The Principle of Energetic Consistency: Application to the Shallow-Water Equations

Stephen E. Cohn
Global Modeling and Assimilation Office
NASA Goddard Space Flight Center
Greenbelt, Maryland 20771

Popular Summary

It has often been said that the reason we cannot predict tomorrow's weather is that we don't know today's weather. More precisely, if the complete state of the earth's atmosphere (e.g., pressure, temperature, winds and humidity, everywhere throughout the atmosphere) were known at any particular initial time, then solving the equations that govern the dynamical behavior of the atmosphere would give the complete state at all subsequent times. Part of the difficulty of weather prediction is that the governing equations can only be solved approximately, which is what weather prediction models do. But weather forecasts would still be far from perfect even if the equations could be solved exactly, because the atmospheric state is not and cannot be known completely at any initial forecast time. Rather, the initial state for a weather forecast can only be estimated from incomplete observations taken near the initial time, through a process known as data assimilation.

Weather prediction models carry out their computations on a grid of points covering the earth's atmosphere. The formulation of these models is guided by a mathematical convergence theory which guarantees that, given the exact initial state, the model solution approaches the exact solution of the governing equations as the computational grid is made more fine. For the data assimilation process, however, there does not yet exist a convergence theory. For instance, it is not yet known how to formulate a data assimilation method in such a way that increasing the number of observations available to estimate the initial state is guaranteed to improve the accuracy of the estimated state, even in a statistical sense. Instead, the development of data assimilation methods has proceeded on the basis of a number of ad hoc assumptions and approximations.

This book chapter represents an effort to begin establishing a convergence theory for data assimilation methods. The main result, which is called

the principle of energetic consistency, provides a necessary condition that a convergent method must satisfy. Current methods violate this principle, as shown in earlier work of the author, and therefore are not convergent. The principle is illustrated by showing how to apply it as a simple test of convergence for proposed methods.

The Principle of Energetic Consistency: Application to the Shallow-Water Equations[1]

Stephen E. Cohn
Global Modeling and Assimilation Office
NASA Goddard Space Flight Center
Greenbelt, Maryland 20771

Draft of February 26, 2009

[1]prepared for *Data Assimilation: Making Sense of Observations*, W. Lahoz, B. Khattatov and R. Ménard (eds.), Springer

1 Introduction

The statement of conservation of total energy for nonlinear stochastic dynamical systems, when expressed in the natural energy variables of the system, provides an exact dynamical link between just the first two moments of the state of the system. This statement is what will be called here the *principle of energetic consistency*. This principle should be useful to the data assimilation community, because most current four-dimensional data assimilation methods are in fact based on an approximate evolution of the first two moments, conditioned on the observations. In particular, the principle provides one simple test of how well current methods approximate the actual evolution.

Suppose that the system state $\mathbf{s} = \mathbf{s}(t)$ is a vector governed by a nonlinear conservative system of ordinary differential equations

$$\frac{d\mathbf{s}}{dt} + \mathbf{f}(\mathbf{s}, t) = \mathbf{0},$$

where t is time, and suppose that the state variables have been chosen in such a way that the conserved quantity is $E = \mathbf{s}^T\mathbf{s}$, where the superscript T denotes transposition. Thus the statement of energy conservation is $E(t) = E(t_0)$. Now suppose that the initial state $\mathbf{s}(t_0)$ is a vector-valued random variable, with mean $\overline{\mathbf{s}}(t_0)$ and covariance matrix $\mathbf{P}(t_0)$. Then the principle of energetic consistency says that, under certain hypotheses,

$$\overline{\mathbf{s}}^T(t)\overline{\mathbf{s}}(t) + \operatorname{tr}\mathbf{P}(t) = \overline{\mathbf{s}}^T(t_0)\overline{\mathbf{s}}(t_0) + \operatorname{tr}\mathbf{P}(t_0),$$

where $\operatorname{tr}\mathbf{A}$ is the trace, or sum of the diagonal elements, of a matrix $\mathbf{A}$. The trace of the covariance matrix $\mathbf{P}(t)$ is sometimes called the total variance of the system state. Thus the principle of energetic consistency says that any increase (decrease) in the uncertainty in the state of the system, as measured by the total variance, is compensated for exactly by a corresponding decrease (increase) in the energy of the mean state. Some ramifications of the principle of energetic consistency for ordinary differential equations, in the context of both data assimilation schemes and predictability theory, were given in a recent paper of Cohn (2008). The principle holds also for nonlinear, conservative discrete-time systems.

The purpose of this chapter is to establish the principle of energetic consistency for a class of hyperbolic partial differential equations, and in particular, to determine precise conditions under which it holds. Ultimately a rigorous convergence theory for data assimilation methods will be needed. Given the role that energy considerations play in the convergence theory for discretizations of partial differential equations, the principle of energetic consistency is likely to play a role in a convergence theory for data assimilation methods.

The principle of energetic consistency for infinite-dimensional spaces is given as Theorem 1 in Section 2.3. It states that under appropriate hypotheses,

$$\|\overline{\mathbf{s}}_t\|^2 + \operatorname{tr}\mathcal{P}_t = \|\overline{\mathbf{s}}_{t_0}\|^2 + \operatorname{tr}\mathcal{P}_{t_0},$$

where the norm is a Hilbert space norm and $\mathcal{P}_t$ is the covariance operator of the system state. Hypotheses needed for the principle of energetic consistency to hold, for ordinary differential equations and for classical solutions of symmetric hyperbolic partial differential equations, are examined in Section 3. A main result is that for the latter case, the system state cannot be Gaussian-distributed, and the class of probability distributions that the system state can have is identified. The global nonlinear shallow-water equations are treated as an example in Section 4. There hypotheses are given so that the principle of energetic consistency holds when the state variables are taken to be the natural energy variables. It is shown that $\operatorname{tr} \mathcal{P}_t$ takes the simple form

$$\operatorname{tr} \mathcal{P}_t = \int \operatorname{tr} \mathbf{P}_t(\mathbf{x}, \mathbf{x})\, a \cos\phi \, d\phi \, d\lambda,$$

where $\mathbf{P}_t(\mathbf{x}, \mathbf{y})$ is the covariance matrix of the shallow-water system state.

The rigorous theory established in this chapter requires some mathematical machinery. The natural framework for the principle of energetic consistency is the theory of Hilbert space-valued random variables, which is covered in Appendix A. Appendix B covers the theory of families of Hilbert spaces which is needed to handle spherical geometry conveniently. Appendix C summarizes mathematical basics needed throughout the text.

2 The principle of energetic consistency

2.1 Problem setting

Let $\mathcal{H}$ be a real, separable Hilbert space, with inner product and corresponding norm denoted by $(\cdot, \cdot)$ and $\|\cdot\|$, respectively. Recall that every separable Hilbert space has a countable orthonormal basis, and that every orthonormal basis of a separable Hilbert space has the same number of elements $N \leq \infty$, the dimension of the space. Let $\{\mathbf{h}_i\}_{i=1}^N$ be an orthonormal basis for $\mathcal{H}$, where $N = \dim \mathcal{H} \leq \infty$ is the dimension of $\mathcal{H}$.

Let $\mathcal{S}$ be any nonempty set in $\mathcal{B}(\mathcal{H})$, where $\mathcal{B}(\mathcal{H})$ denotes the Borel field generated by the open sets in $\mathcal{H}$, i.e., $\mathcal{B}(\mathcal{H})$ is the smallest σ-algebra of subsets of $\mathcal{H}$ containing all the sets that are open in $\mathcal{H}$. In particular, $\mathcal{S} \subset \mathcal{H}$, $\mathcal{S}$ can be all of $\mathcal{H}$, and $\mathcal{S}$ can be any open or closed set in $\mathcal{H}$.

Let t_0 and T be two times with $-\infty < t_0 < T < \infty$, and let $\mathcal{T}$ be a time set bounded by and including t_0 and T. For instance, $\mathcal{T} = [t_0, T]$ in the case of continuous-time dynamics, and $\mathcal{T} = [t_0, t_1, \ldots, t_K = T]$ in the discrete-time case. The set $\mathcal{T}$ is allowed to depend on the set $\mathcal{S}$, $\mathcal{T} = \mathcal{T}(\mathcal{S})$.

Let $\mathbf{N}_{t,t_0}$ be a map from $\mathcal{S}$ into $\mathcal{H}$ (written $\mathbf{N}_{t,t_0} : \mathcal{S} \to \mathcal{H}$) for all times $t \in \mathcal{T}$, i.e., for all $\mathbf{s}_{t_0} \in \mathcal{S}$ and $t \in \mathcal{T}$, $\mathbf{N}_{t,t_0}(\mathbf{s}_{t_0})$ is defined and

$$\mathbf{s}_t = \mathbf{N}_{t,t_0}(\mathbf{s}_{t_0}) \tag{1}$$

is in $\mathcal{H}$, $\|\mathbf{s}_t\| < \infty$. Assume that $\mathbf{N}_{t,t_0}$ is continuous and bounded for all $t \in \mathcal{T}$. Continuity means that for every $t \in \mathcal{T}$, $\mathbf{s}_{t_0} \in \mathcal{S}$ and $\epsilon > 0$, there is a $\delta > 0$

such that if $\|\mathbf{s}_{t_0} - \mathbf{s}'_{t_0}\| < \delta$ and $\mathbf{s}'_{t_0} \in \mathcal{S}$, then $\|\mathbf{N}_{t,t_0}(\mathbf{s}_{t_0}) - \mathbf{N}_{t,t_0}(\mathbf{s}'_{t_0})\| < \epsilon$. Boundedness means that there is a constant $M = M_{t,t_0}$ such that

$$\|\mathbf{N}_{t,t_0}(\mathbf{s}_{t_0})\| \leq M_{t,t_0}\|\mathbf{s}_{t_0}\|$$

for all $\mathbf{s}_{t_0} \in \mathcal{S}$ and $t \in \mathcal{T}$. Continuity and boundedness are equivalent if $\mathbf{N}_{t,t_0}$ is a linear operator.

In most applications, $\mathbf{N}_{t,t_0}$ will be a nonlinear operator. Typically it will be the solution operator of a well-posed initial-value problem, for the state vector $\mathbf{s}$ of a nonlinear, deterministic system of partial ($\dim \mathcal{H} = \infty$) or ordinary ($\dim \mathcal{H} < \infty$) differential equations ($\mathcal{T} = [t_0, T]$).[1] Recall that continuity of the solution operator is part of the (Hadamard) definition of well-posedness of the initial-value problem for continuous-time or discrete-time dynamical systems: not only must there exist sets $\mathcal{S}$ and $\mathcal{T} = \mathcal{T}(\mathcal{S})$, taken here to be defined as above, and a unique solution $\mathbf{s}_t \in \mathcal{H}$ for all $\mathbf{s}_{t_0} \in \mathcal{S}$ and $t \in \mathcal{T}$, which taken together define the solution operator, but the solution must also depend continuously on the initial data.

The operator $\mathbf{N}_{t,t_0}$ is called isometric or conservative (in the norm $\|\cdot\|$ on $\mathcal{H}$) if

$$\|\mathbf{N}_{t,t_0}(\mathbf{s}_{t_0})\| = \|\mathbf{s}_{t_0}\|$$

for all $\mathbf{s}_{t_0} \in \mathcal{S}$ and $t \in \mathcal{T}$, and the differential (or difference) equations that express the dynamics of a well-posed initial-value problem are called conservative if the solution operator of the problem is conservative. With $\mathbf{s}_t \in \mathcal{H}$ defined for all $\mathbf{s}_{t_0} \in \mathcal{S}$ and $t \in \mathcal{T}$ by Eq. (1), the quantity

$$E_t = \|\mathbf{s}_t\|^2 = (\mathbf{s}_t, \mathbf{s}_t) < \infty \tag{2}$$

satisfies $E_t \leq M^2_{t,t_0} E_{t_0}$ for all $t \in \mathcal{T}$ under the assumption of boundedness, and is constant in time, $E_t = E_{t_0}$ for all $t \in \mathcal{T}$, in the conservative case.

In essence, the principle of energetic consistency is a statement about continuous transformations of Hilbert space which are conservative. Applied to solution operators, it becomes a statement about well-posed initial-value problems for conservative dynamics. It is important to recognize that the quantity E_t defined in Eq. (2) is quadratic in $\mathbf{s}_t$. For nonlinear systems of differential equations that express physical laws, there is usually a choice of dependent (state) variables such that E_t is the physical total energy. Then the dynamics are conservative in the norm on $\mathcal{H}$ if the physical system is closed, and the principle of energetic consistency applies.

2.2 Scalar and Hilbert space-valued random variables

Before stating the principle of energetic consistency, some probability concepts will first be summarized. For details, see Appendices A.1–A.3 and C.3.

[1] As discussed further in Section 3, for partial differential equations $\mathcal{H}$ will usually be the space $L^2(D)$ of square-integrable vectors on the spatial domain D of the problem and $\mathcal{S}$ will be an appropriate Sobolev or Sobolev-like space, while for ordinary differential equations $\mathcal{H}$ will usually be Euclidean space $\mathbb{R}^N$ and $\mathcal{S}$ will be an appropriate open set in $\mathbb{R}^N$.

Let $(\Omega, \mathcal{F}, P)$ be a complete probability space, with Ω the sample space, $\mathcal{F}$ the event space and P the probability measure. The event space consists of subsets of the set Ω, called events or measurable sets, which are those subsets on which the probability measure is defined. Denote by $\mathcal{E}$ the expectation operator.

A (scalar) random variable is a map $r : \Omega \to \mathbb{R}^e$ that is measurable, i.e., an extended real-valued function r, defined for all $\omega \in \Omega$, that satisfies

$$\{\omega \in \Omega : r(\omega) \leq x\} \in \mathcal{F}$$

for all $x \in \mathbb{R}$. Thus, if r is a random variable then its probability distribution function

$$F_r(x) = P(\{\omega \in \Omega : r(\omega) \leq x\})$$

is defined for all $x \in \mathbb{R}$. If r is a random variable then r^2 is a random variable.

