values of the field. (Another way to look at the disparity between the current proposal for field inference and traditional proposals is that the traditional approaches are sufficient only for band-limited fields.)

In section 3 the GKF is specialized to height fields, where an example, surface representation and learning, of the GKF paradigm is described. (The approach taken in this section is to specialize to a case that is then easily seen to generalize to the general continuous basis field inference paradigm.) The next section continues with observations on the update scheme. Further sections continue with the example special case for surface distributions with particularly tractable mathematics, and final sections provide explicit forms for the general GKF equations, a discussion on their relationship to the standard Kalman filter, a discussion on the amount of information learned at each update, and a search heuristic. Extensive appendices provide supporting mathematics for the derivations.

3 Surface representation and inference

In this section the main ideas of the Bayesian surface representation and inference paradigm presented in this paper are given. The technique is general, though: section 4 discusses the extension to an arbitrary-basis, arbitrary-dimension field.

3.1 Surface distributions

The surface and height field distributions (the prior, likelihood, and posterior surface and height field distributions) are discussed in this section.

3.1.1 Surface and height field prior distributions

Consider a set S of surfaces where each element $\boldsymbol{s} \in S$ is a height field, i.e. such that $\boldsymbol{s} = s(x, y)$ is real function of two variables. Write the prior probability distribution for surfaces in S given the parameters θ which determine the prior distribution as

$$P(\boldsymbol{s} \mid \theta). \tag{2}$$

Consider a vector $\boldsymbol{v} = (v_1, \ldots, v_n)$ of discrete (x, y) points, $v_i = (x_i, y_i)$. For any given surface $\boldsymbol{s}$ denote the associated vector of heights by $\boldsymbol{h}(\boldsymbol{s}, \boldsymbol{v}) = (h_1(\boldsymbol{s}, \boldsymbol{v}), \ldots, h_n(\boldsymbol{s}, \boldsymbol{v}))$. Write the prior

distribution of the surface heights at the chosen points $\boldsymbol{v}$ as $P(\boldsymbol{h}_v \mid \theta)$. This discrete height distribution may be found as follows:

$$\begin{aligned} P(\boldsymbol{h}_v \mid \theta) &= \int P(\boldsymbol{h}_v \mid \boldsymbol{s}, \theta)\, P(\boldsymbol{s} \mid \theta)\, d\boldsymbol{s} && (3) \\ &= \int P(\boldsymbol{h}_v \mid \boldsymbol{s})\, P(\boldsymbol{s} \mid \theta)\, d\boldsymbol{s} && (4) \\ &= \int \boldsymbol{\delta}(\boldsymbol{h}_v - \boldsymbol{h}(\boldsymbol{s}, \boldsymbol{v}))\, P(\boldsymbol{s} \mid \theta)\, d\boldsymbol{s} && (5) \end{aligned}$$

where the vector delta-function is defined as

$$\boldsymbol{\delta}(\boldsymbol{h}_v - \boldsymbol{h}(\boldsymbol{s}, \boldsymbol{v})) = \Pi_{i=1}^{n} \delta(h_{v,i} - h_i(\boldsymbol{s}, \boldsymbol{v})) \quad (6)$$

Now, given that what is known is the surface heights $\boldsymbol{h}_v$ at a vector $\boldsymbol{v}$ of discrete (x, y) points, the posterior distribution of surfaces is found from Bayes' theorem as

$$\begin{aligned} P(\boldsymbol{s} \mid \boldsymbol{h}_v, \theta) &= \frac{P(\boldsymbol{h}_v \mid \boldsymbol{s}, \theta)\, P(\boldsymbol{s} \mid \theta)}{P(\boldsymbol{h}_v \mid \theta)} && (7) \\ &= \frac{P(\boldsymbol{h}_v \mid s)\, P(\boldsymbol{s} \mid \theta)}{P(\boldsymbol{h}_v \mid \theta)} && (8) \\ &= \frac{\boldsymbol{\delta}(\boldsymbol{h}_v - \boldsymbol{h}(\boldsymbol{s}, \boldsymbol{v}))\, P(\boldsymbol{s} \mid \theta)}{\int \boldsymbol{\delta}(\boldsymbol{h}_v - \boldsymbol{h}(\boldsymbol{s}, \boldsymbol{v}))\, P(\boldsymbol{s} \mid \theta)\, d\boldsymbol{s}} && (9) \end{aligned}$$

where the denominator distribution was found in equation 5.

3.1.2 Measurements: The Likelihood

In general, a surface $\boldsymbol{s}$ and some other parameters ϕ not dependent upon $\boldsymbol{s}$ (i.e. camera point spread function, camera position and direction, lighting position and direction, etc.) specify the probability distribution for data (likelihood)

$$P(\boldsymbol{x} \mid \boldsymbol{s}, \phi, \theta) = P(\boldsymbol{x} \mid \boldsymbol{s}, \phi) \quad (10)$$

where the data distribution is independent of θ once $\boldsymbol{s}$ is known.

3.1.3 Conditioning on data: Surface and height field posterior distributions

Given data, the surface posterior distribution is inferred using Bayes' theorem as

$$\begin{aligned} P(\boldsymbol{s} \mid \boldsymbol{x}, \phi, \theta) &= \frac{P(\boldsymbol{x} \mid \boldsymbol{s}, \phi, \theta)\, P(\boldsymbol{s} \mid \phi, \theta)}{P(\boldsymbol{x} \mid \phi, \theta)} && (11) \\ &= \frac{P(\boldsymbol{x} \mid \boldsymbol{s}, \phi)\, P(\boldsymbol{s} \mid \theta)}{\int P(\boldsymbol{x} \mid \boldsymbol{s}, \phi)\, P(\boldsymbol{s} \mid \theta)\, d\boldsymbol{s}} && (12) \end{aligned}$$

The distribution of the surface posterior marginalized to a set of discrete points may be written using equations 11–12, doing steps similar to those taken in equations 3–5, as

$$P(\boldsymbol{h}_v \mid \boldsymbol{x}, \phi, \theta) = \int P(\boldsymbol{h}_v \mid \boldsymbol{s}, \boldsymbol{x}, \phi, \theta)\, P(\boldsymbol{s} \mid \boldsymbol{x}, \phi, \theta)\, d\boldsymbol{s} \tag{13}$$

$$= \int P(\boldsymbol{h}_v \mid \boldsymbol{s})\, P(\boldsymbol{s} \mid \boldsymbol{x}, \phi, \theta)\, d\boldsymbol{s} \tag{14}$$

$$= \int \boldsymbol{\delta}(\boldsymbol{h}_v - \boldsymbol{h}(\boldsymbol{s}, \boldsymbol{v}))\, P(\boldsymbol{s} \mid \boldsymbol{x}, \phi, \theta)\, d\boldsymbol{s} \tag{15}$$

In steps similar to equations 7–9 the surface posterior when a height field is also known is given by

$$P(\boldsymbol{s} \mid \boldsymbol{h}_v, \boldsymbol{x}, \phi, \theta) = \frac{P(\boldsymbol{h}_v, \boldsymbol{x} \mid \boldsymbol{s}, \phi, \theta)\, P(\boldsymbol{s} \mid \phi, \theta)}{P(\boldsymbol{h}_v, \boldsymbol{x} \mid \phi, \theta)} \tag{16}$$

$$= \frac{P(\boldsymbol{h}_v \mid \boldsymbol{s})\, P(\boldsymbol{x} \mid \boldsymbol{s}, \phi)\, P(\boldsymbol{s} \mid \theta)}{P(\boldsymbol{h}_v, \boldsymbol{x} \mid \phi, \theta)} \tag{17}$$

$$= \frac{\boldsymbol{\delta}(\boldsymbol{h}_v - \boldsymbol{h}(\boldsymbol{s}, \boldsymbol{v}))\, P(\boldsymbol{x} \mid \boldsymbol{s}, \phi)\, P(\boldsymbol{s} \mid \theta)}{\int \boldsymbol{\delta}(\boldsymbol{h}_v - \boldsymbol{h}(\boldsymbol{s}, \boldsymbol{v}))\, P(\boldsymbol{x} \mid \boldsymbol{s}, \phi)\, P(\boldsymbol{s} \mid \theta)\, d\boldsymbol{s}} \tag{18}$$

where we used the facts that, given a surface, the data and the surface heights are independent, and the surface distribution is independent of the camera and lighting parameters ϕ.

3.2 Approximating the posterior

One motivation for approximating the surface distribution is that generally a surface is an uncountably infinite, continuous entity, and therefore there is little else which can be done to represent it exactly other than to go into, literally, infinite detail (requiring an infinite supply of memory). It is therefore useful to have an approximation scheme which, although finite, captures the relevant information provided by data. Another excellent reason for developing an approximation is mathematical tractability. Having a representation scheme which allows a tractable calculation of the posterior is a huge benefit for both computation and communication. Finally, it is of great interest to not waste computational resources while representing learned surface information. The solution to the surface representation problem presented here addresses the competition for representational resources (memory) issue in a unique manner.

