On the uncertainty in the regularized solution

2011 J. Phys. A: Math. Theor. 44 023001

(http://iopscience.iop.org/1751-8121/44/2/023001)

IOP PUBLISHING JOURNAL OF PHYSICS A: MATHEMATICAL AND THEORETICAL

J. Phys. A: Math. Theor. **44** (2011) 023001 (28pp) doi:10.1088/1751-8113/44/2/023001

TOPICAL REVIEW

On the uncertainty in the regularized solution

Daniel Keren[1] **and Ido Nissenboim**

Computer Science Department, Haifa University, Haifa 31905, Israel

E-mail: dkeren@cs.haifa.ac.il and idoniss@gmail.com

Received 22 October 2010
Published 7 December 2010
Online at stacks.iop.org/JPhysA/44/023001

Abstract

Regularization is a fundamental technique in the processing of measurement data. These data are typically noisy and sparse, and in some cases, only a projection of the data on a low-dimensional subspace is available. Very often, direct analysis of measured data yields unphysical results, due to sensitivity to noise and neglect of the underlying distribution of the measured physical phenomena. Therefore, regularization is extensively applied in physics. One of the most common regularization methods was developed by a few researchers (most notably Tikhonov and Phillips) during the 1940s–1960s. In the early 1980s, Geman and Geman reformulated regularization in a Bayesian framework, in which every candidate function $f(x)$ is assigned a probability, with the expectation of the random variable $f(x_0)$ equal to the value of the regularized interpolating function at x_0. Taking this probabilistic approach 'to second order', the *variance* of this random variable, which reflects the amount of uncertainty inherent in the solution, is computed. When estimating physical phenomena, not only the estimated value is important, but also the *uncertainty* associated with it (for example, more measurements should yield a more certain solution). This review provides a formal definition of this uncertainty, explores some numerical issues with its computation, and extends it to compute the uncertainties of the integral and derivative, with the intriguing result that in some cases the uncertainty of the function and the derivative are highly uncorrelated—for example, there are cases in which the point at which the function is *least* reliable is the one at which the derivative is *most* reliable. Lastly, a related definition of optimal sampling is discussed. The relevance of optimal sampling to physics is that it offers a method to determine the location of the measuring devices which yields the most reliable restoration of the measured physical phenomenon.

PACS numbers: 02.60.Ed, 02.60.Gf

(Some figures in this article are in colour only in the electronic version)

[1] Author to whom any correspondence should be addressed.

1. Interpolation and regularization

Interpolation is a general term for algorithms used to estimate an entire function, given measured values at some (typically sparse) set of points. It is widely applied in practically all sciences which use functional representations of data. The problem's input is usually a set of values sampled from the graph of a function (these values may be accurate or contaminated by noise), and the output is a complete function—the interpolant—which should be a good approximation to the sampled, unknown function. The interpolant is restricted to lie in a class of functions which shall be referred to as the *admissible functions*. For simplicity, we will always assume that we are dealing with functions defined on the unit interval [0, 1].

The most direct approach is to fit a simple function—usually a low-degree polynomial—to the data, and use it as an estimate to the unknown function. However, this method is usually over-restrictive because it assumes that the interpolated function belongs to a very small class of functions.

On the other hand, extending the class of functions can also be a problem, because there are simply too many possible interpolants; which one should be chosen? What should be the criterion for preferring one interpolant over the other?

Unless some restrictions are enforced on the oscillation of the interpolating function, nothing meaningful can be achieved; every value is made possible at every interpolated point by choosing an appropriate function. The question is: what kind of restriction should be applied on the oscillation of the interpolant? It is not enough to demand that the admissible functions be infinitely differentiable, as such functions can behave in an oscillatory manner; for example, the infinitely differentiable functions used for *partition of unity* [1] have the property that they assume a value of zero at all points of Euclidean space except for an arbitrarily small box, and in the inner half-box the value is 1. This means that no correlation exists between the function's values at different spatial locations. Such correlation exists (albeit in a restricted manner) for *analytic* functions, for knowing the values of an analytic function at an arbitrary domain of the Euclidean space allows us to reconstruct it in the entire space; however, it is too restrictive to expect natural phenomena to behave in an analytical manner.

A viable class of functions are those with a bounded derivative. Using such functions imposes a bound on the oscillation, and enforces a correlation (decreasing with the distance) between different values of the interpolated function. However, it is not easy to enforce the condition of the bounded derivative; worse, these functions do not form a linear space. A more useful measure for the oscillation of a function is the integral of the square of the first or second derivative. It is also easier to work with such an expression than with the maximal value of the derivative, since taking the derivative's maximal value is not a differentiable operator on the space of functions. Such considerations led to the introduction of *regularized* interpolants (often referred to as *splines*) [33, 45, 46], which will now be described.

1.1. Regularization

In computer vision, regularization is used to reconstruct objects from partial data [8, 21, 43, 44, 46]. Reconstruction of functions given partial information has been studied in many other fields, for example petroleum exploration [37], geology [11], electronics [4] and physics (see section 8). The regularization approach overcomes the difficulty of choosing 'the' best interpolant—and also offers a 'reasonable' solution—by choosing among the possible functions one which approximates the given data and is also 'slowly changing'. This embodies an important assumption that less oscillatory functions are more probable.

Formally, a *cost functional* $M(f)$ is defined for every function f by $M(f) = D(f) + \lambda S(f)$, where $D(f)$ measures the deviation of f from the given measurements,

$S(f)$ measures the roughness of f and λ is a positive parameter. The f chosen is the one minimizing $M()$. A typical definition of these functionals is the following: if the measured (up to some degree of accuracy) values of the function at a set of points $\{x_i\}_{i=1}^{l}$ (from now on called *sampling points*) are denoted by $\{y_i\}_{i=1}^{l}$, then $D(f) = \sum_{i=1}^{l} \frac{1}{2\sigma_i^2}[f(x_i) - y_i]^2$ and $S(f) = \int_0^1 ([f''(v)]^2)\,\mathrm{d}v$, where σ_i is an indicator for the measurement reliability at the point (x_i) (for simplicity it will be assumed hereafter that $2\sigma_i^2 = 1$). As λ increases, the resulting interpolant tends to be smoother—possibly at the cost of not being close to the sample points (here, as is common in the regularization literature, 'smooth' means 'slowly changing', not infinitely differentiable). The f minimizing $M()$ can be found by variational methods.

Once the interpolant is computed, a natural question is: how certain are its values? This review suggests a solution to this question. Suppose f is the interpolant and $f(x) = y$. When there are many functions g such that $M(g)$ is almost equal to $M(f)$ but $g(x)$ substantially differs from y, the value y should be assigned a high degree of uncertainty. When a point x is close to many sampling points, its interpolated value should be more reliable than the values of points situated farther away from the sampling points. Thus, the measure of uncertainty should be pointwise and not global.

From the practical point of view, the importance of calculating this uncertainty is clear: typically, one is interested not just in the function describing some natural phenomena, but also in how reliable the values of this function are. For example, a weather forecast is not worth much without some measure of trust associated with it.

An attempt is made to solve this problem in a probabilistic setting. Each function f is assigned a probability ('prior') proportional to $\exp(-M(f))$, thus generating a probability distribution over all functions. The motivation for this definition follows from a Boltzmann-like distribution [17]. Using this probability structure, the distribution of the f-values at x is computed for each x.

Formally, the *evaluation at x* (assigning each function its value at the point x) is viewed as a random variable on the space of admissible functions, and its variance is computed.

It turns out that the same method used for estimating the uncertainty of the function's value at a point can be extended to estimate the uncertainty of any linear functional on the function space, for example the derivative. It will be demonstrated that sometimes the point at which the function's value is least reliable is the point at which the derivative is most reliable.

2. Previous work

This section surveys several definitions of the notion of uncertainty, in general function spaces and in the framework of regularization and their relation to this review.

Some of the work described here appeared in [22], but in a limited context and without proofs. Other relevant work in which the first author participated is [23]. Other approaches to the problem are now discussed.

2.1. General treatment of uncertainty in Banach spaces

A general approach for estimating the uncertainty of linear functionals on function spaces was introduced in [27] and further developed in [28, 48]. The question posed is: given partial data about a function, what can be said about the value of some linear functional on that function? This problem is answered in two cases: *worst case setting* and *average case setting*.

In the worst case setting, the goal is to give bounds on the value of the functional which are broad enough to include all admissible functions. As a simple example, suppose the

admissible functions consist of the differentiable functions on the interval [0, 1] such that the absolute value of their derivative is bounded by 1, and the partial information given is the value $f(0)$. If, for instance, the linear functional to estimate is the value at the point 1, it can be bounded inside the interval $[f(0) - 1,\ f(0) + 1]$. In a similar fashion, the integral and other linear functionals can be bounded.

It turns out that in order to provide a worst case bound one usually has to impose overly restrictive conditions on the space of functions involved. An especially disturbing fact is that this restrictive space is typically not a linear space. Also, the worst case bound depends only on one function (the one at which this 'worst case' is manifest), which can be a pathological and 'non-probable' element of the function space. In our case, for instance—where the functional is the value of an interpolating function—the worst case uncertainty at each point except for the sample points will be infinite, because there are always functions passing through the sample points but accepting arbitrary values at other points.

For these reasons, the average case error is defined and studied. Here, every function is assigned a weight (in [26] this is the *Wiener measure*) and this weight is taken into account when computing the possible values of a functional. This allows enlarging the class of functions, as pathological functions are usually assigned a low weight and thus do not have much influence on the overall uncertainty.

While [26] shares some ideas with the work presented in this review, it differs in some key points:

(1) the probability overlaid on the space of functions is based on the Wiener measure, which is not appropriate for the interpolation problem;
(2) it does not introduce the concept of a variance and joint normal distribution presented in this work;
(3) the Wiener measure does not favor smooth objects over 'rough' ones, and thus its use in connection with regularization is questionable;
(4) it is not clear how the case of noisy measurements can be incorporated into the Wiener measure formalism.

2.2. *Diffused priors*

A different approach for estimating the uncertainty is adopted in [49]. A prior distribution on the values the admissible functions assume at t is defined as the following stochastic process:

$$X(t) = \sum_{j=1}^{m} \theta_j \phi_j(t) + \sqrt{b} Z(t),$$

where $\phi_j(t) = \frac{t^{j-1}}{(j-1)!}$, $\{\theta_1 \ldots \theta_n\}$ is a Gaussian noise term, b a constant and $Z(t)$ the iterated Wiener process [26, 49]. The first summand is the 'diffused' part of the process, and relates to the 'fractal prior' of section 2.3. This process is used to describe the spline as a stochastic process.

This approach differs from the one presented here in the structure of the probability space, for it does not define a probability distribution on the function space.

2.3. *Fractal priors*

This approach [42] was applied to problems in computer vision and computer graphics. In order to estimate the uncertainty associated with each point in the interpolant, the interpolation process [43] is performed many times, each time with the addition of random noise, thus

generating many samples from the class of all surfaces. After the samples are generated, the spread of the values at each point is used to give a measure of uncertainty at that point—the higher the spread, the larger the uncertainty. It is also proved that the samples thus created are fractals.

This work differs from the one presented here in that it is heuristic in nature and does not formalize the concept of variance. Also, it is not clear why the noised iterative process generates random samples from the space of admissible functions.

3. Mathematical analysis of the uncertainty

In order to compute the uncertainty, first one has to define the probability of a function. According to Bayes' rule

$$\mathbf{Prob}(f/D) = \frac{\mathbf{Prob}(D/f)\mathbf{Prob}(f)}{\mathbf{Prob}(D)},$$

where f is the function and D the sampled data. When Gaussian error is assumed in the measurement process, the following holds:

$$\mathbf{Prob}(D/f) = \exp\left(-\sum_{i=1}^{l}[f(x_i) - y_i]^2\right),$$

where, as noted before, Gaussian noise with variance σ such that $2\sigma^2 = 1$ is assumed.