Suppose that r is a random variable. Then the expectation $\mathcal{E}|r|$ is defined and $\mathcal{E}|r| \leq \infty$. If $\mathcal{E}|r| < \infty$, then the expectation $\mathcal{E}r$ is defined and called the mean of r, and $|\mathcal{E}r| \leq \mathcal{E}|r| < \infty$. If $\mathcal{E}r^2 < \infty$, then r is called second-order, the mean $\overline{r} = \mathcal{E}r$ and variance $\sigma^2 = \mathcal{E}(r - \overline{r})^2$ of r are defined, and

$$\mathcal{E}r^2 = \overline{r}^2 + \sigma^2. \tag{3}$$

An $\mathcal{H}$-valued random variable is a map $\mathbf{r} : \Omega \to \mathcal{H}$ such that

$$\{\omega \in \Omega : \mathbf{r}(\omega) \in B\} \in \mathcal{F}$$

for every set $B \in \mathcal{B}(\mathcal{H})$. A map $\mathbf{r} : \Omega \to \mathcal{H}$ is an $\mathcal{H}$-valued random variable if, and only if, $(\mathbf{h}, \mathbf{r})$ is a scalar random variable for every $\mathbf{h} \in \mathcal{H}$, that is, if and only if

$$\{\omega \in \Omega : (\mathbf{h}, \mathbf{r}(\omega)) \leq x\} \in \mathcal{F}$$

for all $\mathbf{h} \in \mathcal{H}$ and $x \in \mathbb{R}$. If $\mathbf{r}$ is an $\mathcal{H}$-valued random variable then $\|\mathbf{r}\|$ is a scalar random variable. An $\mathcal{H}$-valued random variable $\mathbf{r}$ is called second-order if $\|\mathbf{r}\|$ is a second-order scalar random variable, i.e., if $\mathcal{E}\|\mathbf{r}\|^2 < \infty$. If $\mathbf{r}$ is a second-order $\mathcal{H}$-valued random variable then $(\mathbf{h}, \mathbf{r})$ is a second-order scalar random variable, i.e., $\mathcal{E}(\mathbf{h}, \mathbf{r})^2 < \infty$, for all $\mathbf{h} \in \mathcal{H}$.

Suppose that $\mathbf{r}$ is a second-order $\mathcal{H}$-valued random variable. Then there exists a unique element $\overline{\mathbf{r}} \in \mathcal{H}$, called the mean of $\mathbf{r}$, such that $\mathcal{E}(\mathbf{h}, \mathbf{r}) = (\mathbf{h}, \overline{\mathbf{r}})$ for all $\mathbf{h} \in \mathcal{H}$. Also, $\mathbf{r}' = \mathbf{r} - \overline{\mathbf{r}}$ is a second-order $\mathcal{H}$-valued random variable with mean $\mathbf{0} \in \mathcal{H}$, and

$$\mathcal{E}\|\mathbf{r}\|^2 = \|\overline{\mathbf{r}}\|^2 + \mathcal{E}\|\mathbf{r}'\|^2.$$

Furthermore, there exists a unique bounded linear operator $\mathcal{P} : \mathcal{H} \to \mathcal{H}$, called the covariance operator of $\mathbf{r}$, such that

$$\mathcal{E}(\mathbf{g}, \mathbf{r}')(\mathbf{h}, \mathbf{r}') = (\mathbf{g}, \mathcal{P}\mathbf{h})$$

for all $\mathbf{g}, \mathbf{h} \in \mathcal{H}$. The covariance operator $\mathcal{P}$ is self-adjoint and positive semidefinite, i.e., $(\mathbf{g}, \mathcal{P}\mathbf{h}) = (\mathcal{P}\mathbf{g}, \mathbf{h})$ and $(\mathbf{h}, \mathcal{P}\mathbf{h}) \geq 0$ for all $\mathbf{g}, \mathbf{h} \in \mathcal{H}$. It is also trace class, i.e., the sum $\sum_{i=1}^{N} (\mathbf{h}_i, \mathcal{P}\mathbf{h}_i)$ is finite and independent of the orthonormal

basis $\{\mathbf{h}_i\}_{i=1}^N$, $N = \dim \mathcal{H} \leq \infty$, chosen for $\mathcal{H}$. This sum is called the trace of $\mathcal{P}$:

$$\operatorname{tr} \mathcal{P} = \sum_{i=1}^{N} (\mathbf{h}_i, \mathcal{P}\mathbf{h}_i) < \infty.$$

In addition, there exists an orthonormal basis for $\mathcal{H}$ which consists of eigenvectors $\{\tilde{\mathbf{h}}_i\}_{i=1}^N$ of $\mathcal{P}$,

$$\mathcal{P}\tilde{\mathbf{h}}_i = \lambda_i \tilde{\mathbf{h}}_i$$

for $i = 1, 2, \ldots, N$, and the corresponding eigenvalues $\{\lambda_i\}_{i=1}^N$ are all nonnegative. It follows that

$$\lambda_i = (\tilde{\mathbf{h}}_i, \mathcal{P}\tilde{\mathbf{h}}_i) = \mathcal{E}(\tilde{\mathbf{h}}_i, \mathbf{r}')^2 = \sigma_i^2,$$

where σ_i^2 is the variance of the second-order scalar random variable $(\tilde{\mathbf{h}}_i, \mathbf{r})$, for $i = 1, 2, \ldots, N$, and that

$$\operatorname{tr} \mathcal{P} = \sum_{i=1}^{N} \sigma_i^2 = \mathcal{E}\|\mathbf{r}'\|^2.$$

Thus the trace of $\mathcal{P}$ is also called the total variance of the second-order $\mathcal{H}$-valued random variable $\mathbf{r}$, and

$$\mathcal{E}\|\mathbf{r}\|^2 = \|\overline{\mathbf{r}}\|^2 + \mathcal{E}\|\mathbf{r}'\|^2 = \|\overline{\mathbf{r}}\|^2 + \sum_{i=1}^{N} \sigma_i^2 = \|\overline{\mathbf{r}}\|^2 + \operatorname{tr} \mathcal{P}. \tag{4}$$

Equation (4) generalizes Eq. (3), which holds for second-order scalar random variables, to the case of second-order $\mathcal{H}$-valued random variables.

Suppose that $\mathcal{R} \in \mathcal{B}(\mathcal{H})$. An $\mathcal{R}$-valued random variable is a map $\mathbf{r} : \Omega \to \mathcal{R}$ such that

$$\{\omega \in \Omega : \mathbf{r}(\omega) \in C\} \in \mathcal{F}$$

for every set $C \in \mathcal{B}_{\mathcal{R}}(\mathcal{H})$, where

$$\mathcal{B}_{\mathcal{R}}(\mathcal{H}) = \{B \in \mathcal{B}(\mathcal{H}) : B \subset \mathcal{R}\}.$$

Every $\mathcal{R}$-valued random variable is an $\mathcal{H}$-valued random variable, and every $\mathcal{H}$-valued random variable $\mathbf{r}$ with $\mathbf{r}(\omega) \in \mathcal{R}$ for all $\omega \in \Omega$ is an $\mathcal{R}$-valued random variable. An $\mathcal{R}$-valued random variable $\mathbf{r}$ is called second-order if $\|\mathbf{r}\|$ is a second-order scalar random variable. Thus every second-order $\mathcal{R}$-valued random variable is a second-order $\mathcal{H}$-valued random variable, and every second-order $\mathcal{H}$-valued random variable $\mathbf{r}$ with $\mathbf{r}(\omega) \in \mathcal{R}$ for all $\omega \in \Omega$ is a second-order $\mathcal{R}$-valued random variable. Finally, if $\mathbf{r}$ is an $\mathcal{R}$-valued random variable and $\mathbf{N}$ is a continuous map from $\mathcal{R}$ into $\mathcal{H}$, then $\mathbf{N}(\mathbf{r})$ is an $\mathcal{H}$-valued random variable.

2.3 The principle of energetic consistency in Hilbert space

Referring now back to Section 2.1, consider for $\mathbf{s}_{t_0}$ not just a single element of $\mathcal{S}$, but rather a whole collection of elements $\mathbf{s}_{t_0}(\omega)$ indexed by the probability variable $\omega \in \Omega$. Suppose at first that $\mathbf{s}_{t_0}$ is simply a map $\mathbf{s}_{t_0} : \Omega \to \mathcal{S}$, i.e., that $\mathbf{s}_{t_0}(\omega)$ is defined for all $\omega \in \Omega$ and $\mathbf{s}_{t_0}(\omega) \in \mathcal{S}$ for all $\omega \in \Omega$. Then since $\mathbf{N}_{t,t_0} : \mathcal{S} \to \mathcal{H}$ for all $t \in \mathcal{T}$, it follows that $\mathbf{s}_t = \mathbf{N}_{t,t_0}(\mathbf{s}_{t_0}) : \Omega \to \mathcal{H}$ for all $t \in \mathcal{T}$, with

$$\mathbf{s}_t(\omega) = \mathbf{N}_{t,t_0}(\mathbf{s}_{t_0}(\omega))$$

and $\|\mathbf{s}_t(\omega)\| < \infty$, for all $\omega \in \Omega$ and $t \in \mathcal{T}$.

Suppose further that $\mathbf{s}_{t_0}$ is an $\mathcal{S}$-valued random variable. Then it follows from the continuity assumption on $\mathbf{N}_{t,t_0}$ that $\mathbf{s}_t$ is an $\mathcal{H}$-valued random variable, and therefore that $E_t = \|\mathbf{s}_t\|^2$ is a scalar random variable, for all $t \in \mathcal{T}$.

Suppose still further that $\mathbf{s}_{t_0}$ is a second-order $\mathcal{S}$-valued random variable, $\mathcal{E}E_{t_0} = \mathcal{E}\|\mathbf{s}_{t_0}\|^2 < \infty$. Then from the boundedness assumption on $\mathbf{N}_{t,t_0}$,

$$\|\mathbf{s}_t(\omega)\|^2 \leq M_{t,t_0}^2 \|\mathbf{s}_{t_0}(\omega)\|^2$$

for all $\omega \in \Omega$ and $t \in \mathcal{T}$, it follows that

$$\mathcal{E}E_t = \mathcal{E}\|\mathbf{s}_t\|^2 \leq M_{t,t_0}^2 \mathcal{E}\|\mathbf{s}_{t_0}\|^2 < \infty$$

for all $t \in \mathcal{T}$. Therefore, $\mathbf{s}_t$ is a second-order $\mathcal{H}$-valued random variable, with mean $\overline{\mathbf{s}}_t \in \mathcal{H}$, covariance operator $\mathcal{P}_t : \mathcal{H} \to \mathcal{H}$, and

$$\mathcal{E}\|\mathbf{s}_t\|^2 = \|\overline{\mathbf{s}}_t\|^2 + \operatorname{tr} \mathcal{P}_t,$$

for all $t \in \mathcal{T}$. Thus the principle of energetic consistency has been established:

Theorem 1 *Let $\mathcal{H}$, $\mathcal{S}$, $\mathcal{T}$ and $\mathbf{N}_{t,t_0}$ be as stated in Section 2.1, with $\mathbf{N}_{t,t_0}$ continuous and bounded for all $t \in \mathcal{T}$, and let $\mathcal{E}$ be the expectation operator on a complete probability space $(\Omega, \mathcal{F}, P)$. If $\mathbf{s}_{t_0}$ is a second-order $\mathcal{S}$-valued random variable, then for all $t \in \mathcal{T}$, (i) $\mathbf{s}_t = \mathbf{N}_{t,t_0}(\mathbf{s}_{t_0})$ is a second-order $\mathcal{H}$-valued random variable, (ii) $E_t = \|\mathbf{s}_t\|^2$ is a scalar random variable, (iii) $\mathbf{s}_t$ has mean $\overline{\mathbf{s}}_t \in \mathcal{H}$ and covariance operator $\mathcal{P}_t : \mathcal{H} \to \mathcal{H}$, (iv)*

$$\mathcal{E}E_t = \|\overline{\mathbf{s}}_t\|^2 + \operatorname{tr} \mathcal{P}_t,$$

and (v)

$$\|\overline{\mathbf{s}}_t\|^2 + \operatorname{tr} \mathcal{P}_t \leq M_{t,t_0}^2 (\|\overline{\mathbf{s}}_{t_0}\|^2 + \operatorname{tr} \mathcal{P}_{t_0}). \tag{5}$$

If, in addition, $\mathbf{N}_{t,t_0}$ is conservative, then (vi)

$$\|\overline{\mathbf{s}}_t\|^2 + \operatorname{tr} \mathcal{P}_t = \|\overline{\mathbf{s}}_{t_0}\|^2 + \operatorname{tr} \mathcal{P}_{t_0} \tag{6}$$

for all $t \in \mathcal{T}$.

It is in the conservative case that the principle of energetic consistency is most useful, because in that case, Eq. (6) provides an equality against which, for instance, approximate moment evolution schemes can be compared. In case $\mathbf{N}_{t,t_0}$ is only bounded, for example in the presence of dissipation, or for initial-boundary value problems with a net flux of energy across the boundaries, Eq. (5) still provides an upper bound on the total variance $\operatorname{tr} \mathcal{P}_t$.