3.2.1 The knowledge representation distribution

The full posterior may be written in the form

$$P(\boldsymbol{s} \mid \boldsymbol{x}, \phi, \theta) = \int P(\boldsymbol{s} \mid \boldsymbol{h}_v, \boldsymbol{x}, \phi, \theta)\, P(\boldsymbol{h}_v \mid \boldsymbol{x}, \phi, \theta)\, d\boldsymbol{h}_v \tag{19}$$

where the distributions inside the integral appear in equations 13–18. The issue of generating a finite representation is not yet resolved via equation 19 however, since storing information sufficient to determine the distributions $P(\boldsymbol{s} \mid \boldsymbol{x}, \phi, \theta)$, and $P(\boldsymbol{s} \mid \boldsymbol{h}_v, \boldsymbol{x}, \phi, \theta)$ generally requires storing an infinite set of values in a finite amount of memory, or requires that all data be stored, disallowing any discarding of data and the incremental updating of the representation. Instead, consider the following approximation where the prior conditioned on a set of heights, along with a new distribution, the *knowledge representation* distribution $\hat{P}(\boldsymbol{h}_v \mid \boldsymbol{x}, \phi, \theta)$, are substituted for the distributions inside the integral of equation 19.

$$\hat{P}(\boldsymbol{s} \mid \hat{P}(\boldsymbol{h}_v \mid \boldsymbol{x}, \phi, \theta)) = \int P(\boldsymbol{s} \mid \boldsymbol{h}_v, \theta) \hat{P}(\boldsymbol{h}_v \mid \boldsymbol{x}, \phi, \theta) \, d\boldsymbol{h}_v \tag{20}$$

It is important to note at this point that *any* suitable surface distribution may be substituted into the right-hand side of equation 20 for $P(\boldsymbol{s} \mid \boldsymbol{h}_v, \theta)$, since it is important only that the resulting integral be capable of making a good approximation to the true posterior. Further, it is not necessary to restrict the basis $\boldsymbol{v}$ to discrete height field basis points, any suitable basis may be taken, for instance Fourier components. Although all of the calculations of this paper are carried thru with the form of 20, other forms may prove more convenient, and it is not difficult to suggest others. In particular, since equation 20 will be used in an iterative update loop later, updates that take for the right-hand side prior term the last posterior term appear quite reasonable (the corresponding GKF update equations may be found immediately from those presented later).

Although conditioning on the KR distribution $\hat{P}(\boldsymbol{h}_v \mid \boldsymbol{x}, \phi, \theta)$ may seem strange, a good way to understand the meaning is that it is the KR distribution which is being used as a statistic for the learned surface information. The key thing to notice in equation 20 is that, with reasonable regularity conditions, choosing the points of $\boldsymbol{v}$ sufficiently dense, the approximation desired to the full posterior may become arbitrarily good. The trick will be to choose $\boldsymbol{v}$ appropriately, properly weighting the competing need to approximate arbitrarily well everywhere with the limited resources that are imposed when a finite amount of storage is available, i.e. when the dimensionality of $\boldsymbol{v}$ is fixed. This will be addressed in the next section. In the case of simple imaging systems, the point spread function and pixel diameter are good indicators of the necessary sampling scale for $\boldsymbol{v}$. In the super-resolved case, the resolution expected available from the data is the appropriate scale for $\boldsymbol{v}$.

The approximation to the posterior of 20 has several properties which make it valuable:

- The prior distribution $P(\boldsymbol{s} \mid \boldsymbol{h}_v, \theta)$ which supplies the uncertainties associated with points of the surface not in the vector $\boldsymbol{v}$ may be chosen to have a simple form (see appendix 12.1) that is easily encoded algorithmically in finite memory.

- There is a clear separation between what was already known - the prior $P(\boldsymbol{s} \mid \boldsymbol{h}_v, \theta)$, and what has been learned - the KR distribution $\hat{P}(\boldsymbol{h}_v \mid \boldsymbol{x}, \phi, \theta)$.

- There is a clear description of the scale at which information has been acquired in terms of the density and uncertainties associated with the points $(\boldsymbol{v}, h(\boldsymbol{s}, \boldsymbol{v}))$ on the surface, and in terms of the uncertainties of their positions as encoded in the KR distribution.

In practice, it is useful to take a multinormal distribution over the discrete-point height field as the KR distribution. Let the parameterization of the KR distribution be Θ_v. For example, if the KR is taken to be multinormal then the parameters of that distribution are

$$\Theta_v(\boldsymbol{x}) = (\boldsymbol{\mu}_v(\boldsymbol{x}), \Sigma_v(\boldsymbol{x})), \tag{21}$$

the mean and covariance matrix of the multinormal, where the functional dependence on $\boldsymbol{x}$ indicates a data dependency through the update procedure, and the subscript $\boldsymbol{v}$ indicates that the parameters parameterize a distribution of heights at points $\boldsymbol{v}$. Because the KR distribution and its parameters are related by a one-to-one mapping, re-write equation 20 as

$$\hat{P}(\boldsymbol{s} \mid \Theta_v, \theta) \;=\; \int P(\boldsymbol{s} \mid \boldsymbol{h}_v, \theta)\, \hat{P}(\boldsymbol{h}_v \mid \Theta_v)\, d\boldsymbol{h}_v. \tag{22}$$

In summary, we have arrived at an approximation to the surface posterior distribution, via the KR distribution, parameterized by Θ_v.

3.3 Updating the knowledge representation

Now we discuss updating Θ_v when new data are acquired. Temporarily restrict attention to the fixed $\boldsymbol{v}$ case. During this and the next sections refer to figure 1 for a flowchart of the general GKF update process.

3.3.1 Bayes' theorem

Having acquired $\Theta_v^n = \Theta_v(x^n)$, from previously seen data $\boldsymbol{x}^n = (\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n)$ and upon seeing new data $\boldsymbol{x}_{n+1}$, the goal is to find Θ_v^{n+1} such that the surface distribution given Θ_v^{n+1} approximates the surface distribution given $\boldsymbol{x}_{n+1}$ and Θ_v^n. Given new data $\boldsymbol{x}_{n+1}$ in the context of the previously seen data $\boldsymbol{x}^n$ summarized by Θ_v^n, our updated surface distribution is found via Bayes' theorem

$$\begin{aligned}
\hat{P}(\boldsymbol{s} \mid \boldsymbol{x}_{n+1}, \Theta_v^n, \phi, \theta) &= \frac{P(\boldsymbol{x}_{n+1} \mid \boldsymbol{s}, \Theta_v^n, \phi, \theta)\hat{P}(\boldsymbol{s} \mid \Theta_v^n, \phi, \theta)}{\hat{P}(\boldsymbol{x}_{n+1} \mid \Theta_v^n, \phi, \theta)} \\
&= \frac{P(\boldsymbol{x}_{n+1} \mid \boldsymbol{s}, \phi)\hat{P}(\boldsymbol{s} \mid \Theta_v^n, \theta)}{\hat{P}(\boldsymbol{x}_{n+1} \mid \Theta_v^n, \phi, \theta)} \\
&= \frac{P(\boldsymbol{x}_{n+1} \mid \boldsymbol{s}, \phi)\hat{P}(\boldsymbol{s} \mid \Theta_v^n, \theta)}{\int P(\boldsymbol{x}_{n+1} \mid \boldsymbol{s}, \phi)\hat{P}(\boldsymbol{s} \mid \Theta_v^n, \theta)\, d\boldsymbol{s}}
\end{aligned} \tag{23}$$

where we defined

$$\hat{P}(\boldsymbol{x}_{n+1} \mid \Theta_v^n, \phi, \theta) = \int P(\boldsymbol{x}_{n+1} \mid \boldsymbol{s}, \phi) \hat{P}(\boldsymbol{s} \mid \Theta_v^n, \theta)\, d\boldsymbol{s}. \tag{24}$$

The updated posterior $\hat{P}(\boldsymbol{s} \mid \Theta_v^n, \boldsymbol{x}_{n+1}, \phi, \theta)$ will be approximated by the Θ_v^{n+1} parameterized KR distribution of equation 22 as

$$\hat{P}(\boldsymbol{s} \mid \Theta_v^{n+1}, \theta) \;\; = \;\; \int P(\boldsymbol{s} \mid \boldsymbol{h}_v, \theta)\, \hat{P}(\boldsymbol{h}_v \mid \Theta_v^{n+1})\, d\boldsymbol{h}_v. \tag{25}$$

The approximation condition for determining Θ_v^{n+1} is then written

$$\hat{P}(\boldsymbol{s} \mid \Theta_v^{n+1}, \theta) \;\; \approx \;\; \hat{P}(\boldsymbol{s} \mid \boldsymbol{x}_{n+1}, \Theta_v^n, \phi, \theta) \tag{26}$$

Equation 26 suggests we try to minimize various measures of the closeness of the two distributions. For example, one measure is the average square difference of the two distributions,

$$\int |P_1(\boldsymbol{s}) - P_2(\boldsymbol{s})|^2\, d\boldsymbol{s} \tag{27}$$

but there is (apparently) no good first-principles reason to use this form. In the next section we discuss the measure of distance which lead$\boldsymbol{s}$ to the *maximally informative* choice of Θ_v^{n+1}.