A crucial question is how to define the probability of a function ('prior on the functions' in the Bayesian terminology). Here the following definition is used:

$$\mathbf{Prob}(f) = \exp\left(-\lambda \int [f''(v)]^2 \, \mathrm{d}v\right),$$

where λ is a positive constant. Since the integral of the second derivative squared is a rough approximation to the bending energy of a thin rod, the probability can be viewed as a Boltzmann distribution with λ proportional to the inverse of the temperature. Combining both terms yields

$$\mathbf{Prob}(f/D) \propto \exp(-M(f)), \tag{1}$$

where

$$M(f) = \sum_{i=1}^{l}[f(x_i) - y_i]^2) + \lambda \int [f''(v)]^2 \, \mathrm{d}v. \tag{2}$$

This justifies—in hindsight—the choice of the interpolant as the function which minimizes $M()$, which was common long before the Bayesian interpretation was introduced.

In physics, the restored function represents a physical phenomenon, and the prior should be chosen to reflect the real-life behavior, or distribution, of this phenomenon. Choosing the best prior is a fascinating and crucial question, and it greatly affects the restoration. We discuss this issue in section 8.1. The prior we use here, which penalizes functions with a large second derivative, is often referred to as the *second-order smoothness term* (see the following section), but the technique presented in this review can be applied to handle other priors (section 8.2).

3.1. First versus second derivative

The prior we study here makes use of the integral of the second derivative squared. The Wiener measure uses a first derivative. The methods developed in this review can be applied

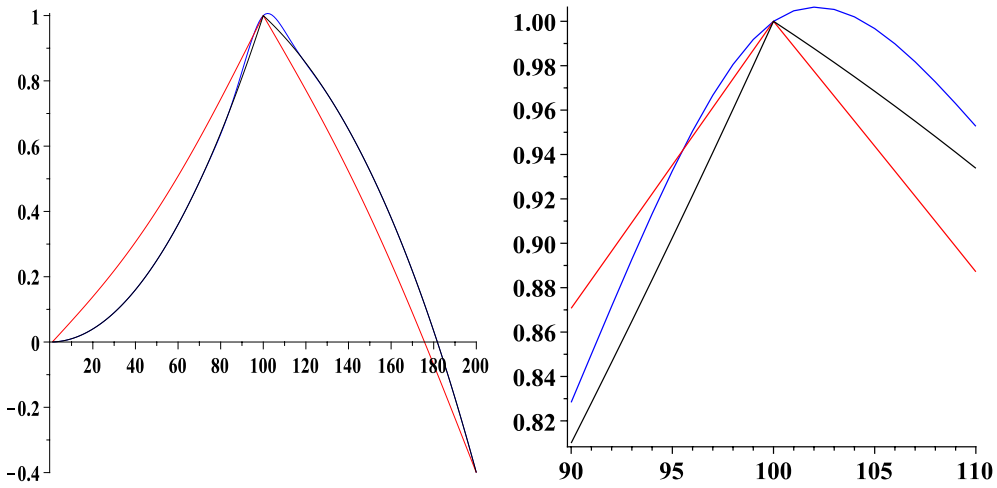

Figure 1. Left: in black, a function defined (up to a scaling factor) by t^2 in the interval [0, 100], and $1.2 - 0.2t^3$ in the interval [100, 200]. In red, the function after being smoothed with a first-order smoothness term. In blue, after applying second-order smoothing. While the first-order smoothing yields a piecewise linear function, second-order smoothing preserves the general shape of the curve, 'smoothing out' only the cusp in the middle. Right: same, with a zoom on the middle part of the curves.

to either case; the choice of what smoothness term to use depends on the natural phenomena in question. The first-order term is known to 'push' the solution towards a linear piecewise function, while the second-order term is 'softer', allowing 'curvy' solutions. In computer vision, the common wisdom is that when restoring surfaces, the second-order term yields better results. Figure 1 contains typical results of regularization results with first- and second-order terms. For applications of the second-order smoothness term in physics, see e.g. [7, 13, 15, 41]. In the frequency domain, both priors can be viewed as enforcing a higher penalty on higher frequencies in the restored function, with the first-order penalty of the frequency k being proportional to k and that of the second-order term being proportional to k^2. This may also be thought of as a 'soft threshold' on high frequencies, as opposed to e.g. the sharp ultraviolet cutoff. It can also be viewed as a mathematical representation of optical or electronic filters used to attenuate high-frequency components. For a general discussion of priors used in physics, see section 8.1.

3.2. The case of a finite-dimensional subspace

In this section the two questions posed in the introduction—what interpolant should be chosen and how to define the uncertainty associated with it—will be solved when the interpolants are restricted to a finite-dimensional subspace. Given a set of points $\{(x_i, y_i)\}_{i=1}^l$, $0 \leqslant x_i \leqslant 1$, the regularization approach is to find the function f minimizing the cost functional $M(f) = \sum_{i=1}^l [f(x_i) - y_i]^2 + \lambda \int_0^1 [f''(v)]^2 \, dv$. To simplify indexing, the solution will be carried out for a two-dimensional subspace, span$\{f, g\}$, where f and g are linearly independent. For every function in this space $M(af+bg) = \sum_{i=1}^l [af(x_i)+bg(x_i) - y_i]^2 + \lambda \int_0^1 [af''(v)+bg''(v)]^2 \, dv$. This is a bilinear form in a and b which can be written as $(a, b)\Delta(a, b)^T + C_a a + C_b b + C$, where Δ is a 2×2 matrix given by

$$\Delta_{11} = \sum_{i=1}^{l} f^2(x_i) + \lambda \int_0^1 [f''(v)]^2 \,\mathrm{d}v, \quad \Delta_{22} = \sum_{i=1}^{l} g^2(x_i) + \lambda \int_0^1 [g''(v)]^2 \,\mathrm{d}v,$$

$$\Delta_{12} = \Delta_{21} = \sum_{i=1}^{l} f(x_i)g(x_i) + \lambda \int_0^1 f''(v)g''(v) \,\mathrm{d}v,$$

$$C_a = -2\sum_{i=1}^{l} y_i f(x_i), \quad C_b = -2\sum_{i=1}^{l} y_i g(x_i), \quad C = \sum_{i=1}^{l} y_i^2.$$

As noted before, the *probability* of $af + bg$ is defined to be $\frac{1}{K}\,\mathrm{e}^{-M(af+bg)}$, and the *expectation at the point* x, denoted by E_x, will therefore be $\frac{1}{K}\int_{\mathcal{H}} h(x)\,\mathrm{e}^{-M(h)}\,\mathrm{d}h$ where $\mathcal{H} = \mathrm{span}\{f, g\}$ and $K = \int_{\mathcal{H}} \mathrm{e}^{-M(h)}\,\mathrm{d}h$ is the normalizing factor. Substituting the explicit expression for $M()$ gives

$$E_x = \frac{\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}[af(x) + bg(x)]\,\mathrm{e}^{-[(a,b)\Delta(a,b)^T + C_a a + C_b b + C]}\,\mathrm{d}a\,\mathrm{d}b}{\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\mathrm{e}^{-[(a,b)\Delta(a,b)^T + C_a a + C_b b + C]}\,\mathrm{d}a\,\mathrm{d}b}$$

(note that the natural identification of $\mathcal{H}$ and $\mathcal{R}^2$ is used). Using standard Gaussian integration methods, E_x turns out to be $-\frac{1}{2}(C_a, C_b)\Delta^{-1}(f(x), g(x))^T$. However, this result can be obtained more easily by looking for the pair a, b minimizing $M(af + bg)$; they are readily seen to be $(a, b) = -\frac{1}{2}(C_a, C_b)\Delta^{-1}$. So the value at x of the cost minimizing function is $-\frac{1}{2}(C_a, C_b)\Delta^{-1}(f(x), g(x))^T$, which is equal to the expectation. This is also a standard result in Gaussian integral theory.

The measure of uncertainty, or *variance at* x, V_x, is defined as

$$V_x = \frac{\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}[af(x) + bg(x) - E_x]^2\,\mathrm{e}^{-[(a,b)\Delta(a,b)^T + C_a a + C_b b + C]}\,\mathrm{d}a\,\mathrm{d}b}{\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\mathrm{e}^{-[(a,b)\Delta(a,b)^T + C_a a + C_b b + C]}\,\mathrm{d}a\,\mathrm{d}b}.$$

After some manipulations this expression turns out to be $\frac{1}{2}(f(x), g(x))\Delta^{-1}(f(x), g(x))^T$. Similarly this result is true for any finite-dimensional subspace.

Lemma 1. *If* $\mathcal{H} = span\{f_1, f_2 \ldots f_n\}$ *where* $\{f_1, f_2 \ldots f_n\}$ *are linearly independent and the* $n \times n$ *matrix* Δ *is defined by* $\Delta_{i,j} = \sum_{k=1}^{l} f_i(x_k) f_j(x_k) + \lambda \int_0^1 f_i''(v) f_j''(v)\,\mathrm{d}v$, *and* $C_j = -2\sum_{i=1}^{l} y_i f_j(x_i)$, *then*

$$\begin{aligned} E_x &= -\tfrac{1}{2}(C_1, C_2 \ldots C_n)\Delta^{-1}(f_1(x), f_2(x) \ldots f_n(x))^T \\ V_x &= \tfrac{1}{2}(f_1(x), f_2(x) \ldots f_n(x))\Delta^{-1}(f_1(x), f_2(x) \ldots f_n(x))^T. \end{aligned} \tag{3}$$

Note that V_x does not depend on the measured values $\{y_i\}_{i=1}^{l}$, but only on the sampling point locations $\{x_i\}_{i=1}^{l}$.

A simple example follows: suppose $\mathcal{H} = \mathrm{span}\{1, x^2\}$ and $l = 2$. Then

$$V_x = \frac{4\lambda + 2x^4 - 2\left(x_1^2 + x_2^2\right)x^2 + x_1^4 + x_2^4}{8\lambda + \left(x_1^2 - x_2^2\right)^2}.$$

So if the two points x_1 and x_2 are very close, the estimate of the function's value is very unreliable. Also, as the weight λ given to the 'roughness term' of the cost functional is smaller, the reliability decreases.

An interesting quantity, measuring the uncertainty over the entire interval, is

$$\int_0^1 V_x \mathrm{d}x = \frac{60\lambda + 15\left(x_1^4 + x_2^4\right) - 10\left(x_1^2 + x_2^2\right)}{120\lambda + 15\left(x_1^2 - x_2^2\right)^2}$$

(this is an ordinary integral over the unit interval, unrelated to the probability measure).

When for example $\lambda = 1$ and $x_1 = 0$, this expression has a minimum for $x_2 = 0.83$. So if the function was already sampled at the point 0, sampling it again at 0.83 will minimize the overall uncertainty. This is of course not true for all functions spaces; in section 6 the question of *optimal sampling* in the general case will be treated.

4. The infinite-dimensional case

In this section the problem is treated in its full generality, and the class of admissible functions is extended to include all the functions for which the roughness term $\int_0^1 [f''(v)]^2 \, dv$ is defined. After a brief review of Hilbert space theory, this *Sobolev space* of admissible functions is introduced, as well as some notions of integration on Hilbert space. The computation of the variance of a linear functional is provided in an abstract setting using a reproducing kernel and the norm of a linear functional with an appropriate inner product. Then, some computational issues are addressed; it is shown that the variance is a sixth-degree polynomial, and a lower bound is provided on the difference between the variance on the entire space and certain finite-dimensional subspaces.

4.1. Some notions of Hilbert space

In order to extend the definitions of expectation and variance to the space of all admissible functions, one should work in the appropriate Hilbert space. Some basic definitions are now provided.

A *real Hilbert space* H is a vector space over the reals $\mathcal{R}$ with an *inner product* $(\cdot, \cdot) : H \times H \to \mathcal{R}$ satisfying for all $x, y, z \in H$ and $t \in \mathcal{R}$:
$(x, x) \geqslant 0$ and $(x, x) = 0$ iff $x = 0$,
$(x, y) = (y, x)$,
$(tx, y) = t(x, y)$,
$(x, y + z) = (x, y) + (x, z)$,
for example, for $x = (x_1, x_2 \ldots x_n)$, $y = (y_1, y_2 \ldots y_n) \in \mathcal{R}^n$, $(x, y) = \sum_{i=1}^n x_i y_i$ defines an inner product on $\mathcal{R}^n$.