2.4 A natural restriction on $\mathcal{S}$

Suppose for the moment that $\mathbf{s}_{t_0}$ is an $\mathcal{S}$-valued random variable, not necessarily second-order. When the squared norm on $\mathcal{H}$ represents a physical total energy, it is natural to impose the restriction that every possible initial state $\mathbf{s}_{t_0}(\omega)$, $\omega \in \Omega$, has total energy less than some finite maximum amount, say $E_* < \infty$, i.e. that $\mathcal{S} \subset \mathcal{H}_{E_*}$, where $\mathcal{H}_E$ is defined for all $E > 0$ as the open set

$$\mathcal{H}_E = \{\mathbf{s} \in \mathcal{H} : \|\mathbf{s}\|^2 < E\}. \tag{7}$$

Otherwise, given any total energy E, no matter how large, there would be a nonzero probability that $\mathbf{s}_{t_0}$ has total energy greater than or equal to E:

$$P(\{\omega \in \Omega : \|\mathbf{s}_{t_0}(\omega)\|^2 \geq E\}) > 0.$$

Of course, it can be argued that since this probability would be very small for E very large, it may be acceptable as an approximation not to impose this restriction. On the other hand, as discussed in Section 3.2 and illustrated in Section 4, for classical solutions of hyperbolic systems of partial differential equations, it is necessary to require that $\mathcal{S} \subset \mathcal{H}_{E_*}$ for some $E_* < \infty$ just to ensure well-posedness. Thus the restriction is often not only natural, but also necessary. It also simplifies matters, as discussed next, for it makes $\mathbf{s}_{t_0}$ second-order automatically and gives $\mathbf{s}_t = \mathbf{N}_{t,t_0}(\mathbf{s}_{t_0})$ some additional desirable properties, and it also yields a convenient characterization of $\mathbf{s}_{t_0}$ and $\mathbf{s}_t$.

Suppose that $\mathbf{s}_{t_0}$ is an $\mathcal{S}$-valued random variable, and that $\mathcal{S} \subset \mathcal{H}_{E_*}$ for some $E_* < \infty$. Thus $\|\mathbf{s}_{t_0}(\omega)\|^2 < E_*$ for all $\omega \in \Omega$, and therefore $\mathcal{E}\|\mathbf{s}_{t_0}\|^2 < E_*$, i.e., $\mathbf{s}_{t_0}$ is a second-order $\mathcal{S}$-valued random variable. Therefore, for all $t \in \mathcal{T}$, $\mathbf{s}_t = \mathbf{N}_{t,t_0}(\mathbf{s}_{t_0})$ is a second-order $\mathcal{H}$-valued random variable, in fact with $\mathbf{s}_t(\omega) \in \mathcal{H}_E$ for all $\omega \in \Omega$, where $E = E_*$ in the conservative case and $E = M_{t,t_0}^2 E_*$ in the merely bounded case. Since $\mathcal{H}_E$ is an open set in $\mathcal{H}$, $\mathcal{H}_E \in \mathcal{B}(\mathcal{H})$. Therefore, for all $t \in \mathcal{T}$, $\mathbf{s}_t$ is an $\mathcal{H}_E$-valued random variable. Further, for all $p > 0$ and $t \in \mathcal{T}$, $\mathcal{E}\|\mathbf{s}_t\|^p < E^{p/2}$. Thus $\|\mathbf{s}_t\|$ has finite moments of all orders, for all $t \in \mathcal{T}$.

Now suppose that $\mathbf{s}$ is an $\mathcal{H}_E$-valued random variable, for some $E < \infty$. Then since $\mathbf{s}$ is also an $\mathcal{H}$-valued random variable, $(\mathbf{h}_i, \mathbf{s})$ is a scalar random variable for $i = 1, \ldots, N$, where $\{\mathbf{h}_i\}_{i=1}^N$ is any orthonormal basis for $\mathcal{H}$ and $N = \dim \mathcal{H} \leq \infty$. Since $\mathbf{s}(\omega) \in \mathcal{H}$ for all $\omega \in \Omega$, $\mathbf{s}(\omega)$ has the representation

$$\mathbf{s}(\omega) = \sum_{i=1}^{N} (\mathbf{h}_i, \mathbf{s}(\omega))\mathbf{h}_i$$

for each $\omega \in \Omega$, and by Parseval's relation,

$$\|\mathbf{s}(\omega)\|^2 = \sum_{i=1}^{N} (\mathbf{h}_i, \mathbf{s}(\omega))^2 < E$$

for each $\omega \in \Omega$.

It is shown in Appendix A.4 that if $\{s_i\}_{i=1}^N$ is any collection of scalar random variables with $\sum_{i=1}^N \mathcal{E} s_i^2 < \infty$, where $N = \dim \mathcal{H} \leq \infty$, then there is a second-order $\mathcal{H}$-valued random variable $\tilde{\mathbf{s}}$ such that $(\mathbf{h}_i, \tilde{\mathbf{s}}(\omega)) = s_i(\omega)$ for $i = 1, \ldots, N$ and for all $\omega \in \Omega$ with $\sum_{i=1}^N s_i^2(\omega) < \infty$. Therefore, if $\{s_i\}_{i=1}^N$ is any collection of scalar random variables with $\sum_{i=1}^N s_i^2(\omega) < E$ for all $\omega \in \Omega$, then there is a second-order $\mathcal{H}$-valued random variable $\tilde{\mathbf{s}}$ such that $(\mathbf{h}_i, \tilde{\mathbf{s}}(\omega)) = s_i(\omega)$ for $i = 1, \ldots, N$ and for *all* $\omega \in \Omega$, in which case $\tilde{\mathbf{s}}(\omega) = \sum_{i=1}^N s_i(\omega)\mathbf{h}_i$ for all $\omega \in \Omega$, and so by Parseval's relation, this $\tilde{\mathbf{s}}$ is an $\mathcal{H}_E$-valued random variable.

Thus, a map $\mathbf{s} : \Omega \to \mathcal{H}$ is an $\mathcal{H}_E$-valued random variable if, and only if,

$$\mathbf{s}(\omega) = \sum_{i=1}^N s_i(\omega)\mathbf{h}_i$$

for all $\omega \in \Omega$, where $\{s_i\}_{i=1}^N$ is a collection of scalar random variables with

$$\sum_{i=1}^N s_i^2(\omega) < E$$

for all $\omega \in \Omega$, in which case

$$s_i(\omega) = (\mathbf{h}_i, \mathbf{s}(\omega))$$

for $i = 1, \ldots, N$ and for all $\omega \in \Omega$. In particular, $|s_i(\omega)| < E^{1/2}$ for $i = 1, \ldots, N$ and for all $\omega \in \Omega$, which is a strong restriction on the scalar random variables $s_i = (\mathbf{h}_i, \mathbf{s})$. It implies immediately that the probability distribution functions

$$F_{(\mathbf{h}_i, \mathbf{s})}(x) = P(\{\omega \in \Omega : (\mathbf{h}_i, \mathbf{s}(\omega)) \leq x\})$$

must satisfy

$$F_{(\mathbf{h}_i, \mathbf{s})}(x) = \begin{cases} 0 & \text{if } x \leq -E^{1/2} \\ 1 & \text{if } x \geq E^{1/2} \end{cases}$$

for $i = 1, \ldots, N$. Thus $(\mathbf{h}_i, \mathbf{s})$ cannot be Gaussian-distributed, for instance, for any $i = 1, \ldots, N$. Also, since $\|\mathbf{s}(\omega)\| < E^{1/2}$ for all $\omega \in \Omega$, the probability distribution function

$$F_{\|\mathbf{s}\|}(x) = P(\{\omega \in \Omega : \|\mathbf{s}(\omega)\| \leq x\})$$

of the scalar random variable $\|\mathbf{s}\|$ must satisfy

$$F_{\|\mathbf{s}\|}(x) = \begin{cases} 0 & \text{if } x \leq 0 \\ 1 & \text{if } x \geq E^{1/2} \end{cases} .$$

The characterization of $\mathcal{H}_E$-valued random variables given above will be used in Section 4 to construct an $\mathcal{H}_E$-valued random variable $\mathbf{s}_{t_0}$ for the shallow-water equations. This will guarantee directly that the random initial geopotential field is positive.

3 The principle of energetic consistency for differential equations

3.1 Ordinary differential equations

Consider a nonlinear system of ordinary differential equations

$$\frac{d\mathbf{s}}{dt} + \mathbf{f}(\mathbf{s}, t) = \mathbf{0}, \tag{8}$$

where $\mathbf{f} : \mathcal{S}_1 \times \mathcal{T}_1 \to \mathbb{R}^N$, with $\mathcal{S}_1$ an open connected set in $\mathbb{R}^N$, possibly all of $\mathbb{R}^N$, and with $\mathcal{T}_1 = [t_0, T_1]$ and $0 < T_1 - t_0 < \infty$. Take $\mathcal{H} = \mathbb{R}^N$, with $(\cdot, \cdot)$ denoting the Euclidean inner product, $(\mathbf{g}, \mathbf{h}) = \mathbf{g}^T\mathbf{h}$ for all $\mathbf{g}, \mathbf{h} \in \mathbb{R}^N$, and $\|\cdot\|$ the Euclidean norm, $\|\mathbf{h}\| = (\mathbf{h}^T\mathbf{h})^{1/2}$ for all $\mathbf{h} \in \mathbb{R}^N$.

Assume that $\mathbf{f}$ is of class $C(\mathcal{S}_1 \times \mathcal{T}_1)$, i.e., that $\mathbf{f}$ is continuous on its domain of definition $\mathcal{S}_1 \times \mathcal{T}_1$. Assume also that $\mathbf{f}$ is bounded on $\mathcal{S}_1 \times \mathcal{T}_1$:

$$\|\mathbf{f}(\mathbf{s}, t)\| < C_1$$

for some constant C_1, for all $\mathbf{s} \in \mathcal{S}_1$ and $t \in \mathcal{T}_1$. Note that the latter assumption follows from the former one if $\mathcal{S}_1 \subset \mathcal{H}_E$ for some $E < \infty$, where $\mathcal{H}_E$ was defined in Eq. (7). Assume finally that $\mathbf{f}$ is Lipschitz continuous in its first argument, uniformly in time, i.e., that there is a constant C_2 such that

$$\|\mathbf{f}(\mathbf{r}, t) - \mathbf{f}(\mathbf{s}, t)\| \le C_2\|\mathbf{r} - \mathbf{s}\|$$

for all $\mathbf{r}, \mathbf{s} \in \mathcal{S}_1$ and $t \in \mathcal{T}_1$.

A real N-vector function $\mathbf{s} = \mathbf{s}(t)$ defined on an interval $\mathcal{T}_* = [t_0, T_*]$, $T_* \in (t_0, T_1]$, is called a (continuous) solution of Eq. (8) if, for all $t \in \mathcal{T}_*$, (i) $\mathbf{s}(t) \in \mathcal{S}_1$, (ii) $\mathbf{s}(t)$ is continuous, and (iii) $\mathbf{s}(t)$ satisfies Eq. (8) pointwise. It follows from the continuity assumption on $\mathbf{f}$ that if $\mathbf{s}$ is a solution on an interval $\mathcal{T}_*$, then $d\mathbf{s}/dt$ is continuous on $\mathcal{T}_*$, and so

$$\frac{d}{dt}\|\mathbf{s}\|^2 = \frac{d}{dt}(\mathbf{s}, \mathbf{s}) = 2(\mathbf{s}, \frac{d\mathbf{s}}{dt}) = -2(\mathbf{s}, \mathbf{f}(\mathbf{s}, t)) \tag{9}$$

is also continuous on $\mathcal{T}_*$, hence integrable on $\mathcal{T}_*$. Similarly, if $\mathbf{r}$ and $\mathbf{s}$ are two solutions on an interval $\mathcal{T}_*$, then by the Schwarz inequality,

$$\|\mathbf{r} - \mathbf{s}\| \left|\frac{d\|\mathbf{r} - \mathbf{s}\|}{dt}\right| = |(\mathbf{r} - \mathbf{s}, \mathbf{f}(\mathbf{r}, t) - \mathbf{f}(\mathbf{s}, t))| \le \|\mathbf{r} - \mathbf{s}\|\,\|\mathbf{f}(\mathbf{r}, t) - \mathbf{f}(\mathbf{s}, t)\|$$

for all $t \in \mathcal{T}_*$, and so by integrating it follows from the Lipschitz continuity assumption that

$$\|\mathbf{r}(t) - \mathbf{s}(t)\| \le e^{C_2(t-t_0)}\|\mathbf{r}(t_0) - \mathbf{s}(t_0)\|$$

for all $t \in \mathcal{T}_*$. Thus, if $\mathbf{r}(t_0) = \mathbf{s}(t_0)$ then $\mathbf{r}(t) = \mathbf{s}(t)$ for all $t \in \mathcal{T}_*$: for each $\mathbf{s}_{t_0} \in \mathcal{S}_1$ there exists at most one solution $\mathbf{s}(t)$ defined on an interval $\mathcal{T}_*$, such

that $\mathbf{s}(t_0) = \mathbf{s}_{t_0}$. The inequality also shows that if such a solution exists, then it depends continuously on $\mathbf{s}_{t_0}$, for all $t \in \mathcal{T}_*$.

The continuity and boundedness assumptions on $\mathbf{f}$ together imply that, for each $\mathbf{s}_{t_0} \in \mathcal{S}_1$, there does exist a solution $\mathbf{s}(t)$ with $\mathbf{s}(t_0) = \mathbf{s}_{t_0}$, and that it remains in existence either until time $t = T_1$ or the first time that the solution hits the boundary $\partial\mathcal{S}_1$ of $\mathcal{S}_1$, where $\mathbf{f}$ may not be defined, whichever is smaller (e.g. Coddington and Levinson (1955, pp. 6, 15)). Thus, if $\mathcal{S}_1 = \mathbb{R}^N$, then the solution exists until time T_1. This time can be arbitrarily large, for instance if $\mathbf{f}$ is independent of time. More generally, a minimum existence time can be found by noting that the solution $\mathbf{s}(t)$ with $\mathbf{s}(t_0) = \mathbf{s}_{t_0} \in \mathcal{S}_1$ must satisfy the integral equation

$$\mathbf{s}(t) = \mathbf{s}_{t_0} - \int_{t_0}^{t} \mathbf{f}(\mathbf{s}(\tau), \tau)\, d\tau,$$

and so

$$\|\mathbf{s}(t) - \mathbf{s}_{t_0}\| < C_1(t - t_0)$$

by the boundedness assumption on $\mathbf{f}$, for as long as the solution exists. Denoting by $\rho(\mathbf{s}_{t_0})$ the Euclidean distance from any $\mathbf{s}_{t_0} \in \mathcal{S}_1$ to $\partial\mathcal{S}_1$,

$$\rho(\mathbf{s}_{t_0}) = \inf_{\mathbf{h} \in \partial\mathcal{S}_1} \|\mathbf{h} - \mathbf{s}_{t_0}\|,$$

it follows that $\|\mathbf{s}(t) - \mathbf{s}_{t_0}\| < \rho(\mathbf{s}_{t_0})$ if $t - t_0 \leq \rho(\mathbf{s}_{t_0})/C_1$, and so the solution exists on $\mathcal{T}_* = [t_0, T_*]$ for

$$T_* = T_*(\mathbf{s}_{t_0}) = \min(T_1, t_0 + \rho(\mathbf{s}_{t_0})/C_1).$$

Note that $\rho(\mathbf{s}_{t_0}) > 0$ for each $\mathbf{s}_{t_0} \in \mathcal{S}_1$ since $\mathcal{S}_1$ is an open set, and therefore $T_* > t_0$.