3.3.2 Maximally informative inference

The measure of distance which leads to the Θ^{n+1} providing the most information about the surface distribution is the *maximally informative* choice for the statistic Θ^{n+1}. The condition for being maximally informative, see [5], is that the Kullback-Leibler distance $D(P_1(\boldsymbol{s}), P_2(\boldsymbol{s}))$ is minimized, where

$$D(P_1(\boldsymbol{s}), P_2(\boldsymbol{s})) = \int P_1(\boldsymbol{s})\, log\left(\frac{P_1(\boldsymbol{s})}{P_2(\boldsymbol{s})}\right) d\boldsymbol{s} \tag{28}$$

and where the P's above are *posterior* distributions of field, that is

$$P_1(\boldsymbol{s}) \;\; = \;\; \hat{P}(\boldsymbol{s} \mid \boldsymbol{x}_{n+1}, \Theta_v^n, \phi, \theta) \tag{29}$$

$$P_2(\boldsymbol{s}) \;\; = \;\; \hat{P}(\boldsymbol{s} \mid \Theta_v^{n+1}, \theta). \tag{30}$$

That is,

Find the Θ^{n+1} *such that*

$$\partial_{\Theta_v^{n+1}} \int \hat{P}(\boldsymbol{s} \mid \Theta_v^n, \boldsymbol{x}_{n+1}, \phi, \theta)\, log\left(\frac{\hat{P}(\boldsymbol{s} \mid \Theta_v^n, \boldsymbol{x}_{n+1}, \phi, \theta)}{\hat{P}(\boldsymbol{s} \mid \Theta_v^{n+1}, \theta)}\right) d\boldsymbol{s} = \mathbf{0} \tag{31}$$

while at the Θ_v^{n+1} *satisfying the derivative condition above*

$$det\left[\partial^2_{\Theta_v^{n+1}} \int \hat{P}(\boldsymbol{s} \mid \Theta_v^n, \boldsymbol{x}_{n+1}, \phi, \theta)\, log\left(\frac{\hat{P}(\boldsymbol{s} \mid \Theta_v^n, \boldsymbol{x}_{n+1}, \phi, \theta)}{\hat{P}(\boldsymbol{s} \mid \Theta_v^{n+1}, \theta)}\right) d\boldsymbol{s}\right] < 0 \tag{32}$$

i.e., the hessian is negative definite and the extremum is a local maximum. If possible, choose the global maximum. Note that the Kullback-Leibler distance is asymmetric. Generally, it is highly relevant which distribution contains the prior information and which distribution is being updated. Maximum entropy techniques reverse the roles of P_1 and P_2 which appear here. For a detailed explanation see [5].

In the following section are some observations on the approach taken to maximally informative surface inference. Section 5 then briefly makes explicit the specific distribution forms which are assumed. The Generalized Kalman Filter update equations for the surface inference example which follow from this approach are then presented in section 6, completing the derivation of the maximally informative approach.

4 Observations on the update scheme

Note the following:

- The updating scheme described here is a maximally informative update scheme and is related to the Kalman filter. The Kalman filter is a minimum variance filtering scheme applicable in the case of fixed representation dimension. The crucial step which has been taken in the current work is the step of allowing the representation scheme to be adaptable. We have adopted the label "Generalized Kalman Filter" (GKF) to describe the idea represented here. The GKF equations are presented in section 6.

- To this point we have only optimized over Θ_v. It is clear that we may also vary the number of vertices $|\boldsymbol{v}|$ of the representation, allowing optimization over the number of vertices. Varying the number of vertices of the representation is absolutely necessary if surface knowledge at scales smaller than the current set of vertices represents is to ever accumulate. In section 6 the GKF update equations are derived assuming that the number of vertices in the representation basis vertex set is arbitrary at each update.

- Beyond allowing the number of vertices to vary, the positions of the vertices may be allowed to vary. In section 6 the GKF update equations are derived assuming that the representation basis vertex set positions are arbitrary.

- Detecting when and where new vertices are necessary is a matter of observing directly in equations 28 or 31 when new data produces a lower surface uncertainty over a region, and when having smaller uncertainty at neighboring vertices is not sufficient to represent this lower uncertainty over the region.

- The vertex representation for the surface knowledge is convenient, but not necessary. For example it is possible to extend a height field to a height-and-reflectance field or "arbitrary dimension field", where the reflectance lies within a many-dimensional space. Reasonable structures for the covariance matrix allow differing correlations between reflectance values and between height values. It will be seen in in section 6 that the GKF update equations are easily used in the "arbitrary dimension field" context.

- In its most abstract form, instead of having a "field", there is simply a set of objects, while for each "object" there is an associated vector of properties, where some of the components of the property vector may be considered a location in space. In this fairly abstracted setting, the collection of objects has an associated joint probability distribution which describes the probability distribution over configurations of objects. It will be seen in in section 6 that the GKF update equations are easily understood in the "object" context.

- Equation 31 which defines the quantity to be minimized is where a penalty term which indicates how many bits in hardware is available in trade for each bit of information learned from data. For example, one might penalize the KL distance by 1/10th the number of bytes it takes to represent the new information gained by extending the number of points represented. The exact form of the information learned about the surface distribution contained in the KR distribution is found in section 8, where the dimensionality of the representation enters directly, and where bits-used penalty-terms may be introduced.

- The previous note points out how a minimum description length method fails for this problem. It is certainly the case that that our update scheme may require much more memory (in bits) to represent the information learned than the information learned (in bits). At some point, if information at small enough scales is desired, MDL would truncate and stop. Clearly, applying MDL would then be a disaster. On the other hand, what seems to work here may be called an adaptive MDL approach.

- Note that a method like maximum entropy is entirely deficient for providing distributions of surfaces: given the constraints implied by the knowledge of the distribution of the heights at discrete points: maximum entropy ignores correlations between nearby surface points no matter how close, an entirely ludicrous situation. On the other hand, a method like relative

maximum entropy, based on inverting the roles of the distributions in equation 28, claims to provide the *least* informative inference relative to the prior information, a heuristic, difficult to justify, at best. Further, such approaches are typically based on likelihood distributions, rather than the posteriors that appear in equation 28.

5 Surface Distribution Forms

5.1 Prior

For simplicity of mathematical presentation *only*, the prior in our surface inference example is taken multinormal over continuous, smooth height fields. One particular, conveniently chosen, representation of the prior distribution is constructed in appendix 12.1. This prior may be written in the shorthand

$$P(\boldsymbol{s} \mid \theta) = N(\boldsymbol{\mu}_s, \Sigma_s)(\boldsymbol{s}) \tag{33}$$

where $\theta = (\boldsymbol{\mu}_s, \Sigma_s)$ is the parameter vector. The density of the height field determined by the prior

$$\begin{aligned} P(\boldsymbol{h}_v \mid \theta) &= \int P(\boldsymbol{h}_v \mid \boldsymbol{s})\, P(\boldsymbol{s} \mid \theta)\, d\boldsymbol{s} && (34) \\ &= \int \boldsymbol{\delta}(\boldsymbol{h}_v - \boldsymbol{h}(\boldsymbol{s}, \boldsymbol{v}))\, P(\boldsymbol{s} \mid \theta)\, d\boldsymbol{s} && (35) \\ &= N(\boldsymbol{\mu}_v, \Sigma_v)(\boldsymbol{h}_v) && (36) \end{aligned}$$

where

$$\begin{aligned} \boldsymbol{\mu}_v &= A_{vs}\boldsymbol{\mu}_s \\ \Sigma_v &= A_{vs}\Sigma_s A_{vs}^T \end{aligned} \tag{37}$$

and the projection onto the height field is given by A_{vs}. Note that equation 37 implies that the surface density covariance is represented differently than a discrete surface distribution covariance matrix. Specifically, the projection matrix A_{vs} is a delta-function-like operator, and Σ_s is a continuous function of two positions. In appendix 12.1 we show that the surface density has a compact continuous power spectrum representation, and there give the explicit form of that representation. Thus the notation of equation 37 must be considered a shorthand for the underlying continuous construct.

5.2 Likelihood

When measurement is modelled as a linear process corrupted by gaussian noise we have

$$\boldsymbol{x} = M\boldsymbol{s} + \epsilon$$

$$\epsilon \sim N(\mathbf{0}, \Sigma_\epsilon). \tag{38}$$

or

$$P(\boldsymbol{x} \mid \boldsymbol{s}, \phi) = N(M\boldsymbol{s}, \Sigma_\epsilon)(\boldsymbol{x}) \tag{39}$$

where $\phi = (M, \Sigma_\epsilon)$ is the parameter vector.

6 The Generalized Kalman Filter equations.

In this section a concise derivation of the Generalized Kalman Filter update equations specialized to the discrete basis multinormal KR distribution of equation 22 are derived. The updated KR need not have the same basis dimension nor position as the previous KR basis, solving the problem of how to allow updates from one representation to the next, same, finer or coarser, representation.

Proceeding, the KR distribution in terms of the parameterized height field of equation 22 is

$$\hat{P}(\boldsymbol{s} \mid \Theta_v^n, \theta) = \int P(\boldsymbol{s} \mid \boldsymbol{h}_v, \theta)\, \hat{P}(\boldsymbol{h}_v \mid \Theta_v^n)\, d\boldsymbol{h}_v \tag{40}$$

The distribution of surface given the height field from equation 9 is

$$\begin{aligned} P(\boldsymbol{s} \mid \boldsymbol{h}_v\theta) &= \frac{P(\boldsymbol{h}_v \mid \boldsymbol{s})\, P(\boldsymbol{s} \mid \theta)}{P(\boldsymbol{h}_v \mid \theta)} \\ &= \frac{\boldsymbol{\delta}(\boldsymbol{h}_v - \boldsymbol{h}(\boldsymbol{s}, \boldsymbol{v}))\, P(\boldsymbol{s} \mid \theta)}{P(\boldsymbol{h}_v \mid \theta)} \end{aligned} \tag{41}$$

Simplify the integral of the KR distribution to find

$$\begin{aligned} \hat{P}(\boldsymbol{s} \mid \Theta_v^n, \theta) &= \int \frac{P(\boldsymbol{h}_v \mid \boldsymbol{s})\, P(\boldsymbol{s} \mid \theta)}{P(\boldsymbol{h}_v \mid \theta)}\, \hat{P}(\boldsymbol{h}_v \mid \Theta_v^n)\, d\boldsymbol{h}_v \\ &= P(\boldsymbol{s} \mid \theta) \int \boldsymbol{\delta}(\boldsymbol{h}_v - \boldsymbol{h}(\boldsymbol{s}, \boldsymbol{v})) \frac{\hat{P}(\boldsymbol{h}_v \mid \Theta_v^n)}{P(\boldsymbol{h}_v \mid \theta)}\, d\boldsymbol{h}_v \\ &= P(\boldsymbol{s} \mid \theta) \frac{\hat{P}(\boldsymbol{h}(\boldsymbol{s}, \boldsymbol{v}) \mid \Theta_v^n)}{P(\boldsymbol{h}(\boldsymbol{s}, \boldsymbol{v}) \mid \theta)} \end{aligned} \tag{42}$$