The *norm* of $x \in H$, $\|x\|$, is defined as $\sqrt{(x, x)}$.

Two norms, $\|x\|_1$ and $\|x\|_2$, on H are called *equivalent* if there exist constants C and K such that for all $x \in H$ the following inequality holds: $C\|x\|_1 \leqslant \|x\|_2 \leqslant K\|x\|_1$.

A norm $\| \ \|$ can be used to define a *metric*—e.g. a distance function—on the set of pairs $\{(x, y)\}$ of elements of H, where the distance between x and y is $\|x - y\|$. It is easy to see that if two norms are equivalent then the topologies induced by them are the same.

A well-known theorem in Hilbert space theory is the *Cauchy–Schwarz inequality*: $|\ (x, y)\ | \leqslant \|x\|\|y\|$.

There is another requirement for H to be a Hilbert space, demanding that Cauchy sequences converge in it. For an introduction to Hilbert space, see [50].

Let us also recall the notion of a *Gram matrix* of elements $\{x_1, x_2, \ldots x_n\}$ of H: it is defined as the $n \times n$ matrix G with $G_{i,j} = (x_i, x_j)$. It is well known that if the elements $\{x_1, x_2, \ldots x_n\}$ are linearly independent, their Gram matrix is positive definite.

A *linear functional* on H is a linear mapping $L : H \to \mathcal{R}$. The *norm* of L, $\|L\|$, is defined as $\sup_{x \in H} \frac{L(x)}{\|x\|}$. If it is finite L will be called *bounded*. An important theorem about bounded linear functionals is the following.

Riesz representation theorem: if $L{:}H \to \mathcal{R}$ is a bounded linear functional, there exists a unique element y of H such that $(x, y) = L(x)$ for all $x \in H$. Furthermore, $\|L\| = \|y\|$.

4.2. Probability distributions in the infinite-dimensional case

In order to generalize the results of the previous section to the infinite-dimensional case, some notions about measure, probability and integration in Hilbert spaces need to be introduced.

A measure that assigns to a function f a probability proportional to $\exp(-M(f))$ is required. It is known that there exists a measure assigning to f a probability proportional to $\exp(-\|f\|^2)$, for every norm $\|f\|$. It will be shown that it is possible to define an inner product on the function space such that $M(f)$ and $\|f\|^2$ differ only by a translation, and thus this measure is good enough.

In the finite-dimensional case the required measure is quite simple. Suppose the space, H, is equal to $\text{span}\{f_1, \ldots, f_n\}$. Then it can be naturally identified with $\mathcal{R}^n$, simply by identifying $f = a_1 f_1 + \cdots + a_n f_n$ with the vector $(a_1 \ldots a_n) \in \mathcal{R}^n$. As was the case in the two-dimensional example in the preceding section, the probability density of f (up to translation and a multiplicative constant) is $\exp(-(a_1 \ldots a_n)\Delta(a_1 \ldots a_n)^T)$, where Δ is a positive definite matrix. So, an inner product on $\mathcal{R}^n$ can be defined by $(x, y) = x\Delta y^T$. Using this inner product, the probability of a set of vectors E (which are also viewed as functions) is, up to a normalizing factor, $\int_E \exp(-\|x\|^2)\,\mathrm{d}x$.

It is desirable to extend this probability measure to the infinite-dimensional case. The question is: how an integral over a subset E of an infinite-dimensional function space can be computed? The main results needed to define and compute such integrals are now introduced; for a more thorough survey, see [18, 25, 27].

The first step in the construction of any probability measure over a space is the definition of 'atomic' elements of the space and the definition of their measure. In $\mathcal{R}^n$ these are boxes, e.g., products of intervals; the 'atoms' of the infinite-dimensional space are 'generalized boxes'. Formally, a *tame set* in a real separable Hilbert space H is the collection of elements h satisfying $((y_1, h) \ldots (y_n, h)) \in E$, where $y_i \in H$ are linearly independent and E is a Borel set in $\mathcal{R}^n$. The measure of such a set is defined by

$$\nu\{h \in H : ((y_1, h), (y_2, h) \ldots (y_n, h)) \in E\} = \frac{|\Sigma|^{-1/2}}{(\sqrt{2\pi})^n} \int_E \mathrm{e}^{-\frac{1}{2} x \Sigma^{-1} x^t} \mathrm{d}x \qquad (4)$$

where $y_i \in H$, $\Sigma_{i,j} = (y_i, y_j)$ (the Gram matrix of y_i). It is not difficult to see that the definition does not depend on the choice of y_i's.

To understand the meaning of this measure, one starts with the finite-dimensional case, i.e. when it is defined over $\mathcal{R}^n$. Suppose for simplicity that y_i is the projection on the ith coordinate. When viewed as a vector (via the Riesz representation theorem) y_i is the ith unit vector. Now the condition $((y_1, h) \ldots (y_n, h)) \in E$ simply translates to $h \in E$, Σ is the identity matrix and thus—up to a constant—$\nu(E) = \int_E \exp(-\|x\|^2/2)\,\mathrm{d}x$. This is called a *Gaussian measure* on $\mathcal{R}^n$. In the infinite-dimensional case, the situation is different: if one looks at the space as a box with an infinite number of edges, then the limitation $((y_1, h) \ldots (y_n, h)) \in E$ restricts only a finite number of these edges, and what is left is a box with an infinite number of unbounded edges. Actually, this 'unbounded part' of the box is discarded, and the measure of the finite-dimensional bounded part is computed as if it was contained in $\mathcal{R}^n$. The intuitive explanation is that, as the integrand is an exponential function, the integral can be viewed as the product of two integrals, one over the bounded part and other over the unbounded one. However, since the integral has to be normalized by the integral of the Gaussian over the entire space, then it follows that the integral over the unbounded part cancels out and the bounded part is left, e.g., an integral over $\mathcal{R}^n$. This is not a formal proof; however, for that one can consult the literature.

Next, the class of functions that can be integrated with this measure is introduced. Again, the key point is that only functions that depend only on a finite number of coordinates can be treated, and these functions are integrated as functions over $\mathcal{R}^n$.

A function f on H will be called *tame* if there is a finite-dimensional projection P on H (e.g. a function $P : H \to H$ satisfying $P^2 = P$ and with a finite-dimensional image) such that $f(Ph) = f(h)$ (this means that $f(h)$ depends only on a finite number of coordinates of h). If $(e_1, e_2 \dots e_n)$ is an orthonormal basis for $P(H)$ and $\Phi_P(u_1, u_2 \dots u_n) = f(u_1 e_1 + \cdots + u_n e_n)$, then

$$\int_H f(h)\nu(\mathrm{d}h) = (2\pi)^{-n/2} \int_{\mathcal{R}^n} \Phi_P(u)\, \mathrm{e}^{-\|u\|^2/2}\, \mathrm{d}u, \tag{5}$$

where $\|u\|$ is the Euclidean norm on $\mathcal{R}^n$ and ν is the measure defined in equation (4).

4.3. The Hilbert space of admissible functions

The Hilbert space suitable for regularization is the space $L_2^{(2)}$ of all functions f with a square integrable distributional second derivative. This is because the cost-functional $M()$ of equation (2) contains the integral of the square of the second derivative. This is an example of a *Sobolev space* [1]. As a result of the *Sobolev embedding theorem* it follows that these functions are continuous. The inner product on this space is defined as

$$(f, g)_1 = \int_0^1 f(v)g(v)\,\mathrm{d}v + \int_0^1 f'(v)g'(v)\,\mathrm{d}v + \int_0^1 f''(v)g''(v)\,\mathrm{d}v. \tag{6}$$

However, the inner product used here—which is a translation of the cost functional $M()$—is

$$(f, g)_2 = \sum_{i=1}^{l} f(x_i)g(x_i) + \lambda \int_0^1 f''(v)g''(v)\,\mathrm{d}v. \tag{7}$$

Lemma 2. *For $l \geqslant 2$ equation (7) defines an inner product on $L_2^{(2)}$.*

Proof. The only condition not trivially satisfied is that $(f, f) = 0$ iff $f = 0$. However, if $(f, f) = 0$, then $f'' = 0$ almost everywhere so f is a linear function. But it must be zero at the l sample points; since $l \geqslant 2$, f is identically zero. □

In order to use the theory of Sobolev spaces, it is necessary to show that the two inner products in equations (6) and (7)—the standard one and the 'hybrid' one used here—define the same topology. For this the next theorem suffices.

Theorem 1. *The inner products of equations (6) and (7) define equivalent norms.*

Proof. The proof will make use of the following lemmas. □

Lemma 3. *For every $0 \leqslant x \leqslant 1$, the linear functional $L_x : L_2^{(2)} \to \mathcal{R}$ defined by $L_x(f) = f(x)$ is bounded, where the inner product used is $()_2$.*

Proof. It is necessary to prove that there exists a positive constant C such that for every $f \in L_2^{(2)}$ $|f(x)| \leqslant C\|f\|$, or $|f(x)| \leqslant C\sqrt{\sum_{i=1}^{l} f^2(x_i) + \lambda\int_0^1 [f''(v)]^2\,\mathrm{d}v}$. It is enough to prove this for $l = 2$, because adding sampling points only increases $\|f\|$. Clearly, $|f(x)| = |f(x_2) + \int_{x_2}^{x} f'(u)\mathrm{d}u|$. By the mean-value theorem, there exists a point $x_1 \leqslant \eta \leqslant x_2$ such that $\frac{f(x_2)-f(x_1)}{x_2-x_1} = f'(\eta)$. Thus, for every u,

$$|f'(u)| = |f'(\eta) + \int_{\eta}^{u} f''(v)\,\mathrm{d}v| \leqslant \frac{|f(x_2) - f(x_1)|}{x_2 - x_1} + \int_0^1 |f''(v)|\,\mathrm{d}v.$$

Since by the Cauchy–Schwarz inequality for integrals $|\int_0^1 g(v)\,\mathrm{d}v| \leqslant \sqrt{\int_0^1 g^2(v)\,\mathrm{d}v}$ for every g and since $|a-b| \leqslant |a|+|b|$ for all a, b the right-hand side of the last inequality is bounded by

$$\frac{|f(x_2)|+|f(x_1)|}{x_2-x_1} + \sqrt{\int_0^1 [f''(v)]^2\,\mathrm{d}v}.$$

Now for every x

$$\begin{aligned}
|f(x)| &= |f(x_2) + \int_{x_2}^{x} f'(u)\,\mathrm{d}u| \leqslant |f(x_2)| + \int_0^1 |f'(u)|\,\mathrm{d}u \\
&\leqslant |f(x_2)| + \frac{|f(x_2)|+|f(x_1)|}{x_2-x_1} + \sqrt{\int_0^1 [f''(v)]^2\,\mathrm{d}v} \\
&\leqslant 2\frac{|f(x_2)|+|f(x_1)|}{x_2-x_1} + \sqrt{\int_0^1 [f''(v)]^2\,\mathrm{d}v} \\
&\leqslant \frac{4}{x_2-x_1}\sqrt{f^2(x_1)+f^2(x_2)+\int_0^1 [f''(v)]^2\,\mathrm{d}v} \\
&\leqslant \frac{4}{(x_2-x_1)\min\{1,\sqrt{\lambda}\}}\sqrt{f^2(x_1)+f^2(x_2)+\lambda\int_0^1 [f''(v)]^2\,\mathrm{d}v}
\end{aligned}$$

and thus it is proved that $\|L_x\| \leqslant \frac{4}{(x_2-x_1)\min\{1,\sqrt{\lambda}\}}$. □

It follows immediately from the proof that there exist constants C_1 and C_2 such that $C_1(f,f)_2 \geqslant \int_0^1 f^2(v)\,\mathrm{d}v$ and $C_2(f,f)_2 \geqslant \int_0^1 (f'(v))^2\,\mathrm{d}v$, and so there exists a constant C such that $C(f,f)_1 \leqslant (f,f)_2$. Now it remains to prove the opposite that $(f,f)_2$ is bounded by $(f,f)_1$.