The principle of energetic consistency requires a set $\mathcal{S} \in \mathcal{B}(\mathcal{H}) = \mathcal{B}(\mathbb{R}^N)$ for the initial data and a time interval $\mathcal{T} = [t_0, T]$ such that, for *every* $\mathbf{s}_{t_0} \in \mathcal{S}$, the corresponding solution exists on $\mathcal{T}$, i.e., every solution must exist for the same minimum amount of time $T - t_0 > 0$, independently of the location of $\mathbf{s}_{t_0} \in \mathcal{S}$. If $\mathcal{S}_1 = \mathbb{R}^N$, then take $\mathcal{S} = \mathbb{R}^N$ and $\mathcal{T} = [t_0, T_1]$. Otherwise, let $\mathcal{S}$ be any open set in $\mathbb{R}^N$ which is contained in the interior of $\mathcal{S}_1$, and denote by $\rho_{\mathcal{S}}$ the minimum Euclidean distance from the boundary of $\mathcal{S}$ to that of $\mathcal{S}_1$. Then

$$\rho(\mathbf{s}_{t_0}) > \rho_{\mathcal{S}} = \inf_{\mathbf{s} \in \mathcal{S}} \rho(\mathbf{s}) > 0$$

for all $\mathbf{s}_{t_0} \in \mathcal{S}$, and setting

$$T = T_{\mathcal{S}} = \min(T_1, t_0 + \rho_{\mathcal{S}}/C_1)$$

and $\mathcal{T} = \mathcal{T}(\mathcal{S}) = [t_0, T]$, it follows that the unique solution $\mathbf{s}(t)$ corresponding to each $\mathbf{s}_{t_0} \in \mathcal{S}$ exists for all $t \in \mathcal{T}$. Denoting this solution by $\mathbf{s}_t = \mathbf{N}_{t,t_0}(\mathbf{s}_{t_0})$, it follows that $\mathbf{N}_{t,t_0}$ is defined uniquely on $\mathcal{S}$, as a continuous map from $\mathcal{S}$ into $\mathcal{H} = \mathbb{R}^N$, for all $t \in \mathcal{T}$.

It follows from Eq. (9) that the solution operator $\mathbf{N}_{t,t_0}$ is conservative if $(\mathbf{s}, \mathbf{f}(\mathbf{s},t)) = 0$ for all $\mathbf{s} \in \mathcal{S}_1$ and $t \in \mathcal{T}$. More generally, it follows that if there is a constant C_3 such that

$$|(\mathbf{s}, \mathbf{f}(\mathbf{s},t))| \leq C_3 \|\mathbf{s}\|^2$$

for all $\mathbf{s} \in \mathcal{S}_1$ and $t \in \mathcal{T}$, then $\mathbf{N}_{t,t_0}$ is bounded, with

$$\|\mathbf{s}_t\| = \|\mathbf{N}_{t,t_0}(\mathbf{s}_{t_0})\| \leq e^{C_3(t-t_0)} \|\mathbf{s}_{t_0}\|$$

for all $t \in \mathcal{T}$. Note that if $\mathbf{f}(\mathbf{s},t)$ depends linearly on $\mathbf{s}$ near the origin of coordinates $\mathbf{0} \in \mathbb{R}^N$, or if $\mathbf{0} \notin \mathcal{S}_1$, then by the boundedness assumption on $\mathbf{f}$ there is a constant C_3 such that $\|\mathbf{f}(\mathbf{s},t))\| \leq C_3\|\mathbf{s}\|$ for all $\mathbf{s} \in \mathcal{S}_1$ and $t \in \mathcal{T}$, so that by the Schwarz inequality,

$$|(\mathbf{s}, \mathbf{f}(\mathbf{s},t))| \leq \|\mathbf{s}\| \, \|\mathbf{f}(\mathbf{s},t))\| \leq C_3 \|\mathbf{s}\|^2$$

for all $\mathbf{s} \in \mathcal{S}_1$ and $t \in \mathcal{T}$, and therefore $\mathbf{N}_{t,t_0}$ is bounded for all $t \in \mathcal{T}$.

3.2 Symmetric hyperbolic partial differential equations

3.2.1 The deterministic initial-value problem

Consider now a nonlinear system of partial differential equations

$$\frac{\partial \mathbf{s}}{\partial t} + \mathbf{G}\mathbf{s} = \mathbf{0}, \tag{10}$$

where $\mathbf{G} = \mathbf{G}(\mathbf{s}) = \mathbf{G}(\mathbf{s}, \mathbf{x}, t)$ is a linear differential operator of first order in space variables $\mathbf{x} = (x_1, \ldots, x_d)^T$,

$$\mathbf{G} = \sum_{j=1}^{d} \mathbf{A}_j(\mathbf{s}, \mathbf{x}, t) \frac{\partial}{\partial x_j} + \mathbf{A}_{d+1}(\mathbf{s}, \mathbf{x}, t),$$

and $\mathbf{A}_1, \ldots, \mathbf{A}_{d+1}$ are real $n \times n$ matrices. For simplicity assume that the d-dimensional spatial domain D of the problem is

$$D = \{\mathbf{x} \in \mathbb{R}^d : |x_j| \leq L_j, j = 1, \ldots, d\},$$

with periodic boundary conditions at the endpoints $x_j = \pm L_j$, $j = 1, \ldots, d$. Consider endpoint $x_j = L_j$ to be identified with endpoint $x_j = -L_j$, for each $j = 1, \ldots, d$, so that a continuous function on D satisfies the periodic boundary conditions automatically. (Spherical geometry will be treated in Section 4.) Take $\mathcal{H} = L^2(D)$, the Hilbert space of real, Lebesgue square-integrable n-vectors on D, with inner product

$$(\mathbf{g}, \mathbf{h}) = \int_D \mathbf{g}^T(\mathbf{x}) \mathbf{h}(\mathbf{x}) \, dx_1 \cdots dx_d$$

for all $\mathbf{g}, \mathbf{h} \in L^2(D)$, and corresponding norm $\|\mathbf{h}\| = (\mathbf{h}, \mathbf{h})^{1/2}$ for all $\mathbf{h} \in L^2(D)$.

Assume that the matrices $\mathbf{A}_1, \ldots, \mathbf{A}_{d+1}$ are defined on all of $\mathbb{R}^n \times D \times \mathcal{T}_1$, where $\mathcal{T}_1 = [t_0, T_1]$ and $0 < T_1 - t_0 < \infty$. Assume further that each matrix is of class $C^\infty(\mathbb{R}^n \times D \times \mathcal{T}_1)$, i.e., that all of the matrix elements and all of their partial derivatives are continuous functions on $\mathbb{R}^n \times D \times \mathcal{T}_1$ and satisfy the periodic boundary conditions in the space variables. Assume finally that $\mathbf{A}_1, \ldots, \mathbf{A}_d$ (but not $\mathbf{A}_{d+1}$) are symmetric matrices.

A real n-vector function $\mathbf{s} = \mathbf{s}(\mathbf{x}, t)$ defined on $D \times \mathcal{T}_*$, with $\mathcal{T}_* = [t_0, T_*]$ and $T_* \in (t_0, T_1]$, is called a classical solution of the symmetric hyperbolic system Eq. (10) if (i) $\mathbf{s} \in C^1(D) \cap C^1(\mathcal{T}_*)$ and (ii) $\mathbf{s}$ satisfies Eq. (10) pointwise in $D \times \mathcal{T}_*$. The condition $\mathbf{s} \in C^1(D) \cap C^1(\mathcal{T}_*)$ means that the components of the vector $\mathbf{s}$ and their first time and space derivatives are continuous on D for each fixed $t \in \mathcal{T}_*$, are continuous on $\mathcal{T}_*$ for each fixed $\mathbf{x} \in D$, and satisfy the periodic boundary conditions. The initial condition for a classical solution is a real n-vector function $\mathbf{s}_{t_0} \in C^1(D)$.

Suppose for the moment that $\mathbf{s} = \mathbf{s}(\mathbf{x}, t)$ is a classical solution on $D \times \mathcal{T}_*$. Then

$$\frac{d}{dt}\|\mathbf{s}\|^2 = \frac{d}{dt}(\mathbf{s}, \mathbf{s}) = 2(\mathbf{s}, \frac{\partial \mathbf{s}}{\partial t}) = -2(\mathbf{s}, \mathbf{G}(\mathbf{s})\mathbf{s}) \tag{11}$$

is continuous on $\mathcal{T}_*$. Also, by using the symmetry of $\mathbf{A}_1, \ldots, \mathbf{A}_d$ and the periodic boundary conditions, an integration by parts gives

$$(\mathbf{s}, \mathbf{G}(\mathbf{s})\mathbf{s}) = \int_D \mathbf{s}^T(\mathbf{x}, t)\mathbf{B}(\mathbf{x}, t)\mathbf{s}(\mathbf{x}, t)\, dx_1 \cdots dx_d \tag{12}$$

for all $t \in \mathcal{T}_*$, where

$$\mathbf{B}(\mathbf{x}, t) = \mathbf{A}_{d+1}(\mathbf{s}, \mathbf{x}, t) - \tfrac{1}{2}\sum_{j=1}^{d} \frac{d\mathbf{A}_j(\mathbf{s}, \mathbf{x}, t)}{dx_j}$$

and

$$\frac{d\mathbf{A}_j}{dx_j} = \sum_{i=1}^{n} \frac{\partial \mathbf{A}_j}{\partial s_i}\frac{\partial s_i}{\partial x_j} + \frac{\partial \mathbf{A}_j}{\partial x_j}.$$

Further, since $\mathbf{s} \in C^1(D) \cap C^1(\mathcal{T}_*)$, the components of $\mathbf{s}$ and their first partial derivatives with respect to the space variables are bounded functions on $D \times \mathcal{T}_*$. Define $\beta_0 = \beta_0(\mathbf{s})$ by

$$\beta_0 = \max_{D \times \mathcal{T}_*} \sum_{i=1}^{n} |s_i(\mathbf{x}, t)|,$$

and $\beta_j = \beta_j(\mathbf{s})$ by

$$\beta_j = \max_{D \times \mathcal{T}_*} \sum_{i=1}^{n} \left| \frac{\partial s_i(\mathbf{x}, t)}{\partial x_j} \right|,$$

for $j = 1, \ldots, d$. Then it follows from Eq. (12) and the continuity assumption on the matrices $\mathbf{A}_1, \ldots, \mathbf{A}_{d+1}$ that there is a continuous function $C_1 = C_1(\beta_0, \ldots, \beta_d)$ such that

$$|(\mathbf{s}, \mathbf{G}(\mathbf{s})\mathbf{s})| \le C_1\|\mathbf{s}\|^2$$

for all $t \in \mathcal{T}_*$. Equation (11) then implies that

$$\|\mathbf{s}(\cdot,t)\| \le e^{C_1(t-t_0)}\|\mathbf{s}(\cdot,t_0)\| \tag{13}$$

for all $t \in \mathcal{T}_*$.

A similar argument shows that if $\mathbf{r}$ and $\mathbf{s}$ are two classical solutions on $D\times\mathcal{T}_*$, then there is a continuous function $C_2 = C_2(\beta_0(\mathbf{r}),\ldots,\beta_d(\mathbf{r}),\beta_0(\mathbf{s}),\ldots,\beta_d(\mathbf{s}))$ such that

$$\|\mathbf{r}(\cdot,t)-\mathbf{s}(\cdot,t)\| \le e^{C_2(t-t_0)}\|\mathbf{r}(\cdot,t_0)-\mathbf{s}(\cdot,t_0)\| \tag{14}$$

for all $t \in \mathcal{T}_*$. Therefore, for each $\mathbf{s}_{t_0} \in C^1(D)$, there exists at most one classical solution $\mathbf{s}$ on $D \times \mathcal{T}_*$ such that $\mathbf{s}(\mathbf{x},t_0) = \mathbf{s}_{t_0}(\mathbf{x})$ for all $\mathbf{x} \in D$. This inequality does not imply that if such a solution exists, then it depends continuously on $\mathbf{s}_{t_0}$ in the norm $\|\cdot\|$, unless C_2 can be made to depend only on $\mathbf{r}_{t_0}$ and $\mathbf{s}_{t_0}$. This is accomplished by means of the existence theory itself, discussed next.