Note how the full surface distribution is simply modified by the ratio

$$\frac{\hat{P}(\boldsymbol{h}(\boldsymbol{s}, \boldsymbol{v}) \mid \Theta_v^n)}{P(\boldsymbol{h}(\boldsymbol{s}, \boldsymbol{v}) \mid \theta)} \tag{43}$$

From equation 23 the Bayesian update of the KR distribution is

$$\begin{aligned} \hat{P}(\boldsymbol{s} \mid \boldsymbol{x}_{n+1}, \Theta_v^n, \phi, \theta) &= \frac{P(\boldsymbol{x}_{n+1} \mid \boldsymbol{s}, \phi)\, \hat{P}(\boldsymbol{s} \mid \Theta_v^n, \theta)}{\int P(\boldsymbol{x}_{n+1} \mid \boldsymbol{s}, \phi)\, \hat{P}(\boldsymbol{s} \mid \Theta_v^n, \theta)\, d\boldsymbol{s}} \\ &= \frac{P(\boldsymbol{x}_{n+1} \mid \boldsymbol{s}, \phi)\, \hat{P}(\boldsymbol{s} \mid \Theta_v^n, \theta)}{\hat{P}(\boldsymbol{x}_{n+1} \mid \Theta_v^n, \phi, \theta)} \end{aligned} \tag{44}$$

Rewriting the updated distribution using equation 42 yields

$$\hat{P}(\boldsymbol{s} \mid \boldsymbol{x}_{n+1}, \Theta_v^n, \phi, \theta) \;\; \propto \;\; P(\boldsymbol{x}_{n+1} \mid \boldsymbol{s}, \phi)\, P(\boldsymbol{s} \mid \theta) \times \frac{\hat{P}(\boldsymbol{h}(\boldsymbol{s}, \boldsymbol{v}) \mid \Theta_v^n)}{P(\boldsymbol{h}(\boldsymbol{s}, \boldsymbol{v}) \mid \theta)} \tag{45}$$

For maximally informative inference of the new KR we minimize, from equation 28,

$$\begin{aligned} D(P_1(\boldsymbol{s}), P_2(\boldsymbol{s})) &= D(\hat{P}(\boldsymbol{s} \mid \boldsymbol{x}_{n+1}, \Theta_v^n, \phi, \theta), \hat{P}(\boldsymbol{s} \mid \Theta_{\overline{v}}^{n+1}, \theta)) \\ &= \int \hat{P}(\boldsymbol{s} \mid \boldsymbol{x}_{n+1}, \Theta_v^n, \phi, \theta)\, log\left(\frac{\hat{P}(\boldsymbol{s} \mid \boldsymbol{x}_{n+1}, \Theta_v^n, \phi, \theta)}{\hat{P}(\boldsymbol{s} \mid \Theta_{\overline{v}}^{n+1}, \theta)}\right) d\boldsymbol{s} \end{aligned} \tag{46}$$

Note that it is not assumed here that $\boldsymbol{v}$ and $\overline{\boldsymbol{v}}$ have the same dimension. Expanding the probability distributions within the logarithm appearing above yields

$$\begin{aligned} D(P_1(\boldsymbol{s}), P_2(\boldsymbol{s})) = \int \hat{P}(\boldsymbol{s} \mid \boldsymbol{x}_{n+1}, \Theta_v^n, \phi, \theta) \\ \times [\, -log\,(P(\boldsymbol{h}(\boldsymbol{s}, \boldsymbol{v}) \mid \theta)) \\ +log\,(P(\boldsymbol{h}(\boldsymbol{s}, \overline{\boldsymbol{v}}) \mid \theta)) \\ +log\,(P(\boldsymbol{x}_{n+1} \mid \boldsymbol{s}, \phi)) \\ -log\left(\hat{P}(\boldsymbol{x}_{n+1} \mid \Theta_v^n, \phi, \theta)\right) \\ +log\left(\hat{P}(\boldsymbol{h}(\boldsymbol{s}, \boldsymbol{v}) \mid \Theta_v^n)\right) \\ -log\left(\hat{P}(\boldsymbol{h}(\boldsymbol{s}, \overline{\boldsymbol{v}}) \mid \Theta_{\overline{v}}^{n+1})\right)\,] \; d\boldsymbol{s} \end{aligned} \tag{47}$$

Each term has the form of an information (or uncertainty). Together the six terms paint a descriptive picture of how information is acquired by the maximally informative update when taken as three groups of two terms: Denote by "new KR" the two terms with $\overline{\boldsymbol{v}}$ and $\Theta_{\overline{v}}^{n+1}$, by "previous KR" the two terms with $\boldsymbol{v}$ and Θ_v^n and no data, and by "new data" the two terms with data dependency. Now, noting the signs on these quantities, because D is positive, the whole point of choosing a good Θ^{n+1} approximation by minimizing D is that

$$\begin{aligned} &\textit{Expected information in new KR} \simeq \\ &\quad (\textit{Expected information in previous KR} \\ &\qquad + \textit{Expected information in new data}) \end{aligned} \tag{48}$$

or in very rough terms we may see the update as capturing the sum-total of the available knowledge

$$\textit{Total knowledge} = \textit{Prior knowledge} + \textit{New knowledge from data} \tag{49}$$

Because only terms depending upon the update parameters $\overline{\boldsymbol{v}}$ and $\Theta_{\overline{v}}^{n+1}$ are needed to perform the minimization, we drop the other terms at this point, and after making the multinormal substitutions for the distributions in the above we have

$$\begin{aligned}\bar{D}(P_1(\boldsymbol{s}), P_2(\boldsymbol{s})) &= \int \hat{P}(\boldsymbol{h}_{\overline{v}} \mid \boldsymbol{x}_{n+1}, \Theta_v^n, \phi, \theta)\, log\left(N(\boldsymbol{\mu}_{\overline{v}}, \Sigma_{\overline{v}})(\boldsymbol{h}_{\overline{v}})\right)\, d\boldsymbol{h}_{\overline{v}} \\ &- \int \hat{P}(\boldsymbol{h}_{\overline{v}} \mid \boldsymbol{x}_{n+1}, \Theta_v^n, \phi, \theta)\, log\left(N(\boldsymbol{\mu}_{\overline{v}}^{n+1}, \Sigma_{\overline{v}}^{n+1})(\boldsymbol{h}_{\overline{v}})\right)\, d\boldsymbol{h}_{\overline{v}}\end{aligned} \tag{50}$$

To simplify the $\hat{P}$'s appearing in equation 50, the distribution of surface given old knowledge and new data, marginalized to the height field $\overline{\boldsymbol{v}}$, is useful, as is seen by observing equations 47 and 50. Thus, consider

$$\begin{aligned}\hat{P}(\boldsymbol{s} \mid \boldsymbol{x}_{n+1}, \Theta_v^n, \phi, \theta) &\propto N(M(\phi)\boldsymbol{s}, \Sigma_\epsilon^{n+1})(\boldsymbol{x}_{n+1})\, N(\boldsymbol{\mu}_s, \Sigma_s)(\boldsymbol{s}) \\ &\quad \times \frac{N(\boldsymbol{\mu}_v^n, \Sigma_v^n)(\boldsymbol{h}(\boldsymbol{s}, \boldsymbol{v}))}{N(\boldsymbol{\mu}_v, \Sigma_v)(\boldsymbol{h}(\boldsymbol{s}, \boldsymbol{v}))}\end{aligned} \tag{51}$$

found by making substitutions into 45 for the assumed distributions. Since it is not necessarily the case that $v_i \in \{\overline{v}_j\}$ or that $\overline{v}_i \in \{v_j\}$. proceed by marginalizing to the union of the components of $\boldsymbol{v}$ and $\overline{\boldsymbol{v}}$, which we denote $\boldsymbol{v} \cup \overline{\boldsymbol{v}}$, and then to the $\overline{\boldsymbol{v}}$ components. Let $A_{v\cup\overline{v},s}$ denote the projection from $\boldsymbol{v}_s$ to $\boldsymbol{v} \cup \overline{\boldsymbol{v}}$, $A_{\overline{v},v\cup\overline{v}}$ denote the projection from $\boldsymbol{v} \cup \overline{\boldsymbol{v}}$ to $\overline{\boldsymbol{v}}$, and $A_{\overline{v},v}$ denote the projection from $\boldsymbol{v}$ to $\overline{\boldsymbol{v}}$. In performing the two projections (from $\boldsymbol{v}_s$ to $\boldsymbol{v} \cup \overline{\boldsymbol{v}}$, and then from $\boldsymbol{v} \cup \overline{\boldsymbol{v}}$ to $\overline{\boldsymbol{v}}$) in order we find (not necessarily in most simple form), using results of appendices 12.2–12.5, that

$$\int \hat{P}(\boldsymbol{s} \mid \boldsymbol{x}_{n+1}, \Theta_v^n, \phi, \theta)\, d\boldsymbol{s} \setminus \overline{\boldsymbol{v}} = N(\boldsymbol{\mu}_R, \Sigma_R)(\boldsymbol{h}_{\overline{\boldsymbol{v}}}) \tag{52}$$

where

$$\begin{aligned}\boldsymbol{\mu}_{\overline{v}}^R &= \Sigma_R(\Sigma_Q^{-1}\boldsymbol{\mu}_{\overline{\boldsymbol{v}}}^Q + (\Sigma_{\overline{v}}^n)^{-1}\boldsymbol{\mu}_{\overline{\boldsymbol{v}}}^{\boldsymbol{n}} - \Sigma_{\overline{v}}^{-1}\boldsymbol{\mu}_{\overline{\boldsymbol{v}}}) \\ \Sigma_R^{-1} &= \Sigma_Q^{-1} + (\Sigma_{\overline{v}}^n)^{-1} - \Sigma_{\overline{v}}^{-1}\end{aligned} \tag{53}$$

and where

$$\begin{aligned}\boldsymbol{\mu}_{\overline{v}}^Q &= A_{\overline{v},v\cup\overline{v}}A_{v\cup\overline{v},s}\boldsymbol{\mu}_s^P \\ \Sigma_Q^{-1} &= A_{\overline{v},v\cup\overline{v}}A_{v\cup\overline{v},s}\Sigma_P^{-1}A_{v\cup\overline{v},s}^T A_{\overline{v},v\cup\overline{v}}^T\end{aligned} \tag{54}$$

$$\begin{aligned}\boldsymbol{\mu}_s^P &= \Sigma_P(\Sigma_s^{-1}\boldsymbol{\mu}_s + M^T\Sigma_\epsilon^{-1}\boldsymbol{x}_{n+1}) \\ \Sigma_P^{-1} &= \Sigma_s^{-1} + M^T\Sigma_\epsilon^{-1}M\end{aligned}$$