Lemma 4. *There exists a constant K such that for all* $0 \leqslant x \leqslant 1$,
$f^2(x) \leqslant K\big(\int_0^1 f^2(v)\,\mathrm{d}v + \int_0^1 (f'(v))^2\,\mathrm{d}v\big)$.

Proof. By the mean-value theorem for integrals, there exists an $0 \leqslant x_0 \leqslant 1$ such that $|f(x_0)| \leqslant \int_0^1 |f(v)|\,\mathrm{d}v$. Now, for all $0 \leqslant x \leqslant 1$,

$$\begin{aligned}
|f(x)| &= |f(x_0) + \int_{x_0}^{x} f'(v)\,\mathrm{d}v| \leqslant |f(x_0)| + \int_0^1 |f'(v)|\,\mathrm{d}v \\
&\leqslant 2\sqrt{\left[\int_0^1 |f(v)|\,\mathrm{d}v\right]^2 + \left[\int_0^1 |f'(v)|\,\mathrm{d}v\right]^2} \leqslant 2\sqrt{\int_0^1 f^2(v)\,\mathrm{d}v + \int_0^1 (f'(v))^2\,\mathrm{d}v}.
\end{aligned}$$

□

To conclude, it is only necessary to observe that $(f,f)_2$ is composed of a finite number of the squared values of the function—each bounded by $(f,f)_1$—and the product by λ of the integral of the squared second derivative, which is certainly bounded by $(f,f)_1$. This completes the proof of theorem 1.

By the Riesz representation theorem and lemma 3, there exists for every $0 \leqslant x \leqslant 1$ a function $h_x \in L_2^{(2)}$ such that $(f, h_x) = f(x)$ for every $f \in L_2^{(2)}$ (this function is often referred to as the *reproducing kernel*). Using this fact the cost functional $M()$ can be expressed as $M(f) = \sum_{i=1}^{l} y_i^2 + \|f\|^2 - 2\big(f, \sum_{i=1}^{l} y_i h_{x_i}\big)$. Defining $f_0 = \sum_{i=1}^{l} y_i h_{x_i}$ yields

$$M(f) = \sum_{i=1}^{l} y_i^2 + \|f\|^2 - 2(f, f_0) = \sum_{i=1}^{l} y_i^2 + \|f - f_0\|^2 - \|f_0\|^2. \tag{8}$$

Lemma 5. *$M()$ attains its minimal value at f_0.*

Proof. Obvious from equation (8). □

The expectation E_x and variance V_x will now be computed in the space $L_2^{(2)}$. As the norm differs from the cost functional only by a translation of f_0 and a constant:

$$E_x = \int_{L_2^{(2)}} (f(x) + f_0(x))\nu(\mathrm{d}f) = \int_{L_2^{(2)}} L_x(f + f_0)\nu(\mathrm{d}f) = \int_{L_2^{(2)}} L_x(f)\nu(\mathrm{d}f) + \int_{L_2^{(2)}} L_x(f_0)\nu(\mathrm{d}f) = L_x(f_0) = f_0(x)$$

since the function $L_x(f)$ is odd its integral vanishes. So the expectation is the same as the cost minimizing function. Next, the variance is computed:

$$V_x = \int_{L_2^{(2)}} L_x^2(f + f_0)\nu(\mathrm{d}f) - L_x^2(f_0) = \int_{L_2^{(2)}} L_x^2(f)\nu(\mathrm{d}f) + 2L_x(f_0)\int_{L_2^{(2)}} L_x(f)\nu(\mathrm{d}f) + L_x^2(f_0) - L_x^2(f_0) = \int_{L_2^{(2)}} L_x^2(f)\nu(\mathrm{d}f)$$

again using the fact that $\int_{L_2^{(2)}} L_x(f)\nu(\mathrm{d}f) = 0$. Since $L_x(f) = (h_x, f)$,

$$V_x = \int_{L_2^{(2)}} (f, h_x)^2 \nu(\mathrm{d}f).$$

In order to be able to use the notations of equation (5) without confusion, let us write this integral as $V_x = \int_{L_2^{(2)}} (h, h_x)^2 \nu(\mathrm{d}h)$. Now, using the notations of equation (5), define

$$P : H \to H,\ P(h) = \frac{(h, h_x)h_x}{\|h_x\|^2}.$$

This is the finite-dimensional projection required, and its range is span$\{h_x\}$. An orthonormal basis is given by $\left\{e = \frac{h_x}{\|h_x\|}\right\}$, and as $f(h) = (h, h_x)^2$, then

$$\Phi_P(u) = f(ue) = \left(\frac{uh_x}{\|h_x\|}, h_x\right)^2 = u^2 h_x^2$$

and so

$$\int_H (h, h_x)^2 \nu(\mathrm{d}h) = \frac{1}{\sqrt{2\pi}} \int_{\mathcal{R}} \|h_x\|^2 u^2 \exp(-\|u\|^2/2)\,\mathrm{d}u = \frac{\|h_x\|^2}{2}.$$

Note that the integral could be reduced to an integral over $\mathcal{R}$ because the function (f, h_x) depends only on the projection of f on the subspace span$\{h_x\}$.

Now, $(f, h_x) = f(x)$ for every f, and so $\|h_x\|^2 = (h_x, h_x) = h_x(x)$. From this and from the Riesz representation theorem it follows that

Theorem 2. *For all $0 \leqslant x \leqslant 1$,*

$$V_x = \frac{1}{2}\|h_x\|^2 = \frac{1}{2}h_x(x) = \frac{1}{2}\sup_{f \in L_2^{(2)}} \frac{f^2(x)}{\sum_{i=1}^{l} f^2(x_i) + \lambda \int_0^1 [f''(v)]^2\,\mathrm{d}v}.$$

The rightmost expression in theorem 2 embodies an attractive intuitive meaning. As x is farther away from the sample points x_i, it is possible—for a correct choice of $f()$—to increase $f^2(x)$ relative to $f^2(x_i)$, while also maintaining a small value of $\int_0^1 [f''(v)]^2\,\mathrm{d}v$, which measures the oscillation of $f()$. If x is close to the x_i's, one cannot increase $f^2(x)$

relative to $f^2(x_i)$ without substantially increasing the oscillation term, since in that case $f()$ has to 'bend' quickly, resulting in a large second derivative. Therefore, as x moves away from the sampling points, the uncertainty of the interpolant in x increases.

From theorem 2 and lemma 3 lower and upper bounds on the variance can be derived.

Lemma 6. *For all* $0 \leqslant x \leqslant 1$,

$$\frac{1}{2l} \leqslant V_x \leqslant \frac{8}{\max_{1\leqslant i,j\leqslant n}(x_j - x_i)^2 \min\{1, \lambda\}}.$$

Proof. The right inequality is just lemma 3 The left inequality follows by replacing the space $L_2^{(2)}$ with the smaller space of constant functions in theorem 2; evidently, the supremum over that space is $1/2l$. □

Next it is shown that the values $f(x)$ at a point x obey a normal distribution. If E in equation (4) is chosen to be the interval (a, b), it follows that

Lemma 7.

$$\nu\{f : a \leqslant f(x) \leqslant b\} = \frac{1}{\sqrt{2\pi V_x}} \int_a^b \mathrm{e}^{\frac{-(v-E_x)^2}{2V_x}} \,\mathrm{d}v$$

and so $f(x)$ is normally distributed. This allows us to give confidence intervals on the height of the interpolated function.

In much the same way the covariance of $f(x)$ and $f(y)$, $\mathrm{Cov}(x, y)$, can be computed. By the definition of the covariance and using the same type of arguments as for the variance, it is equal to $\int_{L_2^{(2)}}(h, h_x)(h, h_y)\nu(\mathrm{d}h)$. But

$$(h, h_x)(h, h_y) = \frac{(h, h_x + h_y)^2 - (h, h_x)^2 - (h, h_y)^2}{2}$$

and integrating this equality over $L_2^{(2)}$ yields

$$\frac{\|h_x + h_y\|^2 - \|h_x\|^2 - \|h_y\|^2}{2}$$

so the covariance turns out to be $\frac{(h_x,h_y)}{2} = \frac{h_x(y)}{2}$. To summarize

Theorem 3. *Let L_1 and L_2 be any bounded linear functionals on the function space $L_2^{(2)}$. Then if $L_1(x) = (l_1, x)$ and $L_2(x) = (l_2, x)$, the variance of L_1 is $\frac{\|l_1\|^2}{2}$, and the covariance of L_1 and L_2 is $\frac{(l_1,l_2)}{2}$.*

4.4. Computational issues

The most direct way to compute V_x is to find h_x. From theorem 2 it is enough to compute $h_x(x)$. First, a short reminder: a *spline* is a piecewise polynomial function, with some order of differentiability enforced at the 'knots' (the endpoints of the individual segments). A cubic spline is composed of segments each of which is a cubic polynomial such that at each knot; the function and its first and second derivatives at both sides of the knot are equal. This imposes three equations for each knot.

Theorem 4. *h_x is a cubic spline with knots at $\{x_i\}_{i=1}^l$ and x.*

Proof. Suppose without loss of generality that $0 < x_1 < x_2 \cdots < x_{j-1} < x < x_j < \cdots x_l < 1$, so the interval $[0, 1]$ is partitioned into $l + 2$ segments. From the definition of h_x

$\sum_{i=1}^{l} f(x_i)h_x(x_i) + \lambda \int_0^1 f''(v)h_t''(v)\,\mathrm{d}v = f(x)$ for every $f \in L_2^{(2)}$. Integrating by parts

$$\int_z^y f''(v)h_x''(v)\,\mathrm{d}v = f'(y)h_x''(y) - f'(z)h_x''(z) - f(y)h_x'''(y) + f(z)h_x'''(z) + \int_z^y f(v)h_x^{(4)}(v)\,\mathrm{d}v$$

and if h_x is a cubic spline the last summand vanishes. If the last identity is used in each of the $l+2$ segments, it follows that in order for h_x to satisfy the requirements it is necessary that the following conditions hold.

Spline condition: h_x is a cubic spline. This imposes $3(l+1)$ conditions on h_x.

Boundary ('natural') conditions: $h_x''(0) = h_x'''(0) = h_x''(1) = h_x'''(1) = 0$ (4 conditions).

'Delta' conditions: $h_x^{+'''}(x) - h_x^{-'''}(x) = 1/\lambda$, and for all $1 \leqslant i \leqslant l$ $h_x(x_i) - \lambda h_x^{-'''}(x_i) + \lambda h_x^{+'''}(x_i) = 0$, where the upper + denotes derivative from right and the $-$ from left ($l+1$ conditions).

All in all this gives $4(l+2)$ conditions, which uniquely determine a cubic spline. By the Riesz representation theorem, this spline is h_x. If only V_x is sought, the spline's coefficients in one segment only need to be found. Note that since the cost minimizing function f_0 is a linear combination of the $\{h_{x_i}\}$, it can also be found by calculating the splines. □

The following theorem can save a lot of time when computing V_x for many values of x in the interval.

Theorem 5. *Between any two sample points, the variance V_x is a sixth-degree polynomial in x.*

The proof requires the following lemma.

Lemma 8. *Let $f_{i,j}(x)$ be functions of x for all $1 \leqslant i, j \leqslant n$, and let*

$$F(x) = \begin{vmatrix} f_{1,1}(x) \dots f_{1,n}(x) \\ . \\ . \\ . \\ f_{n,1}(x) \dots f_{n,n}(x) \end{vmatrix};$$

then

$$F'(x) = \begin{vmatrix} f'_{1,1}(x) \dots f'_{1,n}(x) \\ . \\ . \\ . \\ f_{n,1}(x) \dots f_{n,n}(x) \end{vmatrix} + \dots + \begin{vmatrix} f_{1,1}(x) \dots f_{1,n}(x) \\ . \\ . \\ . \\ f'_{n,1}(x) \dots f'_{n,n}(x) \end{vmatrix}.$$

The proof is elementary.