Denote by $H^k = H^k(D)$, for $k = 0, 1, \ldots$, the Sobolev space of real n-vectors on D with k Lebesgue square-integrable derivatives on D. The spaces H^k are Hilbert spaces, with inner product

$$(\mathbf{g},\mathbf{h})_{H^k} = \sum_{l=0}^{k}\sum_{l_1+\cdots+l_d=l}(D^l\mathbf{g}, D^l\mathbf{h})$$

for all $\mathbf{g},\mathbf{h} \in H^k$, where

$$D^l = \frac{\partial^l}{\partial x_1^{l_1}\cdots\partial x_d^{l_d}},$$

and corresponding norm $\|\mathbf{h}\|_{H^k} = (\mathbf{h},\mathbf{h})_{H^k}^{1/2}$ for all $\mathbf{h} \in H^k$. Note that $H^m \subset H^k \subset H^0 = \mathcal{H}$ for $0 \le k \le m$. The Sobolev lemma (e.g. Kreiss and Lorenz (1989, Appendix 3, pp. 371–387)) says that if $\mathbf{h} \in H^k$ and $k \ge [\frac{d}{2}]+1$, where $[y]$ denotes the largest integer less than or equal to y, then $\mathbf{h}$ is a bounded function on D, with bound

$$\max_{\mathbf{x}\in D}\sum_{i=1}^{n}|h_i(\mathbf{x})| \le \alpha_k\|\mathbf{h}\|_{H^k}, \tag{15}$$

where the constant α_k depends on $L_1,\ldots,L_d$ but not on $\mathbf{h}$. It follows that if $\mathbf{h} \in H^k$ and $k \ge [\frac{d}{2}]+l+1$ for some positive integer l, then all of the l^{th}-order partial derivatives of $\mathbf{h}$ are bounded functions on D, with bound

$$\max_{\mathbf{x}\in D}\sum_{i=1}^{n}|D^l h_i(\mathbf{x})| \le \alpha_k\|\mathbf{h}\|_{H^k}, \tag{16}$$

and in particular, $\mathbf{h} \in C^{l-1}(D)$, since otherwise the l^{th}-order partial derivatives of $\mathbf{h}$ are not defined as bounded functions. Thus, for any nonnegative integer l, $H^k = H^k(D) \subset C^l(D)$ if $k \ge [\frac{d}{2}]+l+2$.

Suppose now that $\mathbf{s}_{t_0} \in H^k$ with $k \ge [\frac{d}{2}]+3$. According to the existence theory for linear and quasilinear symmetric hyperbolic systems (e.g. Courant

and Hilbert (1962, pp. 668–676)), there is a time interval $\mathcal{T}_* \subset \mathcal{T}_1$ for which Eq. (10) has a solution $\mathbf{s} \in H^k \cap C^1(\mathcal{T}_*)$ with $\mathbf{s}(\mathbf{x}, t_0) = \mathbf{s}_{t_0}(\mathbf{x})$ for all $\mathbf{x} \in D$, which is the classical solution since $H^k \subset C^1(D)$, and the solution remains in existence as long as $t \leq T_1$ and $\mathbf{s}(\cdot, t) \in H^k$. This is completely analogous to the situation for ordinary differential equations: the first time t such that the solution $\mathbf{s}(\cdot, t) \notin H^k$, if such a time is reached, is the first time the solution hits the "boundary" of H^k, $\|\mathbf{s}(\cdot, t)\|_{H^k} = \infty$. Typically the first partial derivatives of the classical solution become unbounded in finite time, even if $\mathbf{s}_{t_0} \in C^\infty(D)$ (e.g. Lax (1973, Theorem 6.1, p. 37)).

A minimum existence time for the solution $\mathbf{s} \in H^k \cap C^1(\mathcal{T}_*)$, $k \geq [\frac{d}{2}] + 3$, can be found in the following way. For any $\mathbf{s} \in H^k \cap C^1(\mathcal{T}_*)$,

$$\frac{d}{dt}\|\mathbf{s}\|^2_{H^k} = -2(\mathbf{s}, \mathbf{G}(\mathbf{s})\mathbf{s})_{H^k}$$

is continuous on $\mathcal{T}_*$, as in Eq. (11). An integration by parts using the symmetry of the matrices $\mathbf{A}_1, \ldots, \mathbf{A}_d$, along with the Sobolev inequalities Eqs. (15, 16), shows that there is a function $\phi \in C^1([0, \infty))$ such that

$$|(\mathbf{s}, \mathbf{G}(\mathbf{s})\mathbf{s})_{H^k}| \leq \phi(\|\mathbf{s}\|_{H^k})\|\mathbf{s}\|_{H^k};$$

see Kreiss and Lorenz (1989, pp. 190–196) for details. It follows that the solution $\mathbf{s}(\cdot, t)$ exists in H^k as long as $t \leq T_1$ and the solution $y(t)$ of the ordinary differential equation $dy/dt = \phi(y)$ with $y(t_0) = \|\mathbf{s}_{t_0}\|_{H^k}$ remains finite. Further, there is a time $T_2 > t_0$ depending continuously on $\|\mathbf{s}_{t_0}\|_{H^k}$, $T_2 = T_2(\|\mathbf{s}_{t_0}\|_{H^k}) \leq T_1$, for which $\|\mathbf{s}(\cdot, t)\|_{H^k}$ can be bounded in terms of $\|\mathbf{s}_{t_0}\|_{H^k}$, say

$$\|\mathbf{s}(\cdot, t)\|_{H^k} \leq \sqrt{2}\,\|\mathbf{s}_{t_0}\|_{H^k},$$

for all $t \in [t_0, T_2]$ (e.g. Kreiss and Lorenz (1989, Lemma 6.4.4, p. 196)). Then by the continuity of $T_2(\|\mathbf{s}_{t_0}\|_{H^k})$, it follows that if $\mathbf{s}_{t_0}$ is restricted to be in any bounded set in H^k, say if $\mathbf{s}_{t_0} \in H^k_E$ for some $E < \infty$, where

$$H^k_E = \{\mathbf{h} \in H^k : \|\mathbf{h}\|^2_{H^k} < E\},$$

then T_2 becomes independent of $\mathbf{s}_{t_0}$ (but depends on E), and the solution $\mathbf{s}$ corresponding to any $\mathbf{s}_{t_0} \in H^k_E$ satisfies

$$\|\mathbf{s}(\cdot, t)\|^2_{H^k} < 2E$$

for all $t \in [t_0, T_2]$. Also, since H^k_E is open as a set in H^k, and since $\|\mathbf{h}\| \leq \|\mathbf{h}\|_{H^k}$ for all $\mathbf{h} \in H^k$, H^k_E is open as a set in $L^2(D)$, and therefore $H^k_E \in \mathcal{B}(L^2(D))$.

3.2.2 The solution operator

Thus take $\mathcal{S} = H^k_E$ for any $E < \infty$ and $k \geq [\frac{d}{2}] + 3$, and take $\mathcal{T} = \mathcal{T}(\mathcal{S}) = [t_0, T_2]$. Then $\mathcal{S} \in \mathcal{B}(\mathcal{H}) = \mathcal{B}(L^2(D))$, and the unique classical solution $\mathbf{s}(\cdot, t)$ corresponding to each $\mathbf{s}_{t_0} \in \mathcal{S}$ exists in $H^k_{2E} \subset H^k \subset \mathcal{H} = L^2(D)$ for all $t \in \mathcal{T}$.

Denoting this solution by $\mathbf{s}_t = \mathbf{N}_{t,t_0}(\mathbf{s}_{t_0})$, it follows that $\mathbf{N}_{t,t_0}$ is defined uniquely on $\mathcal{S}$, as a map from $\mathcal{S}$ into $\mathcal{H}$, for all $t \in \mathcal{T}$. Further, since $\mathbf{s}_t \in H^k_{2E}$ for all $t \in \mathcal{T}$, it follows from the Sobolev inequalities that the function C_2 in Eq. (14) depends only on E and α_k, and therefore the map $\mathbf{N}_{t,t_0}$ is continuous in the norm $\|\cdot\|$, for all $t \in \mathcal{T}$. Note that $\mathcal{S} = H^k_E \subset \mathcal{H}_E$, where $\mathcal{H}_E$ was defined in Eq. (7), since $\|\mathbf{h}\| \leq \|\mathbf{h}\|_{H^k}$ for all $\mathbf{h} \in H^k$. It was necessary to define $\mathcal{S}$ as a bounded set in H^k, and therefore as a bounded set in $L^2(D)$.

The solution operator $\mathbf{N}_{t,t_0}$ is bounded not only as a map from $\mathcal{S}$ into $\mathcal{H}$, with the function C_1 in Eq. (13) now being a constant depending only on E and α_k, but also as a map from $\mathcal{S}$ into H^k, with

$$\|\mathbf{s}_t\|_{H^k} = \|\mathbf{N}_{t,t_0}(\mathbf{s}_{t_0})\|_{H^k} \leq \sqrt{2}\,\|\mathbf{s}_{t_0}\|_{H^k}$$

for all $t \in \mathcal{T}$. According to Eq. (11), the solution operator is conservative if the differential operator $\mathbf{G}$ is skew-symmetric,

$$(\mathbf{s}, \mathbf{G}(\mathbf{r})\mathbf{s}) = 0$$

for all $\mathbf{r}(\cdot,t), \mathbf{s}(\cdot,t) \in H^k$ and $t \in \mathcal{T}$. This conservation condition is met for an important class of symmetric hyperbolic systems (Lax (1973, p. 31)), but often a change of dependent variables which destroys symmetry of the matrices $\mathbf{A}_1, \ldots, \mathbf{A}_d$ is necessary to obtain conservation in $\mathcal{H} = L^2(D)$, as will be the case for the shallow-water equations.

It has been shown that, for each $\mathbf{s}_{t_0} \in \mathcal{S} = H^k_E$, with $E < \infty$, $k \geq [\frac{d}{2}] + l + 2$ and $l \geq 1$, the unique corresponding solution $\mathbf{s} = \mathbf{s}(\mathbf{x}, t)$ is of class $C^l(D) \cap C^1(\mathcal{T})$, for $\mathcal{T} = [t_0, T]$ and an appropriately defined T depending on α_k and E, and that $\|\mathbf{s}(\cdot,t)\|^2_{H^k} < 2E$ for all $t \in \mathcal{T}$. It is important to have a condition to guarantee further that $\mathbf{s} \in C^l(D \times \mathcal{T})$, particularly for the shallow-water example. To this end, denote by $L^2(D \times \mathcal{T})$ the Hilbert space of real, Lebesgue square-integrable n-vectors on $D \times \mathcal{T}$, with inner product

$$(\mathbf{g}, \mathbf{h})_{\mathcal{T}} = \int_{t_0}^{T} (\mathbf{g}, \mathbf{h})\, dt$$

for all $\mathbf{g}, \mathbf{h} \in L^2(D \times \mathcal{T})$, and corresponding norm $\|\mathbf{h}\|_{\mathcal{T}} = (\mathbf{h}, \mathbf{h})^{1/2}_{\mathcal{T}}$ for all $\mathbf{h} \in L^2(D \times \mathcal{T})$. Also, denote by $H^m(D \times \mathcal{T})$, for $m = 0, 1, \ldots$, the Sobolev space of real n-vectors on $D \times \mathcal{T}$ with m Lebesgue square-integrable mixed space and time partial derivatives on $D \times \mathcal{T}$, with the Sobolev inner product and norm. Thus, for any nonnegative integer l, $H^m(D \times \mathcal{T}) \subset C^l(D \times \mathcal{T})$ if $m \geq [\frac{d+1}{2}] + l + 2$. Now, the differential equations Eq. (10) can be used to express all mixed space-time partial derivatives of the solution up to any order m in terms of pure spatial partial derivatives up to order m. But

$$\int_{t_0}^{T} \|\mathbf{s}(\cdot,t)\|^2_{H^k}\, dt < 2E(T - t_0) < \infty$$

since $\|\mathbf{s}(\cdot,t)\|^2_{H^k} < 2E$ for all $t \in \mathcal{T}$, and therefore $\mathbf{s} \in H^k(D \times \mathcal{T})$. Thus, for each $\mathbf{s}_{t_0} \in \mathcal{S} = H^k_E$, with $E < \infty$, $k \geq [\frac{d+1}{2}] + l + 2$ and $l \geq 1$, the unique corresponding solution $\mathbf{s} = \mathbf{s}(\mathbf{x}, t)$ is of class $C^l(D \times \mathcal{T})$.

3.2.3 The stochastic initial-value problem

With $\mathcal{H} = L^2(D)$, $\mathcal{S} = H_E^k$, $E < \infty$, $k \geq [\frac{d}{2}] + l + 2$ and $l \geq 1$, let $\mathcal{T} = [t_0, T]$ and $\mathbf{N}_{t,t_0}$ be as defined in Section 3.2.2, let $t \in \mathcal{T}$, and suppose that $\mathbf{s}_{t_0}$ is an $\mathcal{S}$-valued random variable. Since $\mathcal{S} \subset \mathcal{H}_E$, it follows from the discussion of Section 2.4 that $\mathbf{s}_{t_0}$ is a second-order $\mathcal{S}$-valued random variable. Therefore by Theorem 1, $\mathbf{s}_t = \mathbf{N}_{t,t_0}(\mathbf{s}_{t_0})$ is a second-order $\mathcal{H}$-valued random variable, with mean $\overline{\mathbf{s}}_t \in \mathcal{H}$ and covariance operator $\mathcal{P}_t : \mathcal{H} \to \mathcal{H}$, which are related by

$$\|\overline{\mathbf{s}}_t\|^2 + \operatorname{tr} \mathcal{P}_t \leq e^{2C_1(t-t_0)}(\|\overline{\mathbf{s}}_{t_0}\|^2 + \operatorname{tr} \mathcal{P}_{t_0}),$$

where C_1 is the constant in Eq. (13). In fact,

$$\|\mathbf{s}_t(\omega)\|^2 \leq \|\mathbf{s}_t(\omega)\|_{H^k}^2 \leq 2\|\mathbf{s}_{t_0}(\omega)\|_{H^k}^2 < 2E$$

for all $\omega \in \Omega$, and so $\mathbf{s}_t$ is a second-order H^k-valued random variable with

$$\mathcal{E}\|\mathbf{s}_t\|^2 = \|\overline{\mathbf{s}}_t\|^2 + \operatorname{tr} \mathcal{P}_t \leq \mathcal{E}\|\mathbf{s}_t\|_{H^k}^2 \leq 2\mathcal{E}\|\mathbf{s}_{t_0}\|_{H^k}^2 < 2E.$$

The covariance operator $\mathcal{P}_t$ can be expressed in the following tangible way, which will lead also to a simple expression for $\operatorname{tr} \mathcal{P}_t$. Since $\mathcal{P}_t$ is a trace class operator, $\mathcal{P}_t$ is also a Hilbert-Schmidt operator. Since $\mathcal{H} = L^2(D)$ and $\mathcal{P}_t : \mathcal{H} \to \mathcal{H}$, it follows (e.g. Reed and Simon (1972, Theorem VI.23, p. 210)) that there is a real $n \times n$ matrix function $\mathbf{P}_t \in L^2(D \times D)$, called the covariance matrix of $\mathbf{s}_t$, such that

$$(\mathcal{P}_t \mathbf{h})(\mathbf{x}) = \int_D \mathbf{P}_t(\mathbf{x}, \mathbf{y}) \mathbf{h}(\mathbf{y})\, d\mathbf{y}$$

for all $\mathbf{h} \in \mathcal{H}$, where $d\mathbf{y} = dy_1 \cdots dy_d$, and moreover, that

$$\int_D \int_D \operatorname{tr} \mathbf{P}_t(\mathbf{x}, \mathbf{y}) \mathbf{P}_t^T(\mathbf{x}, \mathbf{y})\, d\mathbf{x}\, d\mathbf{y} = \sum_{i=1}^{\infty} \lambda_i^2(t) < \infty, \tag{17}$$

where $\operatorname{tr} \mathbf{A}$ denotes the trace of a matrix $\mathbf{A}$ and $\{\lambda_i(t)\}_{i=1}^{\infty}$ are the eigenvalues of the covariance operator $\mathcal{P}_t$. Thus

$$\mathcal{E}(\mathbf{g}, \mathbf{s}_t - \overline{\mathbf{s}}_t)(\mathbf{h}, \mathbf{s}_t - \overline{\mathbf{s}}_t) = (\mathbf{g}, \mathcal{P}_t \mathbf{h}) = \int_D \int_D \mathbf{g}^T(\mathbf{x}) \mathbf{P}_t(\mathbf{x}, \mathbf{y}) \mathbf{h}(\mathbf{y})\, d\mathbf{x}\, d\mathbf{y} \tag{18}$$

for all $\mathbf{g}, \mathbf{h} \in \mathcal{H}$. Since $\mathcal{P}_t$ is self-adjoint, the covariance matrix $\mathbf{P}_t$ has the symmetry property $\mathbf{P}_t^T(\mathbf{x}, \mathbf{y}) = \mathbf{P}_t(\mathbf{y}, \mathbf{x})$ for all $\mathbf{x}, \mathbf{y} \in D$.