(55)

$$\begin{aligned} \boldsymbol{\mu}_{\overline{\boldsymbol{v}}}^{\boldsymbol{n}} &= A_{\overline{v},v}\boldsymbol{\mu}_{\boldsymbol{v}}^{\boldsymbol{n}} \\ (\Sigma_{\overline{v}}^{n})^{-1} &= A_{\overline{v},v}(\Sigma_{v}^{n})^{-1}A_{\overline{v},v}^{T} \end{aligned}$$

(56)

$$\begin{aligned} \boldsymbol{\mu}_{\overline{\boldsymbol{v}}} &= A_{\overline{v},v}\boldsymbol{\mu}_{\boldsymbol{v}} \\ \Sigma_{\overline{v}}^{-1} &= A_{\overline{v},v}\Sigma_{v}^{-1}A_{\overline{v},v}^{T} \end{aligned}$$

(57)

$$\begin{aligned} \boldsymbol{\mu}_{v} &= A_{v,s}\boldsymbol{\mu}_{\boldsymbol{s}} \\ \Sigma_{v}^{-1} &= A_{v,s}\Sigma_{s}^{-1}A_{v,s}^{T} \end{aligned}$$

(58)

Using the results of appendix 12.6, the quantities of equation 53 above correspond to the values of the mean and standard deviation parameters of the new KR, found at the minimum Kullback Leibler distance, i.e. the minimization is immediately apparent from those results. Thus:

$$\begin{aligned} \Theta_{\overline{v}}^{n+1} &= (\boldsymbol{\mu}_{\overline{v}}^{n+1}, \Sigma_{\overline{v}}^{n+1}) \\ \boldsymbol{\mu}_{\overline{v}}^{n+1} &= \boldsymbol{\mu}_{\overline{v}}^{R} \\ \Sigma_{\overline{v}}^{n+1} &= \Sigma_{\overline{v}}^{R} \end{aligned} \tag{59}$$

Equations 53 are the Generalized Kalman Filter (GKF) update equations for the surface inference example, yet are quite a bit more general (the necessary change of variables needed when the forward projection is nonlinear appears in appendix 12.10). Having these update equations allows one to consider updating a representation of any dimension relative to the original representation. Thus. knowledge may be represented in finer detail, corresponding to the old representation being contained in the new, knowledge may be represented in the same detail, corresponding to the case when the new representation is the same as the old representation, or knowledge may be tossed, corresponding to the case when the new representation does not contain the old representation. The maximally informative inference approach and its result of the Kullback Leibler distance on conditional posteriors led directly here to deriving the GKF and the solution of the problem of storing knowledge at scales adaptive to the actual needs of the data driving the update. The standard KF is discussed in [1].

7 Specializing the GKF

When the surface of interest is itself a discrete height field, and the KR representation basis never changes in dimension nor position from that height field's basis, then all projections appearing

in equations 53 and following are identities, and the update equations simplify to the standard Kalman filter equations, in effect equations 55 only, given suitable identification of the variables.

8 Information learned

Once a new set of parameters has been chosen, and for the purpose of evaluating the new update in the context of other possible updates at different scales, using different representational bases, it is useful to have the quantity of information about the surface distribution that is contained in the KR at the maximally informative update. Using the results of appendix 12.6 in equation 50 we have this information, up to a constant, is given by

$$\begin{aligned} I_R &= C(\boldsymbol{x}_{n+1}, \Theta_v^n, \phi, \theta) \\ &+ \frac{1}{2}\left(Tr\left[(\Sigma_R + U(\boldsymbol{\mu}_R - \boldsymbol{\mu}_{\overline{v}})) \otimes \Sigma_{\overline{v}}^{-1}\right] + log(|\Sigma_{\overline{v}}|)\right) \\ &- \frac{1}{2}\left(Tr\left[\Sigma_R \otimes \Sigma_R^{-1}\right] + log(|\Sigma_R|)\right) \end{aligned} \tag{60}$$

Note that the d's (representation basis dimensions) from the $dlog(2\pi)$'s of equation 94 have cancelled. However the d's remain hidden within the terms as matrix dimensions. When considering optimizing learned inormation against storage resources, one must weigh a separate cost in bits for the memory used against the bits learned, the expression above. Note also, interestingly the expression above contains a BIC-like $log(d)$ dependence term.

9 Search for update parameters

Now that we know what the update equations for the updating of the KR distribution look like, it is worthwhile considering how an updating scheme might be implemented to acquire information at the appropriate scale. First, we dismiss the notion that we will ever be using the continuous height field $\boldsymbol{v}_s$ (the support of $\boldsymbol{s}$) at any time. None of the update equations force that to happen! Second, since we have concluded that computationally $\boldsymbol{v}_s$ is a discrete set, and since there will always be pathological cases where the surface is much rougher than we care to represent, we acknowledge that fact and proceed by presenting a useful algorithm which allows the updating of the KR while maintaining the ability to explore a large range of scales. The following multigrid-style algorithm provides the general flavor:

- Choose $\boldsymbol{v}_s$ denser by several orders of scale than the current representation, and using other criteria associated with the knowledge of the data acquisition system (see below).
- Choose $\overline{\boldsymbol{v}}$ at regular scales intermediate between $\boldsymbol{v}_s$ and the old KR on $\boldsymbol{v}$, compute the updates on all $\overline{\boldsymbol{v}}$ chosen at these scales.

- Compute the information learned at each scale.
- Plot the information learned as a function of increasing density (decreasing scale).
- Choose, based on exploration of the plot, and costs associated with storing the learned information, whether to explore other octaves of scale. If Choose to explore, repeat above procedure.
- If choice is to pick an informationally and storage attractive KR, do this and update the representation accordingly.

In the surface reconstruction problem data often comes in the form of images. The images may come from devices with vastly different resolutions, and the known parameters of pixel size, point spread function and geometry determine the appropriate reconstruction scale. Finally adapting the surface to resolve at sub-pixel scales requires a memory-aggressive approach which extends the exploration farther out on the learning curve towards smaller, denser representation scales.

10 Conclusion

Field inference has been generalized from the typical discrete fixed-basis setting to a continuous-basis setting. The problem of surface inference was solved in the context of continuous field inference. Using the approach of acquiring the maximally informative KR distribution, the GKF equations were found. The GKF allows the updated KR parameters to be found at any scale and/or "positions" (abstractly, basis components). The approach allows the learning of information at the relevant scales desired. It provides an information-theoretic justification for location-dependent adaptive multi-grid inference. It also effectively provides similar justification for a scale-adaptive MDL method. This is apparently the first time that the maximally informative inference of continuous-basis objects and the multigrid approach have been rigorously justified.

11 Acknowledgements

I thank the members of the Ames Data Understanding group for their interest and comments, especially the invaluable valiant contributions of Dr. Robin D. Morris, who thoughtfully, carefully, and painstakingly spent a week-in-agony checking the maths (any remaining mistakes are fully mine, however), and Drs. Vadim Smelyanskiy and David Maluf for their comments. Finally, immense thanks go to Dr. Peter Cheeseman, for comments, and support. This project was partially supported by the NASA Ames Center for Excellence in Information Technology contract NAS-214217.

12 Appendices

12.1 Construction of a 2D surface prior

In this appendix we first introduce the reader to the fourier representation of a gaussian process, then using the notions developed find the representation for a 2D gaussian process over the plane, where the correlations of the process at points $\boldsymbol{x}$ and $\boldsymbol{y}$ are proportional to $exp(-k\,|\boldsymbol{x}-\boldsymbol{y}|)$, $k>0$, a simple translation-invariant choice for the form of the correlation structure of the probability density of surfaces having the plane as support. The utility for the GKF of having this process is that it serves as a simply computed algorithmic representation of the prior for surfaces having the plane as support.