For the sake of simplicity, a proof will be provided for a special case, as the proof generalizes easily. Suppose there are two sample points, $x_1 \leqslant x_2$, and let $x_1 \leqslant x \leqslant x_2$. Now, h_x is a cubic spline in each of the four following intervals:
in $[0, x_1]$ it is $a_1t^3 + b_1t^2 + c_1t + d_1$.
in $[x_1, x]$ it is $a_2t^3 + b_2t^2 + c_2t + d_2$.
in $[x, x_2]$ it is $a_3t^3 + b_3t^2 + c_3t + d_3$.
in $[x_2, 1]$ it is $a_4t^3 + b_4t^2 + c_4t + d_4$.

Writing the conditions of theorem 4 in terms of the coefficients yields the following linear system:

$$\begin{pmatrix}
x_1^3 & x_1^2 & x_1 & 1 & -x_1^3 & -x_1^2 & -x_1 & -1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
3x_1^2 & 2x_1 & 1 & 0 & -3x_1^2 & -2x_1 & -1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
6x_1 & 2 & 0 & 0 & -6x_1 & -2 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & x^3 & x^2 & x & 1 & -x^3 & -x^2 & -x & -1 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 3x^2 & 2x & 1 & 0 & -3x^2 & -2x & -1 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 6x & 2 & 0 & 0 & -6x & -2 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & x_2^3 & x_2^2 & x_2 & 1 & -x_2^3 & -x_2^2 & -x_2 & -1 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 3x_2^2 & 2x_2 & 1 & 0 & -3x_2^2 & -2x_2 & -1 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 6x_2 & 2 & 0 & 0 & -6x_2 & -2 & 0 & 0 \\
x_1^3-6\lambda & x_1^2 & x_1 & 1 & 6\lambda & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & x_2^3-6\lambda & x_2^2 & x_2 & 1 & 6\lambda & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & -6 & 0 & 0 & 0 & 6 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0
\end{pmatrix}$$

$$\times \begin{pmatrix} a_1 \\ b_1 \\ c_1 \\ d_1 \\ a_2 \\ b_2 \\ c_2 \\ d_2 \\ a_3 \\ b_3 \\ c_3 \\ d_3 \\ a_4 \\ b_4 \\ c_4 \\ d_4 \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 1/\lambda \\ 0 \\ 0 \\ 0 \\ 0 \end{pmatrix}.$$

This system can be solved using Cramer's rule to give the coefficients of the spline h_x. Let us call the coefficient matrix of the system A, and calculate $|A|_x$. Using lemma 8, it is the sum of 16 determinants, each of them identical to A except for one line which is the derivative of the corresponding line in A by x. Of the 16 summands, only three do not have a row equal to zero; these are the ones corresponding to the fourth, fifth and sixth line. In the first summand, however, the derivative of the fourth line is equal to the fifth line; so it vanishes. The second summand vanishes because the derivative of the fifth line is equal to the sixth line, and the third summand vanishes because the derivative of the sixth line is equal to minus the twelfth line. By Cramer's rule, the coefficients are the minors—which are polynomials in x—divided by $|A|$. Since $|A|_x = 0$, $|A|$ is not a function of x, and the coefficients of the h_x are polynomials in x. The theorem follows immediately.

4.5. Reduction to a small subspace

It will now be shown that V_x can be approximated on a finite-dimensional subspace of $L_2^{(2)}$. This allows faster computation in some cases. Since the norm of an operator is to be evaluated, it is desirable to find an increasing sequence of subspaces S_n such that the norm of the operator on the 'supplement' between S_{n-1} and S_n is small. One such sequence is $S_n = \text{span}\{1, x, \sin(\pi x), \sin(2\pi x) \ldots \sin(n\pi x)\}$. $S = \bigcup S_n$ is *dense* in $L_2^{(2)}$, meaning that every element of the space $L_2^{(2)}$ can arbitrarily be approximated in norm by an element of S. This follows from Fourier theory—the odd trigonometric functions are dense in the integral part of the norm, because the integration is performed on a half-period of the trigonometric functions, and by integrating twice completeness in the norm follows. This integration adds the functions 1 and x. The norm of an operator when restricted to a dense subset of a Hilbert space is equal to the norm on the entire space; thus, it suffices to compute the norm on S.

For any $0 \leqslant x \leqslant 1$ denote the linear functional that assigns to every f the value $f(x)$ by L—as the following discussion will not depend on the specific x. Define $M_n = \sup_{f \in S_n} \frac{L^2(f)}{\|f\|^2}$. Let the 'difference' between S_{n-1} and S_n, $\text{span}\{\sin(n\pi x)\}$, be denoted D_n. So

$$M_n = \sup_{f \in S_{n-1}, g \in D_n} \frac{L^2(f+g)}{\|f+g\|^2}.$$

Lemma 9. *For $f \in S_{n-1}, g \in D_n$,*

$$\|f+g\|^2 \geqslant (\|f\|^2 + \|g\|^2)\left(1 - \frac{\sqrt{l}}{\sqrt{\lambda} n^2}\right).$$

Proof. Note that $\int_0^1 f''(v) g''(v)\,\mathrm{d}v = 0$. So

$$\|f+g\|^2 = \|f\|^2 + \|g\|^2 + 2(f,g) \geqslant \|f\|^2 + \|g\|^2 - 2\sum_{i=1}^{l} |f(x_i) g(x_i)|$$

$$\geqslant \|f\|^2 + \|g\|^2 - 2\sqrt{\sum_{i=1}^{l} f^2(x_i)}\sqrt{\sum_{i=1}^{l} g^2(x_i)} \tag{9}$$

using the Cauchy–Schwarz inequality. But clearly $\|f\| \geqslant \sqrt{\sum_{i=1}^{l} f^2(x_i)}$ and $\frac{\sqrt{l}}{\sqrt{\lambda} n^2}\|g\| \geqslant \sqrt{\sum_{i=1}^{l} g^2(x_i)}$ because $\int_0^1 [\sin''(n\pi v)]^2\,\mathrm{d}v \geqslant n^4$ and $g(v)$ is a scalar multiple of $\sin(n\pi v)$. So, the last expression in equation (9) is greater than $\|f\|^2 + \|g\|^2 - \frac{2\sqrt{l}}{\sqrt{\lambda} n^2}\|f\|\|g\|$. But $\|f\|^2 + \|g\|^2 \geqslant 2\|f\|\|g\|$, thus completing the proof. □

Lemma 10. *For every* $g \in D_n$, $\frac{L^2(g)}{\|g\|^2} \leqslant \frac{1}{\lambda n^4}$.

Proof. Obvious. □

Lemma 11. $M_{n-1} \geqslant \frac{1}{2l}$.

Proof. This follows from lemma 6. □

Now, for every $f \in S_{n-1}$ and $g \in D_n$, if $\frac{\sqrt{l}}{\sqrt{\lambda}n^2} < 1$

$$\begin{aligned}
\frac{L^2(f+g)}{\|f+g\|^2} &\leqslant \left(1+\frac{2\sqrt{l}}{\sqrt{\lambda}n^2}\right)\frac{L^2(f)+2L(f)L(g)+L^2(g)}{\|f\|^2+\|g\|^2} \\
&\leqslant \left(1+\frac{2\sqrt{l}}{\sqrt{\lambda}n^2}\right)\frac{M_{n-1}\|f\|^2+2\frac{\sqrt{M_{n-1}}}{\sqrt{\lambda}n^2}\|f\|\|g\|+\frac{\|g\|^2}{\lambda n^4}}{\|f\|^2+\|g\|^2} \\
&\leqslant M_{n-1}\left(1+\frac{2\sqrt{l}}{\sqrt{\lambda}n^2}\right)\left(1+\frac{1}{\sqrt{\lambda}n^2\sqrt{M_{n-1}}}+\frac{1}{\lambda n^4 M_{n-1}}\right) \\
&\leqslant M_{n-1}\left(1+\frac{2\sqrt{l}}{\sqrt{\lambda}n^2}\right)\left(1+\frac{\sqrt{l}}{\sqrt{\lambda}n^2}+\frac{l}{\lambda n^4}\right) = M_{n-1}\left[1+O\left(\frac{\sqrt{l}}{\sqrt{\lambda}n^2}\right)\right]
\end{aligned}$$

(the last inequality uses lemma 11). Since this is true for all $f \in S_{n-1}, g \in D_n$, it follows that $M_{n-1}\left[1+O\left(\frac{\sqrt{l}}{\sqrt{\lambda}n^2}\right)\right] \geqslant M_n$. So the operator's norm in $L_2^{(2)}$ is bounded by the product of the norm in S_{n-1} by $\prod_{k=n}^{\infty}\left[1+O\left(\frac{\sqrt{l}}{\sqrt{\lambda}k^2}\right)\right]$, but this is equal to $\mathrm{e}^{O\left(\frac{\sqrt{l}}{\sqrt{\lambda}n}\right)}$. This proves that to approximate the norm up to a factor of $1+c$, it can be evaluated in a subspace of dimension $\frac{1}{c}\sqrt{\frac{l}{\lambda}}$.

Such approximation schemes can be especially useful for two-dimensional functions, in which the roughness term is $\int_0^1\int_0^1\left(f_{uu}^2+2f_{uv}^2+f_{vv}^2\right)\mathrm{d}u\,\mathrm{d}v$. It turns out that in this case the solution for the reproducing kernel cannot be represented as a spline, so there is no simple solution for computing the variance.

To evaluate V_x over a finite-dimensional subspace, equation (3) (lemma 1) can be used. Note that while x varies, Δ^{-1} does not depend on x and needs to be computed only once. Using a spectral decomposition of Δ^{-1} allows us to replace the expression for V_x by the approximation

$$\sum_{k=1}^{r}\beta_k((f_1(x), f_2(x)\ldots f_n(x))\cdot q_k)^2,$$

where β_k are the leading eigenvalues in the spectral decomposition and q_k are the corresponding eigenvectors. For example, for 30 sampling points evenly spread in the unit interval, and 30 basis functions, the sum of squares of the first nine eigenvalues equals 0.9951 of the sum of squares of all 30 eigenvalues, and the sum of squares of the first two eigenvalues equals 0.9918 of the overall sum, which means that computational complexity can drastically be reduced while maintaining very high accuracy.

5. Uncertainty of other functionals

Next, the uncertainties of two important functionals—the integral and the derivative—are computed. It is demonstrated that, as can be expected, the uncertainty of the derivative is

higher than that of the function; this means, for example, that typically it is harder to estimate the velocity of a moving object than its location. The somewhat surprising result is that the uncertainties are sometimes uncorrelated; the variance of the function can relatively be high at points where the variance of the derivative is low, and vice versa.

First, the variance of the integral is calculated. This is achieved by relating the kernels used to represent the function and the integral via the Riesz representation theorem. It is assumed here that there are two sample points, at 0 and 1.

Lemma 12. *Let $H_x(\xi)$ be the function satisfying $(f, H_x) = \int_0^x f(\xi)\,\mathrm{d}\xi$. Then $H_x(\xi) = \int_0^x h_\eta(\xi)\,\mathrm{d}\xi$, and so $h_x(\xi) = \frac{\partial H_x(\xi)}{\partial x}$, where $h_x(\xi)$ is the function satisfying $(f, h_x) = f(x)$ for all f.*

Proof. For every η the following holds:

$$f(\eta) = f(0)h_\eta(0) + f(1)h_\eta(1) + \lambda \int_0^1 f''(\xi)h''_\eta(\xi)\,\mathrm{d}\xi$$

integrating this equality over the interval $[0, x]$ with respect to η yields

$$\int_0^x f(\eta)\,\mathrm{d}\eta = f(0)\int_0^x h_\eta(0)\,\mathrm{d}\eta + f(1)\int_0^x h_\eta(1)\,\mathrm{d}\eta + \lambda \int_0^1 f''(\xi)\left[\int_0^x h''_\eta(\xi)\,\mathrm{d}\eta\right]\mathrm{d}\xi$$

and defining $H_x(\eta) = \int_0^x h_\xi(\eta)\,\mathrm{d}\xi$ concludes the proof, since h is symmetric. □

This result can of course be extended to the derivative and its uncertainty. In fact

Lemma 13. *The variance of the derivative at the point x is equal to $\frac{\partial^2 h(x,\xi)}{\partial x \partial \xi}|_{\xi=x}$.*

Proof. By lemma 12, the function representing the derivative is $d_x(\xi) \equiv d(x, \xi) = \frac{\partial h(x,\xi)}{\partial x}$. Now, according to theorem 2 the variance of the derivative at x is equal to $\frac{1}{2}(d_x, d_x)$. But for all f, $(f, d_x) = f'(x)$, and so the lemma follows (remember that d_x is a function of ξ). □

The next lemma shows that the second derivative is very unreliable.