Now let $\{\tilde{\mathbf{h}}_i(\cdot, t)\}_{i=1}^{\infty}$ denote the orthonormal eigenvectors (eigenfunctions) of $\mathcal{P}_t$ corresponding to the eigenvalues $\{\lambda_i(t)\}_{i=1}^{\infty}$,

$$\mathcal{P}_t \tilde{\mathbf{h}}_i(\cdot, t) = \lambda_i(t) \tilde{\mathbf{h}}_i(\cdot, t)$$

for $i = 1, 2, \ldots$. The eigenvalues are all nonnegative since the covariance operator is positive semidefinite, and the eigenvectors form an orthonormal basis for

$\mathcal{H}$ since the covariance operator is Hilbert-Schmidt. From Eq. (18) and the orthonormality of the eigenvectors, it follows that

$$\int_D \int_D \tilde{\mathbf{h}}_j^T(\mathbf{x},t)\mathbf{P}_t(\mathbf{x},\mathbf{y})\tilde{\mathbf{h}}_i(\mathbf{y},t)\,d\mathbf{x}\,d\mathbf{y} = (\tilde{\mathbf{h}}_j(\cdot,t), \mathcal{P}_t\tilde{\mathbf{h}}_i(\cdot,t)) = \lambda_i(t)\delta_{ij}$$

for $i,j = 1,2,\ldots$, where δ_{ij} is the Kronecker delta. Therefore, $\mathbf{P}_t$ has the representation

$$\mathbf{P}_t(\mathbf{x},\mathbf{y}) = \sum_{i=1}^{\infty} \lambda_i(t)\tilde{\mathbf{h}}_i(\mathbf{x},t)\tilde{\mathbf{h}}_i^T(\mathbf{y},t), \tag{19}$$

where the convergence is in $L^2(D \times D)$ as indicated in Eq. (17). Since the eigenvectors form an orthonormal basis for $\mathcal{H}$ and since $\mathcal{P}_t$ is trace class,

$$\operatorname{tr} \mathcal{P}_t = \sum_{i=1}^{\infty} \lambda_i(t) < \infty.$$

But according to Eq. (19),

$$\operatorname{tr} \mathbf{P}_t(\mathbf{x},\mathbf{x}) = \sum_{i=1}^{\infty} \lambda_i(t)\tilde{\mathbf{h}}_i^T(\mathbf{x},t)\tilde{\mathbf{h}}_i(\mathbf{x},t),$$

and therefore

$$\int_D \operatorname{tr} \mathbf{P}_t(\mathbf{x},\mathbf{x})\,d\mathbf{x} = \sum_{i=1}^{\infty} \lambda_i(t)$$

by the normality of the eigenvectors. Thus,

$$\operatorname{tr} \mathcal{P}_t = \int_D \operatorname{tr} \mathbf{P}_t(\mathbf{x},\mathbf{x})\,d\mathbf{x}. \tag{20}$$

Now recall that $\mathbf{s}_t$ is a second-order H^k-valued random variable. Therefore $\overline{\mathbf{s}}_t \in H^k$, $\mathcal{P}_t$ maps H^k into H^k, and

$$\mathcal{E}\|\mathbf{s}_t\|_{H^k}^2 = \|\overline{\mathbf{s}}_t\|_{H^k}^2 + \operatorname{tr} \mathcal{P}_t.$$

Also, $\overline{\mathbf{s}}_t \in C^l(D)$ since $H^k \subset C^l(D)$. Further, since $\mathcal{P}_t$ maps H^k into H^k, the eigenvectors of $\mathcal{P}_t$ form an orthonormal basis for H^k, and therefore they are all in $C^l(D)$.

Finally, let $\{\mathbf{h}_i\}_{i=1}^{\infty}$ be an orthonormal basis for H^k. Since $\mathbf{s}_{t_0}$ is an H_E^k-valued random variable, it follows from the discussion of Section 2.4 that

$$\mathbf{s}_{t_0}(\omega) = \sum_{i=1}^{\infty} (\mathbf{h}_i, \mathbf{s}_{t_0}(\omega))_{H^k}\mathbf{h}_i$$

for all $\omega \in \Omega$, with

$$\|\mathbf{s}_{t_0}(\omega)\|_{H^k}^2 = \sum_{i=1}^{\infty} (\mathbf{h}_i, \mathbf{s}_{t_0}(\omega))_{H^k}^2 < E.$$

for all $\omega \in \Omega$. It follows also that if $\{s_i\}_{i=1}^{\infty}$ is any collection of scalar random variables with

$$\sum_{i=1}^{\infty} s_i^2(\omega) < E$$

for all $\omega \in \Omega$, then $\sum_{i=1}^{\infty} s_i(\omega)\mathbf{h}_i$ is an H_E^k-valued random variable.

4 The shallow-water equations

The global nonlinear shallow-water equations written for the zonal and meridional velocity components and geopotential, u, v and Φ, respectively, are the system

$$\frac{\partial u}{\partial t} + \mathbf{V} \cdot \nabla u - (f + \frac{u}{a}\tan\phi)v + \frac{1}{a\cos\phi}\frac{\partial \Phi}{\partial \lambda} = 0$$

$$\frac{\partial v}{\partial t} + \mathbf{V} \cdot \nabla v + (f + \frac{u}{a}\tan\phi)u + \frac{1}{a}\frac{\partial \Phi}{\partial \phi} = 0$$

$$\frac{\partial \Phi}{\partial t} + \mathbf{V} \cdot \nabla \Phi + \Phi \nabla \cdot \mathbf{V} = 0,$$

where $\mathbf{V}$ is the wind vector, ϕ the latitude, λ the longitude, a the Earth radius, and f the Coriolis parameter. The change of variable $\Phi = w^2/4$ yields the symmetric hyperbolic system

$$\frac{\partial \mathbf{s}}{\partial t} + \left[\mathbf{A}\frac{1}{a\cos\phi}\frac{\partial}{\partial \lambda} + \mathbf{B}\frac{1}{a}\frac{\partial}{\partial \phi} + \mathbf{C}\right]\mathbf{s} = 0, \tag{21}$$

where $\mathbf{s} = (u, v, w)^T$,

$$\mathbf{A} = \begin{bmatrix} u & 0 & \frac{1}{2}w \\ 0 & u & 0 \\ \frac{1}{2}w & 0 & u \end{bmatrix},$$

$$\mathbf{B} = \begin{bmatrix} v & 0 & 0 \\ 0 & v & \frac{1}{2}w \\ 0 & \frac{1}{2}w & v \end{bmatrix},$$

$$\mathbf{C} = \begin{bmatrix} 0 & -(f + \frac{u}{a}\tan\phi) & 0 \\ f + \frac{u}{a}\tan\phi & 0 & 0 \\ 0 & -\frac{1}{2}\frac{w}{a}\tan\phi & 0 \end{bmatrix}.$$

Now let $\mathcal{H} = L^2(S)$, the Hilbert space of real square-integrable 3-vectors on the sphere of radius a, with inner product $(\cdot, \cdot)$ and corresponding norm $\|\cdot\|$. Appendix B.2 establishes a Sobolev-type lemma for the family of Hilbert spaces $\{\Phi_p = \Phi_p(S), p \geq 0\}$,

$$\Phi_p = \{\mathbf{h} \in L^2(S) : \|(I - \Delta)^p \mathbf{h}\| < \infty\},$$

where Δ is the Laplacian operator on the sphere, with inner product

$$(\mathbf{g}, \mathbf{h})_p = ((I - \Delta)^p \mathbf{g}, (I - \Delta)^p \mathbf{h})$$

for all $\mathbf{g}, \mathbf{h} \in \Phi_p$, and corresponding norm $\|\mathbf{h}\|_p = (\mathbf{h}, \mathbf{h})_p^{1/2}$ for all $\mathbf{h} \in \Phi_p$. Thus if $\mathbf{h} \in \Phi_p$ and p is a positive integer or half-integer, then all partial derivatives of the components of $\mathbf{h}$ up to order $2p$ are square-integrable. The spaces Φ_p are convenient for spherical geometry since the Laplacian operator is coordinate-free. The existence and uniqueness theory of Section 3.2 based on Sobolev spaces and inequalities carries over to the sphere using the spaces Φ_p for integers and half-integers p.

Tentatively let

$$\mathcal{S} = \{\mathbf{h} \in \Phi_2 : \|\mathbf{h}\|_2^2 < E\}$$

for some $E < \infty$. It follows from Appendix B.2 that

$$\mathcal{S} \subset \Phi_2 \subset C^1(S),$$

as expected from Section 3.2.1. It follows from Section 3.2.2 that for the symmetric hyperbolic system Eq. (21) there is a time interval $\mathcal{T} = \mathcal{T}(\mathcal{S}) = [t_0, T]$ such that, corresponding to each $\mathbf{s}_{t_0} \in \mathcal{S}$, there exists a unique classical solution $\mathbf{s}_t \in \Phi_2$ for all $t \in \mathcal{T}$, and further that $\mathbf{s} \in C^1(S \times \mathcal{T})$.

However, since $w = 2\sqrt{\Phi}$, this solution does not solve the original shallow-water system unless $w \geq 0$ on $S \times \mathcal{T}$. The differential equation for w is

$$\frac{\partial w}{\partial t} + \mathbf{V} \cdot \nabla w + \tfrac{1}{2} w \nabla \cdot \mathbf{V} = 0,$$

and therefore along the curves $\mathbf{x} = \mathbf{x}(t) = (\lambda(t), \phi(t))$ defined by

$$\frac{d\mathbf{x}}{dt} = \mathbf{V}(\mathbf{x}, t), \tag{22}$$

the solution w satisfies the ordinary differential equation

$$\frac{dw}{dt} + \tfrac{1}{2} w \nabla \cdot \mathbf{V} = 0.$$

This guarantees that if $w > 0$ initially, then $w > 0$ for all $t \in \mathcal{T}$. Thus redefine $\mathcal{S}$ as

$$\mathcal{S} = \{\mathbf{h} \in \Phi_2 : \|\mathbf{h}\|_2^2 < E\} \cap \{(u, v, w) \in \Phi_2 : w > 0\}.$$

Note that the latter set is open in $L^2(S)$ since it is open in Φ_2, and therefore $\mathcal{S} \in \mathcal{B}(L^2(S))$ since the intersection of two open sets is open. Also note that the initial-value problem for Eq. (22) is well-posed since $\mathbf{V} \in C^1(S \times \mathcal{T})$.