12.1.1 The discrete gaussian process

Consider $f(n,\boldsymbol{c})$, $n\in Z_N=\{-N,\ldots,-1,0,1,\ldots,N\}$, a discrete process with expression as the fourier expansion

$$f(n,\boldsymbol{c})=\sum_{k=-N}^{N}c_k e^{ikn} \tag{61}$$

where the coefficients $\boldsymbol{c}=(c_k)$ are constrained by $f\in\boldsymbol{R}$ so that $c_k=c_{-k}^*$, and the n and k range over Z_N. Let the coefficients be random variables: $c_k=x_k+iy_k$ with $x_k\sim N(0,\sigma_k)$ and $y_k\sim N(0,\sigma_k)$ both gaussian distributed random variables with mean 0 and standard deviation σ_k. Now, dropping the k's, the joint density of (x,y) is given by

$$P_{x,y}(x,y)=\frac{e^{-x^2/2\sigma^2}}{\sqrt{2\pi}\sigma}\frac{e^{-y^2/2\sigma^2}}{\sqrt{2\pi}\sigma}. \tag{62}$$

From this the joint density of (r,θ) where $r=\sqrt{x^2+y^2}$ and $\theta=\arctan(y/x)$ is given by

$$P_{r,\theta}(r,\theta)=\frac{re^{-r^2/2\sigma^2}}{2\pi\sigma^2}. \tag{63}$$

The density of r is given directly by integrating over θ

$$P_r(r)=\frac{re^{-r^2/2\sigma^2}}{\sigma^2}, \tag{64}$$

while the density of θ is given directly by integrating over r

$$P_\theta(\theta)=\frac{1}{2\pi}. \tag{65}$$

Making a change of variables, the density of $cc^* = x^2 + y^2 = r^2$ is given by the exponential distribution

$$P_{cc^*}(u) = \frac{e^{-u/2\sigma^2}}{2\sigma^2} \tag{66}$$

The distribution of $c_k + c_{-k} = 2Re[c_k] = 2x_k$, $k > 0$ is of interest because the process is real.

$$P_{c+c^*}(u) = \frac{e^{-u^2/2(2\sigma)^2}}{\sqrt{2\pi}2\sigma} \tag{67}$$

which is just a gaussian with zero mean but twice the variance of the components x and y of c. Note that the actual coefficients in equation 61 $c_k e^{ikn} + c_{-k}e^{-ikn} = 2Re[c_k e^{ikn}]$ also have the distribution of equation 67 since the phase of c_k is uniformly distributed in $[0, 2\pi]$.

Now, given a set of integers $\zeta \subset Z_N$ we may ask for the density of the sampled values of the process f at $\boldsymbol{\zeta} = (n_1, n_2, \ldots, n_m)$

$$\boldsymbol{f}(\boldsymbol{\zeta}) = (f(n_1), f(n_2), \ldots, f(n_m)), \tag{68}$$

where $m = |\zeta|, n_i \in Z_N, i = 1, \ldots, m$. Define

$$\boldsymbol{f}(\boldsymbol{\zeta}, \boldsymbol{c}) = (f(n_1, \boldsymbol{c}), f(n_2, \boldsymbol{c}), \ldots, f(n_m, \boldsymbol{c})) \tag{69}$$

Then the probability density function which describes the sampled values is

$$P(\boldsymbol{f}(\boldsymbol{\zeta})) = \int \delta(\boldsymbol{f}(\boldsymbol{\zeta}) - \boldsymbol{f}(\boldsymbol{\zeta}, \boldsymbol{c}))\, P(\boldsymbol{c})\, d\boldsymbol{c} \tag{70}$$

where

$$P(\boldsymbol{c}) = P(c_0) \prod_{k=1}^{N} P(c_k + c_{-k}) \tag{71}$$

Note that that the density of $P(\boldsymbol{f}(\boldsymbol{\zeta}))$ is multivariate gaussian since the representation of $\boldsymbol{f}(\boldsymbol{\zeta}, \boldsymbol{c})$ as a fourier series shows that it is the sum of gaussian random vectors with components $2Re[c_k e^{ikn}]$. The covariances of the process are found as

$$\begin{aligned} \Sigma_{m,n} = E[f(m)f(n)] &= E[f(m)f^*(n)] \\ &= E\left[\sum_{k,l=-N}^{N} c_k c_l^* e^{i(km-ln)}\right] \\ &= \sum_{k=-N}^{N} E[c_k c_k^*] e^{ik(m-n)} \\ &= F[E[c_k c_k^*]](m-n) \end{aligned} \tag{72}$$

where we used the fact that the coefficients of different frequency are uncorrelated for $k \neq l$, i.e $E[c_k c_l^*] = 0$ for $k \neq l$. Define the power spectrum $R(k)$ as

$$R(k) = E[c_k c_k^*] \tag{73}$$

Then we have that the covariance is given by the fourier transform of the power spectrum,

$$\Sigma_{m,n} = E[f(m)f(n)] = F[R](m-n) = \Sigma_{m-n} \tag{74}$$

where we have acknowledged that the covariance structure is dependent only upon the difference $m-n$. From this we see that the inverse fourier transform of the covariance is the power spectrum,

$$F^{-1}\left[\Sigma_u\right](k) = R(k) \tag{75}$$

Finally, note that the density of $c_k c_k^*$ given by equation 66 allows us to infer the parameters σ_k which are the standard deviations of the gaussian processes x_k and y_k underlying the coefficients c_k, since from equation 66

$$E[c_k c_k^*] = \int u \frac{e^{-u/2\sigma_k^2}}{2\sigma_k^2}\, du = 2\sigma_k^2 \tag{76}$$

In the next section the basis for gaussian processes developed here is extended to the continuous 2D case to compute the power spectrum of a process specified by a continuous-basis covariance structure.

12.1.2 The continuous-basis 2D process

Similar to the development in the last section, in two dimensions, given the continuous-basis covariance $\Sigma_{\boldsymbol{x}} = exp(-k\,|\boldsymbol{x}|)$, $k > 0.$, the power spectrum is found as the inverse fourier transform of the covariance, i.e.

$$\begin{aligned} R(\boldsymbol{u} = (u,v)) &= F_2^{-1}[\Sigma_{\boldsymbol{x}}](u,v) \\ &= \int\int e^{-k|(x,y)|} e^{-iux} e^{-ivy}\, dx\, dy \end{aligned} \tag{77}$$

Make the change of variables $(x,y) \rightarrow (r,\theta)$ so that $x = rcos(\theta)$, $y = rsin(\theta)$, then

$$R(u,v) = \int_0^\infty \int_0^{2\pi} e^{-kr} e^{-ir(ucos(\theta)+vsin(\theta))}\, r\, dr\, d\theta \tag{78}$$

For simplicity, make the further change of variables $(u,v) \rightarrow (s,\phi)$ so that $u = scos(\phi)$, $v = ssin(\phi)$, so that

$$R(s,\phi) = \int_0^\infty \int_0^{2\pi} e^{-kr} e^{-irs(cos(\phi)cos(\theta)+sin(\phi)sin(\theta))}\, r\, dr\, d\theta$$

$$\begin{aligned} &= \int_0^\infty \int_0^{2\pi} e^{-kr} e^{-irscos(\theta-\phi)}\, r\, dr\, d\theta \\ &= \int_0^\infty r e^{-kr} \int_0^{2\pi} e^{-irscos(\theta-\phi)}\, d\theta\, dr \\ R(s) &= 2\pi \int_0^\infty r e^{-kr} J_0(rs)\, dr \end{aligned} \tag{79}$$

Finally,

$$R(\boldsymbol{u}) = \frac{2\pi k}{(|\boldsymbol{u}|^2 + k^2)^{3/2}} \tag{80}$$

Note that we have neglected the proportionality constant $1/2\pi$ in the fourier transform, amounting to normalizing the delta function to 2π, and have scaled $\boldsymbol{u}$ to units of cycles per 2π. Note also that both the covariance of the process and the power spectrum scale with the same proportionality constant. Harmonic analysis is discussed in [3]

12.2 Multinormal density MGF

The moment generating function for a probability distribution f is defined as the functional

$$M[f](\boldsymbol{\lambda}) = E_f[e^{Tr[U(\boldsymbol{\lambda},\boldsymbol{x})]}] \tag{81}$$

where $U(\boldsymbol{y},\boldsymbol{z})$ is defined such that $U = [U_{ij}]$ and $U_{ij}(\boldsymbol{y},\boldsymbol{z}) := y_i z_j$, from which holds the property

$$\frac{\partial^k M[f](\boldsymbol{\lambda})}{\partial\lambda_{i_1}\ldots\lambda_{i_k}} \mid_{\boldsymbol{\lambda}=0} = E_f[x_{i_1}\ldots x_{i_k}] \tag{82}$$

i.e the moments are found as derivatives of the MGF with respect to the parameter $\boldsymbol{\lambda}$ at $\boldsymbol{\lambda} = 0$.

Take the multinormal density function for $\boldsymbol{x}$

$$\begin{aligned} P(\boldsymbol{x} \mid \Theta) &= N(\Theta)(\boldsymbol{x}) \\ &= N(\boldsymbol{\mu},\Sigma)(\boldsymbol{x}) \\ &= \frac{1}{(2\pi)^{d/2} \mid \Sigma \mid^{1/2}} exp(-\frac{1}{2} Tr[U(\boldsymbol{x}-\boldsymbol{\mu}) \otimes \Sigma^{-1}]) \end{aligned} \tag{83}$$

where $U(\boldsymbol{y})$ is defined such that $U_{ij}(\boldsymbol{y}) := U_{ij}(\boldsymbol{y},\boldsymbol{y})$ and $d = Dim(\boldsymbol{x})$. The MGF of $N(\Theta)(\boldsymbol{x})$ is then given by

$$\begin{aligned} &M[N(\Theta)(\boldsymbol{x})](\boldsymbol{\lambda}) = E[e^{Tr[U(\boldsymbol{\lambda},\boldsymbol{x})]} \mid \Theta] \\ &= \int \frac{1}{(2\pi)^{d/2} \mid \Sigma \mid^{1/2}} exp(-\frac{1}{2} Tr[U(\boldsymbol{x}-\boldsymbol{\mu}) \otimes \Sigma^{-1}] + Tr[U(\boldsymbol{\lambda},\boldsymbol{x})])\, d\boldsymbol{x} \end{aligned} \tag{84}$$