Lemma 14. *The variance of the second derivative is ∞.*

Proof. According to theorem 2, the variance is equal to the norm of the operator $L(f) = f''(x)$ (for any x) or

$$\sup_{f \in L_2^{(2)}} \frac{[f''(x)]^2}{\sum_{i=1}^l f^2(x_i) + \lambda \int_0^1 [f''(v)]^2\,\mathrm{d}v}; \tag{10}$$

it is well known [1] that for all ϵ_1, ϵ_2 and K there exists a non-negative infinitely differentiable function g satisfying the following:

(1) g is non-zero only in the interval $[x - \epsilon_1, x + \epsilon_1]$;
(2) $\int_0^1 g^2(\xi)\,\mathrm{d}\xi < \epsilon_2$;
(3) $g(x) > K$.

Let f be defined as the function obtained by integrating g twice (e.g. $f'' = g$). It is easy to see that for a suitable choice of ϵ_1, ϵ_2 and K the expression whose supremum is sought in equation (10) can obtain arbitrarily large values, thus completing the proof. □

Let us demonstrate these concepts in the simplest case possible, where there are only two sample points, 0 and 1, and $\lambda = 1$. The kernel representing the integral is

$$H_x(\xi) = \begin{cases} 0 \leqslant \xi \leqslant x: & \frac{\xi^4}{4} + \left(\frac{x^2-2x}{12}\right)\xi^3 + \left(\frac{x^4-4x^3+28x^2-24x}{24}\right)\xi - \frac{x^2-2x}{2} \\ x \leqslant \xi \leqslant 1: & \frac{x^2}{12}\xi^3 - \frac{x^2}{4}\xi^2 + \left(\frac{x^4+28x^2-24x}{24}\right)\xi - \frac{x^4+12x^2-24x}{24} \end{cases}$$

and so the kernel representing the evaluation (substitution) at x is

$$h_x(\xi) = \begin{cases} 0 \leqslant \xi \leqslant x: & \left(\frac{x-1}{6}\right)\xi^3 + \left(\frac{x^3-3x^2+14x-6}{6}\right)\xi + 1 - x \\ x \leqslant \xi \leqslant 1: & \left(\frac{x}{6}\right)\xi^3 - \left(\frac{x}{2}\right)\xi^2 + \left(\frac{x^3+14x-6}{6}\right)\xi - \frac{x^3+6x-6}{6} \end{cases}$$

and the kernel representing the differentiation at x

$$d_x(\xi) = \begin{cases} 0 \leqslant \xi \leqslant x: & \frac{\xi^3}{6} + \left(\frac{3x^2-6x+14}{6}\right)\xi - 1 \\ x \leqslant \xi \leqslant 1: & \frac{\xi^3}{6} - \frac{\xi^2}{2} + \left(\frac{3x^2+14}{6}\right)\xi - \frac{x^2+2}{2}. \end{cases}$$

It is interesting to see what happens when $d_x(\xi)$ is differentiated by x:

$$s_x(\xi) = \begin{cases} 0 \leqslant \xi \leqslant x: & (x-1)\xi \\ x \leqslant \xi \leqslant 1: & x(\xi - 1) \end{cases}$$

by theorem 2, s_x should satisfy for all f

$$(f, s_x) = f(0)s_x(0) + f(1)s_x(1) + \int_0^1 s_x''(\xi) f''(\xi)\,\mathrm{d}\xi = f''(x)$$

now, informally

$$s_x''(\xi) = \begin{cases} x \neq \xi: & 0 \\ x = \xi: & \infty \end{cases}$$

so, since $s_x(0) = s_x(1) = 0$, the identity $\int_0^1 f''(\xi)s_x''(\xi)\,\mathrm{d}\xi = f''(x)$ should hold. But that would imply that $s_x''(\xi)$ is a delta function at x; hence, the variance is ∞.

Another illustration to the variance of the second derivative is now provided. Since

$$f''(x) \approx \frac{f'(x+h) - f'(x)}{h}$$

the variance of $f''(x)$ is approximated by

$$\frac{d_{x+h}(x+h) + d_x(x) - 2d_{x+h}(x)}{h^2}$$

(remember that the covariance of the derivatives at x and $x+h$ is $d_{x+h}(x)$). This expression turns out to be

$$\frac{3hx + 2h^2 - 3h + 7}{3h}$$

and so tends to ∞ as h tends to 0.

5.1. Using an orthonormal basis

Every separable Hilbert space possesses an orthonormal basis [50]. Since the computation of the variance and covariance is especially simple when such a basis is given, it is presented here.

Let $\{e_i\}$ be an orthonormal basis for H, and let $L_1, L_2 : H \to \mathcal{R}$ be the linear functionals on H.

Lemma 15. *The variance of L_1 is equal to $\frac{1}{2}\sum_i L_1^2(e_i)$, and the covariance of L_1 and L_2 equals $\frac{1}{2}\sum_i L_1(e_i)L_2(e_i)$.*

Proof. First, one has to find the elements l_1 and l_2 representing L_1 and L_2. For instance, l_1 has to satisfy $(l_1, e_i) = L(e_i)$ for all i. But by the definition of an orthonormal basis, this means that $l_1 = \sum_i L_1(e_i)e_i$. Hence, $\|l_1\|^2 = \sum_i L_1^2(e_i)$. Similarly, the covariance is $\frac{1}{2}\big(\sum_i L_1(e_i)e_i, \sum_i L_2(e_i)e_i\big) = \frac{1}{2}\sum_i L_1(e_i)L_2(e_i)$. □

Unfortunately, it is not always easy to explicitly present an orthonormal basis. Let us look at the following example where the construction of a basis is easy; this is when there are two sample points, at 0 and 1, and so the inner product is defined by

$$(f, g) = f(0)g(0) + f(1)g(1) + \lambda \int_0^1 f''(\xi)g''(\xi)\,\mathrm{d}\xi.$$

Lemma 16. *An orthonormal basis in this case is given by the following sequence of functions:*

$$\frac{1}{\sqrt{2}}, \frac{1-2\xi}{\sqrt{2}}, \left\{\frac{\sqrt{2}\sin(k\pi\xi)}{\sqrt{\lambda}\pi^2 k^2}\right\}_{k=1}^{\infty}.$$

Proof. It is easy to check that the sequence of functions is indeed orthonormal. It remains to show that it spans the space of admissible functions, but this was shown in the discussion opening section 4.5. □

According to lemma 15, it is now possible to represent the uncertainty of the following functionals via this basis:

$$L(f) = f(x) : \frac{1}{2} + \frac{(1-2x)^2}{2} + \frac{2}{\lambda\pi^4}\sum_{k=1}^{\infty}\frac{\sin^2(k\pi x)}{k^4}$$

$$L(f) = f'(x) : 2 + \frac{2}{\lambda\pi^2}\sum_{k=1}^{\infty}\frac{\cos^2(k\pi x)}{k^2}.$$

What will happen if this approach is attempted to compute the variance of the second derivative?

$$L(f) = f''(x) : \frac{2}{\lambda}\sum_{k=1}^{\infty}\sin^2(k\pi x)$$

and the series diverges to ∞:

$$L(f) = \int_0^x f(\xi)\,\mathrm{d}\xi : \frac{x^2}{2} + \frac{(x-x^2)^2}{2} + \frac{\sqrt{2}}{\sqrt{\lambda}\pi^6}\sum_{k=1}^{\infty}\frac{\cos^6(k\pi x)}{k^6}.$$

It is interesting to see how the intuitive notion that the integral has the smallest variance, and the derivative the largest, expresses itself here: obviously, the size of the variance is inversely proportional to the power of k in the denominator of the infinite series. This power is 2 for the derivative, 4 for the function and 6 for the integral.

6. Optimal sampling

An important question in the natural sciences is: where should a function be sampled in order to obtain its most reliable reconstruction? For example, for the task of reconstructing plasma cross-sections, X-ray 'cameras' are installed inside a Tokamak; what are their ideal locations?

In the model presented here, this could be interpreted as sampling the data in points that would make the uncertainty as small as possible. The measure of uncertainty used is the integral of the point variances over the domain in question. Specifically, the question posed was: for some l, what are the l points that will minimize this integral? This problem poses a numerical difficulty because the function relating the points to the variance is rather complicated. Nonetheless it can be solved using methods that do not require computing the function's derivatives, such as Powell's method [35]. The answer depends considerably on λ, the weight given to the roughness term of the cost functional: as λ decreases the optimal sampling points tend to be closer, while for large λ the points tend to be spread out and frequently lie on the region's boundary. This can be explained as follows: as λ increases, rough functions are assigned very low probabilities, and thus have less effect on the variance. Smooth functions, however, can be sampled at points spatially spread apart without losing too much information.

The following tables give, for some values of l and λ, the optimal points at which to sample a function, given some estimate of its oscillation (this estimate is reflected in the value of λ; the higher it is, the smoother a typical function will be). As λ tends to zero, the points become closer, converging to a fixed location.

	$\lambda = 1$	$\lambda = 0.1$	$\lambda = 0.001$
$l = 2$	0, 1	0.02, 0.98	0.15, 0.85
$l = 3$	0, 0.5, 1	0, 0.5, 1	0.09, 0.5, 0.91
$l = 4$	0, 0, 1, 1	0, 0.18, 0.82, 1	0.05, 0.35, 0.65, 0.95

	$\lambda = 0.0001$	$\lambda = 0.000\,001$
$l = 2$	0.2, 0.8	0.2, 0.8
$l = 3$	0.15, 0.5, 0.85	0.18, 0.5, 0.82
$l = 4$	0.08, 0.35, 0.65, 0.92	0.1, 0.35, 0.65, 0.9

Optimal sampling points for $l = 2, 3, 4$ and a range of λ values. Note that for $l = 4$, $\lambda = 1$ the points are 0,0,1,1; this means that the function has to be sampled twice at 0 and twice at 1. This makes sense, since a noisy measurement process was assumed, with noise between different measurements uncorrelated.

7. Examples

Some example of the uncertainty of the function and derivative are depicted, as well as examples illuminating the concepts of the finite-dimensional approximation and optimal sampling (figures 2–6).

7.1. The effect of the sampling points locations

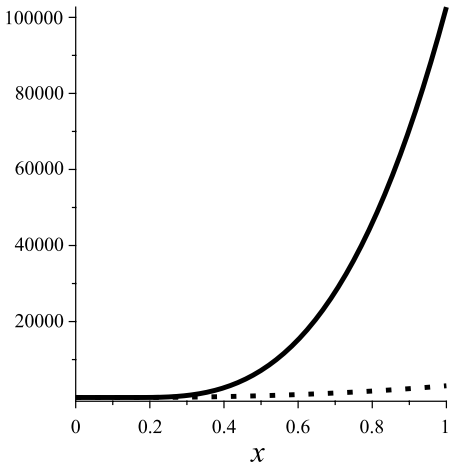

Figure 2. Sampling points are $\{0, 0.1, 0.2\}$, $\lambda = 0.0001$ (continuous line) and $\lambda = 0.01$ (dotted line). Here and hereafter the horizontal axis stands for location, the vertical for the uncertainty (variance) of the function or the derivative. For both values of λ, the uncertainty increases as one moves away from the sampling points, but the rate of increase is much faster for the smaller λ, as it implies a smaller penalty on the oscillation (roughness) of the possible interpolants. To maintain uniformity with figure 3, the variance of the function is scaled by a factor of 50 in all figures.