The classical solutions of the shallow-water equations satisfy the energy equation

$$\frac{\partial}{\partial t}[\Phi(u^2 + v^2) + \Phi^2] + \nabla \cdot \{[\Phi(u^2 + v^2) + 2\Phi^2]\mathbf{V}\} = 0.$$

This suggests introducing a new set of dependent variables, the energy variables $\mathbf{s} = (\alpha, \beta, \Phi)^T$ with $\alpha = u\Phi^{1/2}$ and $\beta = v\Phi^{1/2}$. In the energy variables, the physical total energy is just $\frac{1}{2}\|\mathbf{s}\|^2$, and it is conserved. It can be verified that in terms of the energy variables, the shallow-water system can be written as

$$\frac{\partial \mathbf{s}}{\partial t} + \mathbf{G}\mathbf{s} = 0, \tag{23}$$

where $\mathbf{G} = \mathbf{G}(\mathbf{s})$ has the form

$$\mathbf{G} = \mathbf{A}\frac{1}{a\cos\phi}\frac{\partial}{\partial\lambda} + \mathbf{B}\frac{1}{a}\frac{\partial}{\partial\phi} + \frac{1}{2}\frac{1}{a\cos\phi}\left(\frac{\partial\mathbf{A}}{\partial\lambda} + \frac{\partial\mathbf{B}\cos\phi}{\partial\phi}\right) + \mathbf{C},$$

with

$$\mathbf{A} = \begin{bmatrix} \alpha\Phi^{-1/2} & 0 & \frac{4}{5}\Phi^{1/2} \\ 0 & \alpha\Phi^{-1/2} & 0 \\ \frac{4}{5}\Phi^{1/2} & 0 & \frac{2}{5}\alpha\Phi^{-1/2} \end{bmatrix},$$

$$\mathbf{B} = \begin{bmatrix} \beta\Phi^{-1/2} & 0 & 0 \\ 0 & \beta\Phi^{-1/2} & \frac{4}{5}\Phi^{1/2} \\ 0 & \frac{4}{5}\Phi^{1/2} & \frac{2}{5}\beta\Phi^{-1/2} \end{bmatrix},$$

$$\mathbf{C} = \begin{bmatrix} 0 & -(f + \frac{1}{a}\alpha\Phi^{-1/2}\tan\phi) & 0 \\ f + \frac{1}{a}\alpha\Phi^{-1/2}\tan\phi & 0 & \frac{2}{5}\frac{1}{a}\Phi^{1/2}\tan\phi \\ 0 & -\frac{2}{5}\frac{1}{a}\Phi^{1/2}\tan\phi & 0 \end{bmatrix}.$$

For the system (23) to yield the solution of the original shallow-water system requires being able to recover $u = \alpha\Phi^{-1/2}$ and $v = \beta\Phi^{-1/2}$ from α, β and Φ. Now, products of scalars in Φ_2 are also scalars in Φ_2 since the elements of Φ_2 are all bounded, continuous functions. But $\Phi^{-1/2}$ is not in Φ_2 unless Φ is bounded from below by a positive constant. Thus for the energy variables, the initial space $\mathcal{S}$ is defined as $\mathcal{S} = \mathcal{S}_\gamma$, where

$$\mathcal{S}_\gamma = \{\mathbf{h} \in \Phi_2 : \|\mathbf{h}\|_2^2 < E\} \cap \{(\alpha, \beta, \Phi) \in \Phi_2 : \Phi > \gamma\}$$

for some constant $\gamma > 0$. It follows for the energy variables that for all $t \in \mathcal{T}$, the unique solution $\mathbf{s}_t$ corresponding to each $\mathbf{s}_{t_0} \in \mathcal{S}_\gamma$ is in $\mathcal{S}_\delta$ for some constant $\delta > 0$. The symmetry of the matrices $\mathbf{A}$ and $\mathbf{B}$, the skew-symmetry of the matrix $\mathbf{C}$, and the form of the differential operator $\mathbf{G}$ imply immediately that, for all $\delta > 0$, $\mathbf{G}$ is a skew-symmetric operator on $\mathcal{S}_\delta$:

$$(\mathbf{s}, \mathbf{G}(\mathbf{r})\mathbf{s}) = 0$$

for all $\mathbf{r}, \mathbf{s} \in \mathcal{S}_\delta$. This of course implies energy conservation, as noted in Section 3.2.2.

Now suppose for the energy variables that $\mathbf{s}_{t_0}$ is an $\mathcal{S}_\gamma$-valued random variable. Thus $\mathbf{s}_{t_0}$ is also second-order, and it follows from Theorem 1 that $\mathbf{s}_t = \mathbf{N}_{t,t_0}(\mathbf{s}_{t_0})$ is a second-order $L^2(S)$-valued random variable, with mean $\overline{\mathbf{s}}_t \in L^2(S)$ and covariance operator $\mathcal{P}_t : L^2(S) \to L^2(S)$, which are related by

$$\|\overline{\mathbf{s}}_t\|^2 + \operatorname{tr}\mathcal{P}_t = \|\overline{\mathbf{s}}_{t_0}\|^2 + \operatorname{tr}\mathcal{P}_{t_0} < E$$

for all $t \in \mathcal{T}$. The results of Section 3.2.3 show further that, for all $t \in \mathcal{T}$, $\overline{\mathbf{s}}_t \in \Phi_2 \subset C^1(S)$,

$$\mathcal{E}\|\mathbf{s}_t\|^2 = \|\overline{\mathbf{s}}_t\|^2 + \operatorname{tr}\mathcal{P}_t \le \mathcal{E}\|\mathbf{s}_t\|_2^2 < E,$$

and

$$\operatorname{tr} \mathcal{P}_t = \int_S \operatorname{tr} \mathbf{P}_t(\mathbf{x}, \mathbf{x})\, a \cos\phi \, d\phi \, d\lambda,$$

where $\mathbf{P}_t$ is the covariance matrix of $\mathbf{s}_t$.

The discussion at the end of Section 3.2.3 gives the general form of Φ_2-valued random variables with bounded Φ_2 norm. The Sobolev-type inequality Eq. (35) of Appendix B.2 then suggests a way of ensuring that such a random variable $\mathbf{s}_{t_0}$ is an $\mathcal{S}_\gamma$-valued random variable. Let

$$\Phi_{t_0}(\omega) = \overline{\Phi}_{t_0} + \Phi'_{t_0}(\omega)$$

for all $\omega \in \Omega$. The series expansions in Appendix B.2 give the form of every scalar in Φ_2. Suppose that $\overline{\Phi}_{t_0} \in \Phi_2$ with $\overline{\Phi}_{t_0} > \mu > \gamma > 0$. Equation (35) shows how to ensure that $|\Phi'_{t_0}(\omega)| < \mu - \gamma$ for all $\omega \in \Omega$, and therefore that $\Phi_{t_0}(\omega) > \gamma$ for all $\omega \in \Omega$.

A Random variables taking values in Hilbert space

Appendix A.1 defines Hilbert space-valued random variables and gives some of their main properties. Appendices A.2–A.4 give the definition, main properties and general construction, respectively, of Hilbert space-valued random variables of second order. Definitions of basic terms used in this appendix are provided in Appendix C. Further treatment of Hilbert space-valued random variables, and of random variables taking values in more general spaces, can be found in the books of Itô (1984) and Kallianpur and Xiong (1995).

Hilbert space-valued random variables, like scalar random variables, are defined with reference to some probability space $(\Omega, \mathcal{F}, P)$, with Ω the sample space, $\mathcal{F}$ the event space and P the probability measure. Thus throughout this appendix, a probability space $(\Omega, \mathcal{F}, P)$ is considered to be given. The expectation operator is denoted by $\mathcal{E}$. It is assumed that the given probability space is complete.

A real, separable Hilbert space $\mathcal{H}$ is also considered to be given. The inner product and corresponding norm on $\mathcal{H}$ are denoted by $(\cdot, \cdot)$ and $\|\cdot\|$, respectively. The Borel field generated by the open sets in $\mathcal{H}$ is denoted by $\mathcal{B}(\mathcal{H})$, i.e., $\mathcal{B}(\mathcal{H})$ is the smallest σ-algebra of sets in $\mathcal{H}$ that contains all the open sets in $\mathcal{H}$. Recall that every separable Hilbert space has a countable orthonormal basis, and that every orthonormal basis of a separable Hilbert space has the same number of elements $N \leq \infty$, the dimension of the space. For notational convenience it is assumed in this appendix that $\mathcal{H}$ is infinite-dimensional, with $\{\mathbf{h}_i\}_{i=1}^{\infty}$ denoting an orthonormal basis for $\mathcal{H}$. The results of this appendix hold just as well in the finite-dimensional case, by taking $\{\mathbf{h}_i\}_{i=1}^{N}$, $N < \infty$, as an orthonormal basis for $\mathcal{H}$, and by replacing infinite sums by finite ones.

A.1 $\mathcal{H}$-valued random variables

Recall that if X and Y are sets, $\mathbf{f}$ is a map from X into Y, and B is a subset of Y, then the set

$$\mathbf{f}^{-1}[B] = \{\mathbf{x} \in X : \mathbf{f}(\mathbf{x}) \in B\}$$

is called the inverse image of B (under $\mathbf{f}$). Recall also that the event space $\mathcal{F}$ of the probability space $(\Omega, \mathcal{F}, P)$ consists of the measurable subsets of Ω, which are called events.

Let $(Y, \mathcal{C})$ be a measurable space, i.e., Y is a set and $\mathcal{C}$ is a σ-algebra of subsets of Y. A map $\mathbf{f} : \Omega \to Y$ is called a $(Y, \mathcal{C})$-*valued random variable* if the inverse image of every set C in the collection $\mathcal{C}$ is an event, i.e., if $\mathbf{f}^{-1}[C] \in \mathcal{F}$ for every set $C \in \mathcal{C}$ (e.g. Itô (1984, p. 18), Kallianpur and Xiong (1995, p. 86); see also Reed and Simon (1972, p. 24)).

Thus an $(\mathcal{H}, \mathcal{B}(\mathcal{H}))$-valued random variable is a map $\mathbf{s} : \Omega \to \mathcal{H}$ such that

$$\{\omega \in \Omega : \mathbf{s}(\omega) \in B\} \in \mathcal{F}$$

for every set $B \in \mathcal{B}(\mathcal{H})$. Hereafter, an $(\mathcal{H}, \mathcal{B}(\mathcal{H}))$-valued random variable is called simply an $\mathcal{H}$-valued random variable, with the understanding that this always means an $(\mathcal{H}, \mathcal{B}(\mathcal{H}))$-valued random variable. An equivalent definition of $\mathcal{H}$-valued random variables, expressed in terms of scalar random variables, is given in Appendix A.2.

Let $\mathcal{S}$ be a nonempty set in $\mathcal{B}(\mathcal{H})$. It follows that the collection $\mathcal{B}_{\mathcal{S}}(\mathcal{H})$ of all sets in $\mathcal{B}(\mathcal{H})$ that are subsets of $\mathcal{S}$,

$$\mathcal{B}_{\mathcal{S}}(\mathcal{H}) = \{B \in \mathcal{B}(\mathcal{H}) : B \subset \mathcal{S}\},$$

is a σ-algebra of subsets of $\mathcal{S}$, namely, the collection of all sets C of the form $C = B \cap \mathcal{S}$ with $B \in \mathcal{B}(\mathcal{H})$. Hence $(\mathcal{S}, \mathcal{B}_{\mathcal{S}}(\mathcal{H}))$ is a measurable space, and an $(\mathcal{S}, \mathcal{B}_{\mathcal{S}}(\mathcal{H}))$-valued random variable is a map $\mathbf{s} : \Omega \to \mathcal{S}$ such that

$$\{\omega \in \Omega : \mathbf{s}(\omega) \in C\} \in \mathcal{F}$$

for every set $C \in \mathcal{B}_{\mathcal{S}}(\mathcal{H})$. Hereafter, an $(\mathcal{S}, \mathcal{B}_{\mathcal{S}}(\mathcal{H}))$-valued random variable is called simply an $\mathcal{S}$-valued random variable, with the understanding that this always means an $(\mathcal{S}, \mathcal{B}_{\mathcal{S}}(\mathcal{H}))$-valued random variable.

It follows by definition that every $\mathcal{S}$-valued random variable is an $\mathcal{H}$-valued random variable, for if $\mathbf{s} : \Omega \to \mathcal{S}$ and $\mathbf{s}^{-1}[C] \in \mathcal{F}$ for every set $C \in \mathcal{B}_{\mathcal{S}}(\mathcal{H})$, then $\mathbf{s}^{-1}[B] = \mathbf{s}^{-1}[B \cap \mathcal{S}] \in \mathcal{F}$ for every set $B \in \mathcal{B}(\mathcal{H})$. Also, every $\mathcal{H}$-valued random variable taking values only in $\mathcal{S}$ is an $\mathcal{S}$-valued random variable, for if $\mathbf{s} : \Omega \to \mathcal{S}$ and $\mathbf{s}^{-1}[B] \in \mathcal{F}$ for every set $B \in \mathcal{B}(\mathcal{H})$, then in particular $\mathbf{s}^{-1}[C] \in \mathcal{F}$ for every set $C \in \mathcal{B}_{\mathcal{S}}(\mathcal{H})$.

Finally, let $\mathbf{N}$ be a continuous map from $\mathcal{S}$ into $\mathcal{H}$. It follows that if $\mathbf{s}$ is an $\mathcal{S}$-valued random variable, then $\mathbf{N}(\mathbf{s})$ is an $\mathcal{H}$-valued random variable, i.e. that

$$\{\omega \in \Omega : \mathbf{N}(\mathbf{s}(\omega)) \in B\} \in \mathcal{F}$$

for every set $B \in \mathcal{B}(\mathcal{H})$. To see this, note first that

$$\{\omega \in \Omega : \mathbf{N}(\mathbf{s}(\omega)) \in B\} = \mathbf{s}^{-1}[\mathbf{N}^{-1}[B]],$$

and consider the class of sets E in $\mathcal{H}$ such that $\mathbf{N}^{-1}[E] \in \mathcal{B}_{\mathcal{S}}(\mathcal{H})$. It can be checked that this class of sets is a σ-algebra. Moreover, this class contains all the open sets in $\mathcal{H}$, because if O is an open set in $\mathcal{H}$ then $\mathbf{N}^{-1}[O]$ is also an open set in $\mathcal{H}$ by the continuity of $\mathbf{N}$ (e.g. Reed and Simon (1972, p. 8)) and so

$$C = \mathbf{N}^{-1}[O] = \mathbf{N}^{-1}[O] \cap \mathcal{S} \in \mathcal{B}_{\mathcal{S}}(\mathcal{H}).$$

But $\mathcal{B}(\mathcal{H})$ is the smallest σ-algebra containing all the open sets in $\mathcal{H}$, hence this class includes $\mathcal{B}(\mathcal{H})$, i.e., $\mathbf{N}^{-1}[B] \in \mathcal{B}_{\mathcal{S}}(\mathcal{H})$ for every set $B \in \mathcal{B}(\mathcal{H})$. If $\mathbf{s}$ is an $\mathcal{S}$-valued random variable then $\mathbf{s}^{-1}[C] \in \mathcal{F}$ for every set $C \in \mathcal{B}_{\mathcal{S}}(\mathcal{H})$, and therefore $\mathbf{s}^{-1}[\mathbf{N}^{-1}[B]] \in \mathcal{F}$ for every set $B \in \mathcal{B}(\mathcal{H})$, i.e., $\mathbf{N}(\mathbf{s})$ is an $\mathcal{H}$-valued random variable.