Minus twice the exponent of the integral above may be written as

$$\begin{aligned} Tr[U(\boldsymbol{x}-\boldsymbol{\mu})\otimes\Sigma^{-1}]-2Tr[U(\boldsymbol{\lambda},\boldsymbol{x})] &= Tr[U(\boldsymbol{x}-(\boldsymbol{\mu}-\boldsymbol{\lambda}\,\Sigma))\otimes\Sigma^{-1}] \\ &\quad +Tr[U(\boldsymbol{\mu})\otimes\Sigma^{-1}] \\ &\quad -Tr[U(\boldsymbol{\mu}-\boldsymbol{\lambda}\,\Sigma)\otimes\Sigma^{-1}] \\ &= Tr[U(\boldsymbol{x}-(\boldsymbol{\mu}-\boldsymbol{\lambda}\,\Sigma))\otimes\Sigma^{-1}] \\ &\quad -Tr[U(\boldsymbol{\lambda})\otimes\Sigma] \\ &\quad -2Tr[U(\boldsymbol{\lambda},\boldsymbol{\mu})] \end{aligned} \tag{85}$$

from which the moment generating function is immediately found as

$$M[N(\Theta)(\boldsymbol{x})](\boldsymbol{\lambda}) = exp(\,Tr[U(\boldsymbol{\mu},\boldsymbol{\lambda})]+\frac{1}{2}Tr[U(\boldsymbol{\lambda})\otimes\Sigma]\,) \tag{86}$$

From the above we have

$$\begin{aligned} E[x_i \mid \Theta] &= \mu_i \\ E[(x_i-\mu_i)(x_j-\mu_j) \mid \Theta] &= \Sigma_{ij} \end{aligned} \tag{87}$$

which agrees with the calculation of appendix 12.2. Two things to note: 1. The inverse of Σ is assumed to exist. 2. All moments are determined by simple products and sums of the parameters $(\boldsymbol{\mu},\Sigma)$.

12.3 Multinormal linear change of variables

Letting $\boldsymbol{y}=A\boldsymbol{x}$ be the change of variables, where $P(\boldsymbol{x}\mid\Theta)=N(\Theta)(\boldsymbol{x})$, the MGF of the density $P(\boldsymbol{y}\mid\Theta)$ is found from the MGF of the density for $P(\boldsymbol{x}\mid\Theta)$ in a straightforward manner as

$$\begin{aligned} M[P(\boldsymbol{y}\mid\Theta)](\boldsymbol{\lambda}) &= E[e^{Tr[U(\boldsymbol{\lambda},\boldsymbol{y})]}\mid\Theta] \\ &= E[e^{Tr[U(\boldsymbol{\lambda},A\boldsymbol{x})]}\mid\Theta] \qquad (88) \\ &= E[e^{Tr[U(A^T\boldsymbol{\lambda},\boldsymbol{x})]}\mid\Theta] \\ &= exp(\,Tr[U(\boldsymbol{\mu},A^T\boldsymbol{\lambda})]+\frac{1}{2}Tr[U(A^T\boldsymbol{\lambda})\otimes\Sigma]) \\ &= exp(\,Tr[U(A\boldsymbol{\mu},\boldsymbol{\lambda})]+\frac{1}{2}Tr[U(\boldsymbol{\lambda})\otimes(A\Sigma A^T)]) \qquad (89) \end{aligned}$$

Note that the dropped subscripts x and x of the Θ and λ are easily determined by the context, and that the density used to take the expectation naturally changed in equation 88 from $P(\boldsymbol{y}\mid\Theta)$

to $P(\boldsymbol{x} \mid \Theta)$ without confusion. With this result and referring to equation 86 and preceding we find that the density for $\boldsymbol{y}$ is multinormal with

$$\begin{aligned} \boldsymbol{\mu_y} &= A\boldsymbol{\mu_x} \\ \Sigma_y &= A\Sigma_x A^T \end{aligned} \tag{90}$$

Note that everywhere the condition of A was neither mentioned nor assumed, thus A may be a rectangular matrix or otherwise not of full rank.

12.4 Multinormal projections

Another useful operation is that of projection onto a subset of the components of the argument of the multinormal distribution. Projections may be trivially represented as a linear operation, where the "projection matrix" is typically a rectangular matrix having the form of a unique (single) element of value 1 in each row and column, zeroes elsewhere. Finding the distribution of the projected variables is equivalent to the operation of marginalizing over the components not in the projection. Let A be the projection matrix selecting a subset of the variables of $\boldsymbol{x}$ as $\boldsymbol{y} = A\boldsymbol{x}$. Then, using the result of section 12.3, we immediately find integrals of the form

$$\int N(\boldsymbol{\mu}, \Sigma)(\boldsymbol{x})\, d\boldsymbol{x} \setminus \boldsymbol{y} = N(A\boldsymbol{\mu}, A\Sigma A^T)(\boldsymbol{y}) \tag{91}$$

Both vector $A\boldsymbol{\mu}$ and the matrix $A\Sigma A^T$ are now just appropriately rearranged pieces of the original vector $\boldsymbol{\mu}$ and matrix Σ. Specifically, if $y_k = x_{i_k}$ then $[A\Sigma A^T]_{pq} = \Sigma_{i_p j_q}$.

12.5 Multinormal multiplication

One operation which frequently occurs in Bayesian inference is that of taking the product of two multinormal distributions of the same variable and normalizing that product to find a new distribution. Finding the new $\Theta = (\boldsymbol{\mu}, \Sigma)$ amounts to completing the square, but it is useful to state the result, and we do this here. Let $\Theta_1 = (\boldsymbol{\mu}_1, \Sigma_1)$ and $\Theta_1 = (\boldsymbol{\mu}_1, \Sigma_1)$ be the parameters of the multinormal distributions in the product. Then

$$\begin{aligned} \boldsymbol{\mu} &= \Sigma(\Sigma_1^{-1}\boldsymbol{\mu_1} + \Sigma_2^{-1}\boldsymbol{\mu_2}) \\ \Sigma &= (\Sigma_1^{-1} + \Sigma_1^{-1})^{-1} \end{aligned} \tag{92}$$

12.6 Expected uncertainty in multinormals

It is useful to know the expected uncertainty of one gaussian distribution in the context of another. Consider the quantity

$$E[-log(P(\Theta_2)(\boldsymbol{x})) \mid \Theta_1] = -\int N(\boldsymbol{\mu}_1, \Sigma_1)(\boldsymbol{x})\, log\left(N(\boldsymbol{\mu}_2, \Sigma_2)(\boldsymbol{x})\right) d\boldsymbol{x} \tag{93}$$

which occurs in similar form in the development of the Generalized Kalman Filter (section 6) and represents the expected uncertainty, or entropy, of the surface representation in the context of the updated surface distribution. The value of this integral is found straightforwardly using the results mentioned in appendix 12.2 as

$$\begin{aligned} E[-log(N(\boldsymbol{\mu}_2, \Sigma_2)(\boldsymbol{x})) \mid \Theta_1] &= \frac{1}{2} E\left[Tr[U(\boldsymbol{x} - \boldsymbol{\mu}_2) \otimes \Sigma_2^{-1}]\right] \\ &\quad + \frac{d}{2} log(2\pi) + \frac{1}{2} log(|\Sigma_2|) \\ &= \frac{1}{2} Tr\left[(\Sigma_1 + U(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)) \otimes \Sigma_2^{-1}\right] \\ &\quad + \frac{d}{2} log(2\pi) + \frac{1}{2} log(|\Sigma_2|) \end{aligned} \tag{94}$$

12.7 Maximizing the expected information

Varying Σ_2, the minimum value of the uncertainty above occurs when $\Theta_2 = \Theta_1$. That this is true for the $\boldsymbol{\mu}$ component of Θ_2 is immediate from the positive definite quadratic nature of the first term. For the Σ component the following fact following from the properties of determinants and matrix inverses facilitates the result:

$$\frac{\partial |\Sigma|}{\partial \Sigma_{kl}} = (-1)^{k+l} \frac{Cof_{kl}(\Sigma)}{|\Sigma|} = \Sigma_{kl}^{-1} \tag{95}$$

12.8 Notes on matrix inverses and submatrices

Given the invertible matrix V, composed in the following manner of submatrices V_{11}, V_{12}, V_{21}, V_{22},

$$A = \begin{bmatrix} V_{11} & V_{12} \\ V_{21} & V_{22} \end{bmatrix} \tag{96}$$

and its inverse

$$A^{-1} = \begin{bmatrix} \hat{V}_{11} & \hat{V}_{12} \\ \hat{V}_{21} & \hat{V}_{22} \end{bmatrix} \tag{97}$$

then it is immediate that the following relationships hold among the submatrices

$$\begin{bmatrix} I_{11} & N_{12} \\ N_{21} & I_{22} \end{bmatrix} = \begin{bmatrix} V_{11}\hat{V}_{11} + V_{12}\hat{V}_{21} & V_{11}\hat{V}_{12} + V_{12}\hat{V}_{22} \\ V_{21}\hat{V}_{11} + V_{22}\hat{V}_{21} & V_{21}\hat{V}_{12} + V_{22}\hat{V}_{22} \end{bmatrix} \tag{98}$$

where I and N represent the identity and zero matrices respectively. Any quadratic operator $\boldsymbol{x}^T Q \boldsymbol{x}$ may be decomposed using projection matrices A and $\overline{A}$ where these are diagonal matrices

with one and zero entries only, and where

$$A + \overline{A} = I \tag{99}$$

in the following manner

$$\begin{aligned} \boldsymbol{x}^T Q \boldsymbol{x} &= \boldsymbol{x}^T (A + \overline{A}) Q (A + \overline{A})^T \boldsymbol{x} \\ &= \boldsymbol{x}_A^T Q_{AA} \boldsymbol{x}_A + \boldsymbol{x}_A^T Q_{A\overline{A}} \boldsymbol{x}_{\overline{A}} + \boldsymbol{x}_{\overline{A}}^T Q_{\overline{A}A} \boldsymbol{x}_A + \boldsymbol{x}_{\overline{A}}^T Q_{\overline{A}\overline{A}} \boldsymbol{x}_{\overline{A}} \end{aligned} \tag{100}$$