7.2. Uncertainty of the function versus the derivative

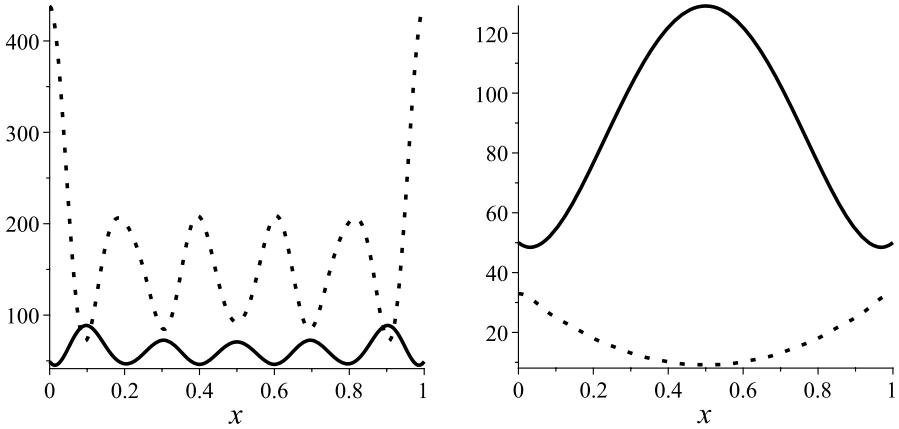

Figure 3. In both figures, the continuous line represents the variance of the function's value at the corresponding spatial location, and the dotted line the variance of the derivative's value. The variance of the function is scaled by a factor of 50 for display purposes. Left: sample points are $\{0, 0.2, 0.4, 0.6, 0.8, 1\}$, and $\lambda = 0.0001$. Note that the local maxima of the derivative's uncertainty coincide with the local minima of the function's uncertainty. Right: sample points are $\{0, 1\}$, and $\lambda = 0.01$. The function is least reliable (highest uncertainty) at $x = 0.5$, and at that point the derivative is most reliable. Note that generally the derivative is far less reliable than the function (the function's uncertainty is scaled by 50), which reaffirms and quantifies the fact, well known amongst the signal processing community that the NSR (noise to signal ratio) is higher in the derivative than in the function.

7.3. Quality of approximation on a finite-dimensional subspace

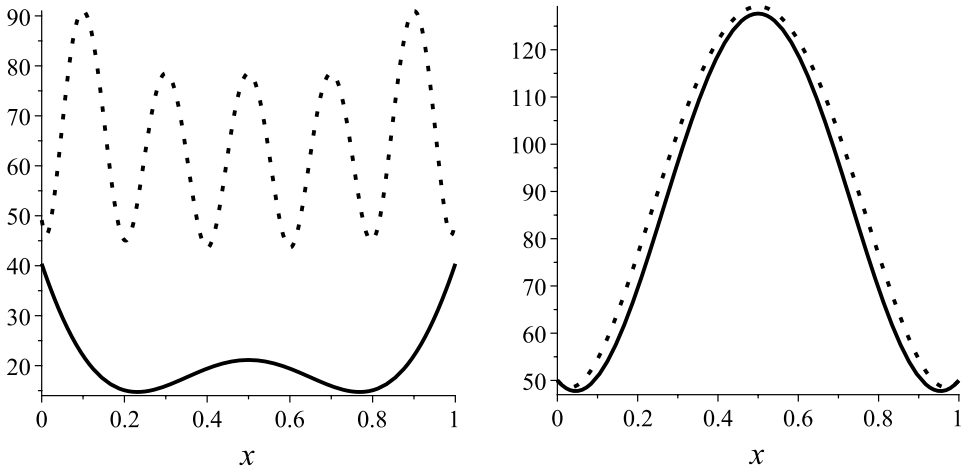

Figure 4. Approximation of the variance on a finite-dimensional subspace. In both plots, the real variance (dotted line) is depicted together with the approximation computed using a three-dimensional subspace, spanned by $\{1, x, \sin(\pi x)\}$ (continuous line). Sample points and λ are as in figures 3(Left) and (Right). From the analysis in section 4.5, the quality of approximation depends on $\sqrt{\frac{l}{\lambda}}$ (l is the number of sample points, λ the weight given to the roughness term), which equals 14.1 for the right plot ($l = 2$, $\lambda = .01$) and 244.9 for the left one ($l = 6$, $\lambda = 0.0001$); hence, while the approximation on the three-dimensional subspace is quite accurate in the right plot, it is rather poor for the upper one.

7.4. Optimal sampling

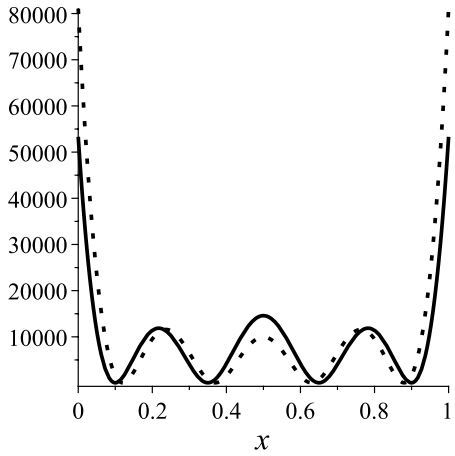

Figure 5. Optimal versus non-optimal sampling, $\lambda = 10^{-6}$, $l = 4$. The optimal sampling points are $\{0.1, 0.35, 0.65, 0.9\}$, and the corresponding variance is the continuous line. The dotted line corresponds to $\{0.12, 0.37, 0.63, 0.88\}$, which are obtained by 'pushing' the optimal points towards the center. While this reduces the variance in the interval's center, the variance sharply increases at the left and right ends, resulting in an increase by a factor of 1.12 in the overall variance (the integral of the pointwise variance). Thus, one may view the location of the optimal sampling points as the one which obtains the optimal balance between the variance at different regions.

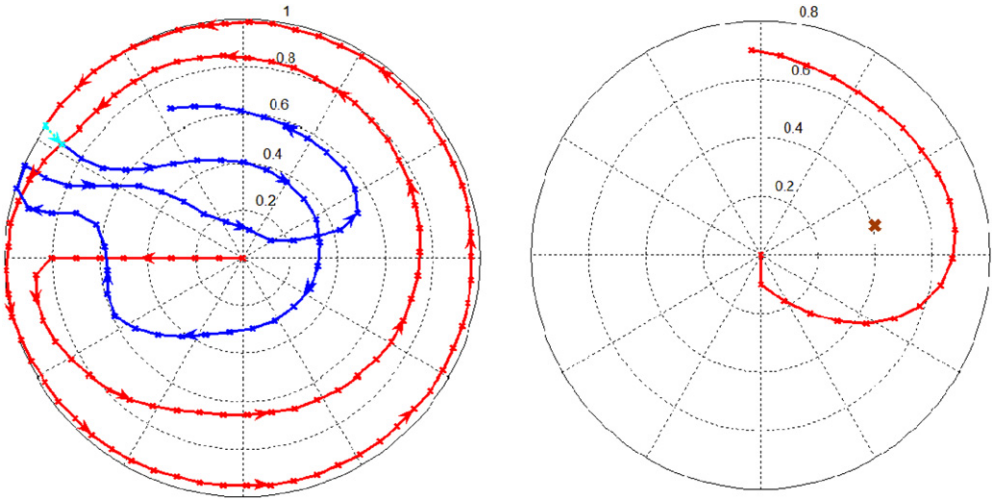

Figure 6. Left: An optimal sampling path, starting at the center of the unit disk. The first step is of course arbitrary, so without loss of generality is was set to the left. Distance between two consecutive points was set to 0.1. The initial part of the curve (red) first advances to the left, in an attempt to 'cover as much area' as possible, and ensure independence between consecutive points. When approaching D's boundary, it starts to curve into a helix-like shape. At the part marked in cyan, the curve sharply bends inwards, covering further regions of D. Right: an optimal sampling path which commenced by moving downwards, and then was informed of a 'preferred' point (marked by 'X').

7.5. Incremental optimal sampling

In the previous section, the question was 'what are the optimal sample points' (i.e. which minimize the overall variance), and one could choose the $(k+1)$st point independently of the kth point. One may assume a different model, in which an optimal sampling *path* is sought, that is, a continuous process which attempts to optimally cover the domain in question. Very loosely speaking, this can be thought of as an agent which attempts to learn its surrounding. In a discrete formulation, this means that there is a fixed distance between consecutive sampling points.

This sampling paradigm is naturally more interesting in two dimensions. The domain tested was the unit disk D, and the roughness term the integral over D of $f_{uu}^2 + 2f_{uv}^2 + f_{vv}^2$. Figure 6(left) depicts the resulting path.

7.5.1. Incremental sampling with a preferred point. Suppose that an optimal sampling path was initiated, and then the optimization process was informed of a 'preferred' point p, by which we mean that it is more important to accurately estimate its value than the value of other points. This can be formalized by modifying the roughness term to $R(p)\left(f_{uu}^2 + 2f_{uv}^2 + f_{vv}^2\right)$, where $R(p)$ is radially increasing around p. This means that smoothness is more strongly enforced as we move away from p; hence, more samples need to be taken in p's vicinity than elsewhere. Figure 6(right) depicts the resulting sampling path.

8. Regularization in physics

In physics, measurement data is typically degraded by noise. It may in addition be corrupted by a convolution or some other operator (in tomography for example, one measures line integrals

of the data). Data measurements may also be sparse, and missing value should be 'filled in'. In some cases [2, 30] the restored object has an infinite number of parameters, which are impossible to extract from a finite number of measurements without additional assumptions.

To solve such problems, regularization is very often applied. We next survey some applications, and then discuss the correct choice of a prior.

In [15] a second-order smoothness term is applied to separate PIXE spectra from the background, by assuming that background spectra is smoother. A second-order smoothness term, identical to the one introduced in section 3, is used. Other applications of regularization to analysis of measurement data include spectroscopy analysis of hard x-ray emission during solar flares [34], and core-shell x-ray spectroscopy [16], in which the measured data is degraded by a convolution; this problem is also studied for gamma-ray spectra in [32]. A Bayesian approach is applied to estimate the maximum diameter of a scatterer in [47]. In [6], neutrino flux data is recovered using regularization with a second-order smoothness term. A method to overcome count losses and noise in photon-number statistics is developed by using regularization techniques in [39]. Further work on the problem of overcoming measurement distortion and noise by applying regularization is discussed in [12, 20].

Regularization has been extensively used in plasma tomography [14, 29] and also in flame tomography [3], in which a two-dimensional function is reconstructed from one-dimensional projections. Regularization of both reduces the effect of measurement noise and also helps to 'fill in' for missing data (typically in these problems, the number of parameters that need to be reconstructed is much higher than the number of measurements, the latter being restricted by e.g. the number of x-ray 'cameras' one can place inside a Tokamak).

More recent applications are to the analysis of cosmological data. A detailed study of the Bayesian approach to the problem of determining cosmological large-scale structure is presented in [24], which also studies different priors (see also section 8.1). Applications of regularization to detecting gravitational waves are presented in [19, 31, 36, 38].

Some other applications of Tikhonov regularization include determining a potential from its normal modes, which leads to an unstable (ill-posed) integral equation [40]. Restoration of potentials from measuring quantum systems is also undertaken in [30]. In [2] regularization is applied to estimating the shape of superconducting cavities in accelerators, which differs from the designed shape due to machine tolerances during the manufacturing process.

A more theoretical application of regularization is presented in [9], which defines a volume operator for loop quantum gravity, and observes a close relation between the inverse of an operator obtained by Tikhonov regularization and the Moore–Penrose pseudoinverse. A somewhat related idea is used in [5], which applies Tikhonov regularization to obtain a regularized Schrödinger equation.

8.1. Choosing the prior for physical problems

Clearly, choosing a prior in order to apply the Bayesian paradigm to a specific physical problem is of crucial importance. The prior reflects our belief as to what properties the correct solution should have, and it affects both the solution and its certainty. And obviously, different problems require different priors.