Now, assume Q is symmetric and that both it and Q_{AA} and $Q_{\overline{A}\overline{A}}$ are invertible, and rewrite this form as the sum of two terms as follows

$$\begin{aligned} \boldsymbol{x}^T Q \boldsymbol{x} &= (\boldsymbol{x}_A - \boldsymbol{\alpha})^T Q_{AA} (\boldsymbol{x}_A - \boldsymbol{\alpha}) + C(\boldsymbol{x}_{\overline{A}}) \\ &= \boldsymbol{x}_A^T Q_{AA} \boldsymbol{x}_A - \boldsymbol{x}_A^T Q_{A\overline{A}} \boldsymbol{x}_{\overline{A}} - \boldsymbol{x}_{\overline{A}}^T Q_{\overline{A}A} \boldsymbol{x}_A + \boldsymbol{\alpha}^T Q_{AA} \boldsymbol{\alpha} + C(\boldsymbol{x}_{\overline{A}}) \end{aligned} \tag{101}$$

where $\boldsymbol{\alpha} = (Q_{AA})^{-1} Q_{A\overline{A}} \boldsymbol{x}_{\overline{A}}$. Thus

$$C(\boldsymbol{x}_{\overline{A}}) = \boldsymbol{x}_{\overline{A}}^T \left(Q_{\overline{A}\overline{A}} - Q_{\overline{A}A} (Q_{AA})^{-1} Q_{A\overline{A}} \right) \boldsymbol{x}_{\overline{A}} \tag{102}$$

Applying the identities of equation 98

$$Q_{AA} \hat{Q}_{A\overline{A}} + Q_{A\overline{A}} \hat{Q}_{\overline{A}\overline{A}} = N_{A\overline{A}} \tag{103}$$

followed by

$$Q_{\overline{A}A} \hat{Q}_{A\overline{A}} + Q_{\overline{A}\overline{A}} \hat{Q}_{\overline{A}\overline{A}} = I_{\overline{A}\overline{A}} \tag{104}$$

find that

$$Q_{\overline{A}\overline{A}} - Q_{\overline{A}A} (Q_{AA})^{-1} Q_{A\overline{A}} = (\hat{Q}_{\overline{A}\overline{A}})^{-1} \tag{105}$$

so that

$$C(\boldsymbol{x}_{\overline{A}}) = \boldsymbol{x}_{\overline{A}}^T (\hat{Q}_{\overline{A}\overline{A}})^{-1} \boldsymbol{x}_{\overline{A}} \tag{106}$$

which immediately provides an alternate method for marginalizing gaussian distributions.

12.9 Alternate inverse forms

In the GKF update equations expressions for updating inverse matrices in terms of the sum of other inverse matrices occur. Because one of the summand matrices may not be well-conditioned, it is of interest to find an expression for the updated matrix in terms of the other matrices, which

explicitly is not a function of the inverse matrices. Thus, let P, Q, R be invertible matrices such that

$$P^{-1} = Q^{-1} + R^{-1} \tag{107}$$

Then we find

$$P = Q - Q(Q+R)^{-1}Q \tag{108}$$

by the following direct substitution

$$\begin{aligned} PP^{-1} &= (Q - Q(Q+R)^{-1}Q)(Q^{-1} + R^{-1}) \\ &= I - Q\left[(Q+R)^{-1}(I + QR^{-1}) - R^{-1}\right] \\ &= I \end{aligned} \tag{109}$$

12.10 Nonlinear forward projection

In the nonlinear forward projection case the projection is given by $\boldsymbol{f}(\boldsymbol{s})$, where $\boldsymbol{f}(\cdot)$ is a nonlinear function of $\boldsymbol{s}$ rather then the linear form $M\boldsymbol{s}$. Because the derivative of the forward projection is often a straightforward object to compute, expand $\boldsymbol{f}(\boldsymbol{s})$ about the mean of the old surface, $\boldsymbol{\mu}_s$

$$\boldsymbol{x} = \boldsymbol{f}(\boldsymbol{\mu}_s) + \frac{\partial \boldsymbol{f}}{\partial \boldsymbol{s}} \mid_{\boldsymbol{\mu}_s} (\boldsymbol{s} - \boldsymbol{\mu}_s) + \epsilon \tag{110}$$

Letting $M = \frac{\partial \boldsymbol{f}}{\partial \boldsymbol{s}} \mid_{\boldsymbol{\mu}_s}$ we have

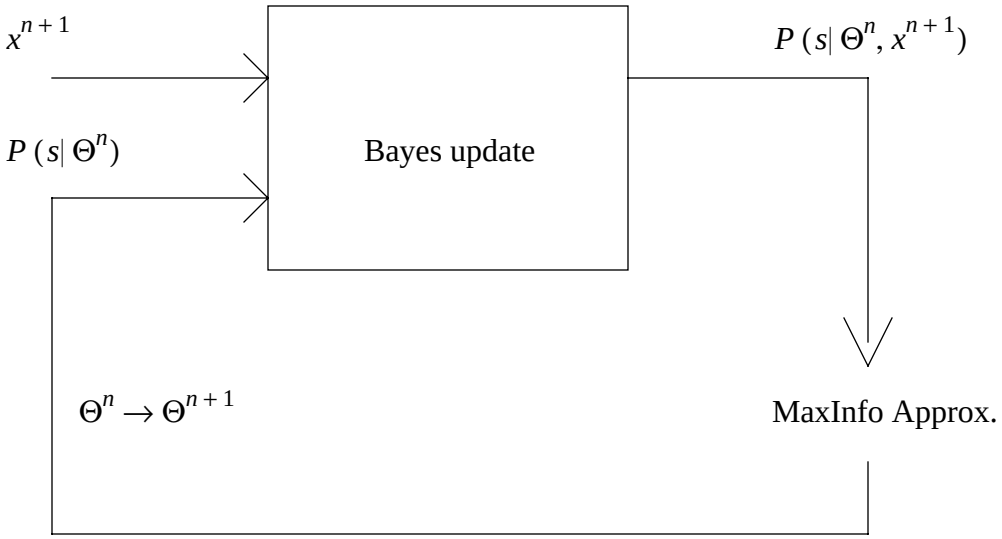
$$\begin{aligned} P(\boldsymbol{x} \mid \boldsymbol{s}, \phi) &= N((\boldsymbol{f}(\boldsymbol{\mu}_s) - M\boldsymbol{\mu}_s) + M\boldsymbol{s}, \Sigma_\epsilon)(\boldsymbol{x}) \\ &= N(M\boldsymbol{s}, \Sigma_\epsilon)(\boldsymbol{x} - (\boldsymbol{f}(\boldsymbol{\mu}_s) - \boldsymbol{M}\boldsymbol{\mu}_s)) \end{aligned} \tag{111}$$

so that the appropriate changes to be made to the GKF update equations are simply

$$\begin{aligned} \boldsymbol{x} &\rightarrow \boldsymbol{x} - (\boldsymbol{f}(\boldsymbol{\mu}_s) - \boldsymbol{M}\boldsymbol{\mu}_s) \\ M &\rightarrow \frac{\partial \boldsymbol{f}}{\partial \boldsymbol{s}} \mid_{\boldsymbol{\mu}_s} \end{aligned} \tag{112}$$

while everything else otherwise remains the same.

GKF Update Loop Equation



The elements going into $P(s|\Theta^n)$ are the prior, restricted to some knowledge H about the field, $P(s|H)$. (In the main text example, H is the set of known surface height field values.) and the Knowledge Representation (KR) distribution is $P(H|\Theta^n)$, which is the learned knowledge about the specifics of the surface at the n'th iteration of the GKF.

These form the approximate posterior $P(s|\Theta^n)$ given by the integral over H of the product of the KR distribution and the prior distribution given H known, that is

$$P(s|\Theta^n) = \int P(s|H)\, P(H|\Theta^n)\, dH \quad (1)$$

At update $n+1$, the new data and the approximate posterior from iteration n are incorporated using the likelihood $P(x^{n+1}|s)$ and Bayes' theorem to produce the data-dependent posterior written $P(s|\Theta^n, x^{n+1})$. Then, the new KR that caputres an approximation to this exact posterior using (1) above with $n \rightarrow n+1$ via Maximally informative statistical inference completes the GKF loop.

Figure 1 - Generalized Kalman Filter Update Loop

References and Notes

1. Brown, R. G. *Introduction to Random Signal Analysis and Kalman Filtering;* Wiley: New York, 1983.
2. Geman, S.; Geman, D. Stochastic relaxation, Gibbs distribution, and the Bayesian restoration of images. *IEEE Trans. Patt. Anal. Mach. Intell.* **1984**, *6*, 721-741.
3. Rosenblatt, M. *Random Processes*; 2nd Ed; Springer-Verlag: New York, 1974.
4. Wolf, D. R. *Information and Correlation in Statistical Mechanical Systems*; PhD Dissertation, Physics Department, University of Texas at Austin, USA, 1996.
5. Wolf, D. R.; George, E. I. Maximally Informative Statistics. *Revista de la Real Academia de Ciencias* (Madrid, Spain. Special edition on Bayesian Statistics) **2000**, *94*, in press.

© 1999 by the author. Reproduction for noncommercial purposes permitted.