Bayesian analysis has been the subject of some criticism in that a choice of a prior is 'subjective' and may lead to bias. In [38], the very interesting observation is made that non-Bayesian approaches also introduce implicit priors into the estimation process, and that these priors often translate to *unphysical* assumptions.

Typically, a prior such as the second-order smoothness one we studied here is applied [7, 13, 15, 41]. The optimal solution is chosen to be the one consistent with the data and also

smooth (of the low curvature). Sometimes, the first derivative is used [3]. A simpler prior assumes that the restored signal must be small, and it proceeds by penalizing solutions with a large norm [2, 34]. A detailed comparison of these different types of priors is presented in [41]. In [24], a very detailed study of different priors is undertaken for the problem of estimating the cosmological large-scale structure.

When the restored function has to be positive, the prior can be chosen so as to exclude negative values [10, 32].

That the choice of prior heavily depends on the type of physical phenomenon at hand is demonstrated in the following example. The *maximum-entropy* prior has been often used. In [13, 29] it is concluded that a smoothness-based prior is superior to the maximum-entropy one for tomography problems, since for the maximum-entropy prior the underlying assumption is that neighboring pixels are uncorrelated; but [10] suggests that the maximum-entropy prior is better than the smoothness-based one for determining pair distance distributions by pulsed electron spin resonance, as it better handles the non-negativity constraint.

A delicate choice of prior is used in [31] to prevent the 'two detector paradox'. If all possible waveforms are allowed, even infinitesimally misaligned detectors may yield very different results—clearly an unphysical phenomenon. In [36], the origin of the two detector paradox is traced back to the ill-posedness of the inverse problem for gravitational-wave bursts, and it is solved by regularizing the network response matrix (reducing its condition number).

Other tunings of the prior, necessitated by physical considerations, are possible. For example, in [13], the regularization parameter (λ in equation (2)) is spatially varied in order to avoid over-smoothing at the edges of the temperature profile, where it is known to be discontinuous. In [30] the goal is to reconstruct quantum potentials from measurements; the priors studied are smoothness-based, Gaussian and a symmetry-based prior; assume that the sought potential V approximately commutes with a unitary symmetry operation S. Then, $< V|(I - S)^{\dagger}(I - S)|V >$ can be chosen as a 'penalty term' instead (or in addition) to the smoothness constraint.

8.2. Adapting the computation of the uncertainty to other priors

In this review the problem of estimating the uncertainty was solved for the second-order smoothness term. For other priors, the solution proceeds by computing the integral for V_x, as in section 4. For a Gaussian prior (e.g. resulting from a zero or first-order smoothness term), the integral can be computed via the theory of Gaussian integrals over a Hilbert space, as discussed here. For the positivity constraint (e.g. [10, 32]), one may replace the 'penalty function' $\lambda \int [f''(v)]^2 \, dv$ (and the prior it induces) by $\lambda \int \theta[f(v)] \, dv$, where θ is the Heaviside step function. The resulting integral can be made Gaussian by using the θ's Fourier transform.

9. Summary

The Bayesian view, in which every candidate function is assigned a probability, allows us to extend the question 'which interpolant should be chosen?' to 'how trustworthy are the interpolated values?'. This is done here by defining the uncertainty at a point as the variance of the random variable defined by the evaluation at that point, with the probability of a function defined by a measure of its roughness and its compatibility with the measured data. The theory of Gaussian measures in reproducing kernel Hilbert spaces is applied to compute the variance.

Further results concern the practical issue in computing the uncertainty, as well as computing the uncertainty of the integral and derivative. A notion of optimal sampling is defined.

Lastly, some examples of application of regularization to physics are provided, and the extension of the notions of this review to other priors discussed.

References

[1] Adams R A 1975 *Sobolev Spaces* (New York: Academic)
[2] Akcelik V, Ko K, Lee L Q, Li Z, Ng C K and Xiao L 2008 Shape determination for deformed electromagnetic cavities *J. Comput. Phys.* **227** 1722–38
[3] Akesson E O and Daun K J 2008 Parameter selection methods for axisymmetric flame tomography through Tikhonov regularization *Appl. Opt.* **47** 407–16
[4] Akima H 1974 Bivariate interpolation and smooth surface fitting based on local procedures *Commun. ACM* **17** 26–31
[5] Baer R 2008 On the mapping of time-dependent densities onto potentials in quantum mechanics *J. Chem. Phys.* **128** 044103
[6] Barger V, Huber P and Marfatia D 2006 Ultra high energy neutrino–nucleon cross section from cosmic ray experiments and neutrino telescopes *Phys. Lett.* B **642** 333–41
[7] Barnabe M and Koopmans L V E 2007 A unifying framework for self-consistent gravitational lensing and stellar dynamics analyses of early-type galaxies *Astrophys. J.* **666** 726–46
[8] Bertero M, Poggio T A and Torre V 1988 Ill-posed problems in early vision *Proc. IEEE* **8** 869–89
[9] Bianchi E 2009 The length operator in loop quantum gravity *Nucl. Phys.* B **807** 591–624
[10] Chiang Y W, Borbat P P and Freed J H 2005 Maximum entropy: a complement to Tikhonov regularization for determination of pair distance distributions by pulsed ESR *J. Magn. Reson.* **177** 184–96
[11] Chorley R J 1972 *Spatial Analysis in Geomorphology* ed R J Chorley (London: Methuen & Co)
[12] Davier M, Hocker A and Zhang Z 2006 The physics of hadronic tau decays *Rev. Mod. Phys.* **78** 1043–109
[13] Denisova N, Haverlag M, Ridderhof E J, Nimalasuriya T and van der Mullen J J A M 2008 X-ray absorption tomography of a high-pressure metal-halide lamp with a bent arc due to lorentz-forces *J. Phys. D: Appl. Phys.* **41** 144021
[14] Dose V 2003 Bayesian inference in physics: case studies *Rep. Prog. Phys.* **66** 1421–61
[15] Fischer R, Hanson K, Dose V and von der Linden W 2000 Background estimation in experimental spectra *Phys. Rev.* E **61** 1152–60
[16] Fister T T, Seidler G T, Rehr J J, Kas J J, Elam W T, Cross J O and Nagle K P 2007 Deconvolving instrumental and intrinsic broadening in core-shell x-ray spectroscopies *Phys. Rev.* B **75**
[17] Geman S and Geman D 1984 Stochastic relaxation, Gibbs distribution and the Bayesian restoration of images *IEEE Trans. Pattern Anal. Mach. Intell.* **6** 721–41
[18] Gross L 1960 Integration and non-linear transformations in Hilbert space *Trans. Am. Math. Soc.* **94** 404–40
[19] Hayama K, Desai S, Mohanty S D, Rakhmanov M, Summerscales T and Yoshida S 2008 Searches for gravitational waves associated with pulsar glitches using a coherent network algorithm *Class. Quantum Gravity* **25** 184016–23
[20] Hoecker A and Kartvelishvili V 1996 SVD approach to data unfolding *Nucl. Instrum. Methods Phys. Res.* **372** 469–81
[21] Horn B K P and Schunck B G 1981 Determining optical flow *Artif. Intell.* **17** 185–203
[22] Keren D and Werman M 1993 Probabilistic analysis of regularization *IEEE Trans. Pattern Anal. Mach. Intell.* **15** 982–95
[23] Keren D and Werman M 1999 A full Bayesian approach to curve and surface reconstruction *J. Math. Imaging Vis.* **11** 27–43
[24] Kitaura F S and Ensslin T A 2008 Bayesian reconstruction of the cosmological large-scale structure: methodology, inverse algorithms and numerical optimization *Mon. Not. R. Astron. Soc.* **389** 497–544
[25] Kuelbs J, Larkin F M and Williamson J A 1972 Weak probability distributions on reproducing kernel Hilbert spaces *Rocky Mt. J. Math.* **2** 369–78
[26] Kuo H H 1975 *Gaussian Measures in Banach Spaces* (Berlin: Springer)
[27] Larkin F M 1972 Gaussian measure in Hilbert space and applications in numerical analysis *Rocky Mt. J. Math.* **2** 379–421
[28] Lee D 1986 Approximation of linear operators on a wiener space *Rocky Mt. J. Math.* **16** 641–59
[29] Lee S H, Kim J, Lee J H and Choe W 2010 Modified Phillips–Tikhonov regularization for plasma tomography *Curr. Appl. Phys.* **10** 893–9
[30] Lemm J C, Uhlig J and Weiguny A 2000 Bayesian approach to inverse quantum statistics *Phys. Rev. Lett.* **84** 2068–71

[31] Mohanty S D, Rakhmanov M, Klimenko S and Mitselmakher G 2006 Variability of signal-to-noise ratio and the network analysis of gravitational wave burst signals *Class. Quantum Gravity* **23** 4799–810
[32] Morhac M 2006 Deconvolution methods and their applications in the analysis of gamma-ray spectra *Nucl. Instrum. Methods Phys. Res.* **559** 119–23
[33] Phillips D L 1962 A technique for the numerical solution of certain integral equations of the first kind *J. Assoc. Comput. Mach.* **9** 84–97
[34] Piana M, Massone A M, Hurford G J, Prato M, Emslie A G, Kontar E P and Schwartz R A 2007 Electron flux spectral imaging of solar flares through regularized analysis of hard x-ray source visibilities *Astrophys. J.* **665** 846–55
[35] Press W H, Flannery B P, Teukolsky S A and Vetterling W T 1986 *Numerical Recipes* (Cambridge: Cambridge University Press)
[36] Rakhmanov M 2006 Rank deficiency and Tikhonov regularization in the inverse problem for gravitational-wave bursts *Class. Quantum Gravity* **23** 673–86
[37] Robinson J E, Charlesworth H A K and Ellis M J 1969 Structural analysis using spatial filtering in interior plans of south-central alberta *Am. Assoc. Pet. Geol. Bull.* **53** 2341–67
[38] Searle A C, Sutton P J, Tinto M and Woan G 2008 Robust Bayesian detection of unmodelled bursts *Class. Quantum Gravity* **25** 114038–46
[39] Starkov V N, Semenov A A and Gomonay H V 2009 Numerical reconstruction of photon-number statistics from photocounting statistics: regularization of an ill-posed problem *Phys. Rev.* A **80** 013813
[40] Sun C P, Young K and Zou J 1999 Numerical algorithm for the determination of the potential of a conservative system from its normal mode spectra *J. Phys. A: Math. Gen.* **32** 3833–49
[41] Suyu S H, Marshall P J, Hobson M P and Blandfod R D 2006 A Bayesian analysis of regularised source inversions in gravitational lensing *Mon. Not. R. Astron. Soc.* **371** 983–98
[42] Szeliski R 1987 *Regularization uses fractal priors Natl Conf. on Artificial Intelligence* pp 749–54
[43] Terzopoulos D 1984 Multi-level surface reconstruction *Multiresolution Image Processing and Analysis* ed A Rosenfeld (Berlin: Springer)
[44] Terzopoulos D 1986 Regularization of visual problems involving discontinuities *IEEE Trans. Pattern Anal. Mach. Intell.* **8** 413–24
[45] Tikhonov A N 1963 Solution of incorrectly formulated problems and the regularization method *Sov. Math.* **4** 501–4 *(Engl. Transl.)*
[46] Tikhonov A N and Arsenin V Y 1977 *Solution of Ill-Posed Problems* (New York: Winston and Sons)
[47] Vestergaard B and Hansen S 2006 Application of Bayesian analysis to indirect fourier transformation in small-angle scattering *J. Appl. Crystallogr.* **39** 797–804
[48] Wasilkowski G W 1986 Optimal algorithms for linear problems with Gaussian measures *Rocky Mt. J. Math.* **16** 727–49
[49] Whaba G 1981 Bayesian confidence intervals for the cross validated smoothing spline *Technical report* University of Wisconsin
[50] Young N 1988 *An Introduction to Hilbert Space* (Cambridge: Cambridge Mathematical Textbooks)