A.2 Second-order $\mathcal{H}$-valued random variables

If $\mathbf{s}$ is an $\mathcal{H}$-valued random variable and $\mathbf{h} \in \mathcal{H}$, then by the Schwarz inequality,

$$|(\mathbf{h}, \mathbf{s}(\omega))| \leq \|\mathbf{h}\| \, \|\mathbf{s}(\omega)\| < \infty \tag{24}$$

for all $\omega \in \Omega$, so for each fixed $\mathbf{h} \in \mathcal{H}$, the inner product $(\mathbf{h}, \mathbf{s})$ is a map from Ω into $\mathbb{R}$. In fact, it can be shown (e.g. Kallianpur and Xiong (1995, Corollary 3.1.1(b), p. 87)) that a map $\mathbf{s} : \Omega \to \mathcal{H}$ is an $\mathcal{H}$-valued random variable if, and only if, $(\mathbf{h}, \mathbf{s})$ is a scalar random variable for every $\mathbf{h} \in \mathcal{H}$. That is, a map $\mathbf{s} : \Omega \to \mathcal{H}$ is an $\mathcal{H}$-valued random variable if, and only if,

$$\{\omega \in \Omega : (\mathbf{h}, \mathbf{s}(\omega)) \leq \alpha\} \in \mathcal{F}$$

for every $\mathbf{h} \in \mathcal{H}$ and every $\alpha \in \mathbb{R}$.

It follows that if $\mathbf{s}$ is an $\mathcal{H}$-valued random variable, then $\|\mathbf{s}\|^2$ is a scalar random variable, that is,

$$\{\omega \in \Omega : \|\mathbf{s}(\omega)\|^2 \leq \alpha\} \in \mathcal{F}$$

for every $\alpha \in \mathbb{R}$. To see this, observe that if $\mathbf{s}$ is an $\mathcal{H}$-valued random variable, then $(\mathbf{h}_i, \mathbf{s})$ for $i = 1, 2, \ldots$ are scalar random variables, hence

$$s_n = \sum_{i=1}^{n} (\mathbf{h}_i, \mathbf{s})^2$$

are scalar random variables with $0 \leq s_n \leq s_{n+1}$ for $n = 1, 2, \ldots$, and by Parseval's relation,

$$\|\mathbf{s}(\omega)\|^2 = \sum_{i=1}^{\infty} (\mathbf{h}_i, \mathbf{s}(\omega))^2 = \lim_{n \to \infty} s_n(\omega)$$

for all $\omega \in \Omega$. Thus $\|\mathbf{s}\|^2$ is the limit of an increasing sequence of nonnegative scalar random variables, and is therefore a (nonnegative) scalar random variable.

If a map $\mathbf{s} : \Omega \to \mathcal{H}$ is an $\mathcal{H}$-valued random variable, then since $\|\mathbf{s}\|^2 \geq 0$ is a scalar random variable, it follows that $\mathcal{E}\|\mathbf{s}\|^2$ is defined and either $\mathcal{E}\|\mathbf{s}\|^2 = \infty$ or $\mathcal{E}\|\mathbf{s}\|^2 < \infty$. An $\mathcal{H}$-valued random variable $\mathbf{s}$ is called *second-order* if $\mathcal{E}\|\mathbf{s}\|^2 < \infty$.

A.3 Properties of second-order $\mathcal{H}$-valued random variables

In this subsection let $\mathbf{s} : \Omega \to \mathcal{H}$ be a second-order $\mathcal{H}$-valued random variable. Since $\mathcal{E}\|\mathbf{s}\|^2 < \infty$, it follows from Eq. (24) that

$$\mathcal{E}(\mathbf{h}, \mathbf{s})^2 \leq \|\mathbf{h}\|^2 \, \mathcal{E}\|\mathbf{s}\|^2 < \infty \tag{25}$$

for each $\mathbf{h} \in \mathcal{H}$. Thus, for each $\mathbf{h} \in \mathcal{H}$, $(\mathbf{h}, \mathbf{s})$ is a second-order scalar random variable, and therefore its mean is defined and finite. The mean of $(\mathbf{h}, \mathbf{s})$ will be denoted by

$$m[\mathbf{h}] = \mathcal{E}(\mathbf{h}, \mathbf{s}),$$

for each $\mathbf{h} \in \mathcal{H}$. Since $\mathcal{E}\|\mathbf{s}\|^2 < \infty$, $\|\mathbf{s}\|$ is a second-order scalar random variable, and its mean $M = \mathcal{E}\|\mathbf{s}\|$ satisfies $0 \leq M \leq \left(\mathcal{E}\|\mathbf{s}\|^2\right)^{1/2} < \infty$. Now

$$|m[\mathbf{h}]| = |\mathcal{E}(\mathbf{h}, \mathbf{s})| \leq \mathcal{E}\,|(\mathbf{h}, \mathbf{s})| \leq M\|\mathbf{h}\|$$

for each $\mathbf{h} \in \mathcal{H}$, by Eq. (24), and also

$$m[\alpha\mathbf{g} + \beta\mathbf{h}] = \alpha m[\mathbf{g}] + \beta m[\mathbf{h}]$$

for each $\mathbf{g}, \mathbf{h} \in \mathcal{H}$ and $\alpha, \beta \in \mathbb{R}$. Thus $m[\cdot]$ is a bounded linear functional on $\mathcal{H}$, and by the Riesz representation theorem for Hilbert space (e.g. Royden (1968, p. 213), Reed and Simon (1972, p. 43)) this implies that there exists a unique element $\overline{\mathbf{s}} \in \mathcal{H}$, called the *mean* of $\mathbf{s}$, such that

$$m[\mathbf{h}] = (\mathbf{h}, \overline{\mathbf{s}})$$

for each $\mathbf{h} \in \mathcal{H}$. Thus the mean $\overline{\mathbf{s}}$ of $\mathbf{s}$ is defined uniquely in $\mathcal{H}$, and satisfies $(\mathbf{h}, \overline{\mathbf{s}}) = \mathcal{E}(\mathbf{h}, \mathbf{s})$ for every $\mathbf{h} \in \mathcal{H}$.

Now let $\mathbf{s}'(\omega) = \mathbf{s}(\omega) - \overline{\mathbf{s}}$ for each $\omega \in \Omega$. Since

$$\|\mathbf{s}'(\omega)\| \leq \|\mathbf{s}(\omega)\| + \|\overline{\mathbf{s}}\| < \infty$$

for each $\omega \in \Omega$, $\mathbf{s}' = \mathbf{s} - \overline{\mathbf{s}}$ is a map from Ω into $\mathcal{H}$. Furthermore, for every $\mathbf{h} \in \mathcal{H}$, $(\mathbf{h}, \mathbf{s})$ is a scalar random variable, $|(\mathbf{h}, \overline{\mathbf{s}})| < \infty$, and $|(\mathbf{h}, \mathbf{s}(\omega))| < \infty$ for each $\omega \in \Omega$. Therefore $(\mathbf{h}, \mathbf{s}') = (\mathbf{h}, \mathbf{s}) - (\mathbf{h}, \overline{\mathbf{s}})$ is a scalar random variable for every $\mathbf{h} \in \mathcal{H}$, and hence $\mathbf{s}'$ is an $\mathcal{H}$-valued random variable. Also,

$$\mathcal{E}(\mathbf{h}, \mathbf{s}') = \mathcal{E}(\mathbf{h}, \mathbf{s}) - (\mathbf{h}, \overline{\mathbf{s}}) = 0$$

for every $\mathbf{h} \in \mathcal{H}$, so the mean of $\mathbf{s}'$ is $\mathbf{0} \in \mathcal{H}$. Thus

$$\mathcal{E}\|\mathbf{s}\|^2 = \mathcal{E}(\overline{\mathbf{s}} + \mathbf{s}', \overline{\mathbf{s}} + \mathbf{s}') = \|\overline{\mathbf{s}}\|^2 + \mathcal{E}\|\mathbf{s}'\|^2 \tag{26}$$

and, in particular, $\mathcal{E}\|\mathbf{s}'\|^2 \leq \mathcal{E}\|\mathbf{s}\|^2 < \infty$. Therefore $\mathbf{s}' : \Omega \to \mathcal{H}$ is a second-order $\mathcal{H}$-valued random variable, and $\|\mathbf{s}'\|$ is a second-order scalar random variable.

Since $\mathbf{s}'$ is a second-order $\mathcal{H}$-valued random variable, $(\mathbf{g}, \mathbf{s}')$ and $(\mathbf{h}, \mathbf{s}')$ are second-order scalar random variables, for each $\mathbf{g}, \mathbf{h} \in \mathcal{H}$. Therefore the expectation

$$C[\mathbf{g}, \mathbf{h}] = \mathcal{E}(\mathbf{g}, \mathbf{s}')(\mathbf{h}, \mathbf{s}')$$

is defined for all $\mathbf{g}, \mathbf{h} \in \mathcal{H}$, and in fact

$$|C[\mathbf{g}, \mathbf{h}]| \leq \mathcal{E}\,|(\mathbf{g}, \mathbf{s}')(\mathbf{h}, \mathbf{s}')| \leq \left[\mathcal{E}(\mathbf{g}, \mathbf{s}')^2\right]^{1/2} \left[\mathcal{E}(\mathbf{h}, \mathbf{s}')^2\right]^{1/2} \leq \|\mathbf{g}\|\,\|\mathbf{h}\|\,\mathcal{E}\|\mathbf{s}'\|^2.$$

The functional C, called the *covariance functional* of $\mathbf{s}$, is also linear in its two arguments. Thus $C[\cdot, \cdot]$ is a bounded bilinear functional on $\mathcal{H} \times \mathcal{H}$. It follows (e.g. Rudin (1991, Theorem 12.8, p. 310)) that there exists a unique bounded linear operator $\mathcal{P} : \mathcal{H} \to \mathcal{H}$, called the *covariance operator* of $\mathbf{s}$, such that

$$C[\mathbf{g}, \mathbf{h}] = (\mathbf{g}, \mathcal{P}\mathbf{h})$$

for each $\mathbf{g}, \mathbf{h} \in \mathcal{H}$. The covariance operator $\mathcal{P}$ is *self-adjoint*, i.e., $(\mathcal{P}\mathbf{g}, \mathbf{h}) = (\mathbf{g}, \mathcal{P}\mathbf{h})$ for all $\mathbf{g}, \mathbf{h} \in \mathcal{H}$, since the covariance functional is symmetric, $C[\mathbf{h}, \mathbf{g}] = C[\mathbf{g}, \mathbf{h}]$ for all $\mathbf{g}, \mathbf{h} \in \mathcal{H}$. The covariance operator is also *positive semidefinite*, i.e., $(\mathbf{h}, \mathcal{P}\mathbf{h}) \geq 0$ for all $\mathbf{h} \in \mathcal{H}$, since

$$(\mathbf{h}, \mathcal{P}\mathbf{h}) = C[\mathbf{h}, \mathbf{h}] = \mathcal{E}(\mathbf{h}, \mathbf{s}')^2 \geq 0$$

for all $\mathbf{h} \in \mathcal{H}$.

Now consider the second-order scalar random variable $\|\mathbf{s}'\|$. By Parseval's relation,

$$\|\mathbf{s}'(\omega)\|^2 = \sum_{i=1}^{\infty} (\mathbf{h}_i, \mathbf{s}'(\omega))^2$$

for all $\omega \in \Omega$, and therefore

$$\mathcal{E}\|\mathbf{s}'\|^2 = \sum_{i=1}^{\infty} \mathcal{E}(\mathbf{h}_i, \mathbf{s}')^2,$$

because $\{(\mathbf{h}_i, \mathbf{s}')^2\}_{i=1}^{\infty}$ is a sequence of nonnegative random variables. Furthermore,

$$\mathcal{E}(\mathbf{h}_i, \mathbf{s}')^2 = (\mathbf{h}_i, \mathcal{P}\mathbf{h}_i) \tag{27}$$

for $i = 1, 2, \ldots$, by definition of the covariance operator $\mathcal{P}$, and therefore

$$\mathcal{E}\|\mathbf{s}'\|^2 = \sum_{i=1}^{\infty} (\mathbf{h}_i, \mathcal{P}\mathbf{h}_i).$$

The summation on the right-hand side, called the *trace* of $\mathcal{P}$ and written $\operatorname{tr}\mathcal{P}$, is independent of the choice of orthonormal basis $\{\mathbf{h}_i\}_{i=1}^{\infty}$ for $\mathcal{H}$, for any positive semidefinite bounded linear operator from $\mathcal{H}$ into $\mathcal{H}$ (e.g. Reed and Simon (1972, Theorem VI.18, p. 206)). Thus

$$\operatorname{tr}\mathcal{P} = \sum_{i=1}^{\infty} (\mathbf{h}_i, \mathcal{P}\mathbf{h}_i) = \mathcal{E}\|\mathbf{s}'\|^2 < \infty,$$

and Eq. (26) can be written as

$$\mathcal{E}\|\mathbf{s}\|^2 = \|\overline{\mathbf{s}}\|^2 + \operatorname{tr}\mathcal{P}, \tag{28}$$

which is a generalization of Eq. (42) to second-order $\mathcal{H}$-valued random variables.

Since $\operatorname{tr}\mathcal{P} < \infty$, $\mathcal{P}$ is a *trace class* operator, and therefore also a *compact* operator (e.g. Reed and Simon (1972, Theorem VI.21, p. 209)). Since $\mathcal{P}$ is self-adjoint in addition to being compact, it follows from the Hilbert-Schmidt theorem (e.g. Reed and Simon (1972, Theorem VI.16, p. 203)) that there exists an orthonormal basis for $\mathcal{H}$ which consists of *eigenvectors* $\{\tilde{\mathbf{h}}_i\}_{i=1}^{\infty}$ of $\mathcal{P}$,

$$\mathcal{P}\tilde{\mathbf{h}}_i = \lambda_i \tilde{\mathbf{h}}_i$$

for $i = 1, 2, \ldots$, where the corresponding *eigenvalues* $\lambda_i = (\tilde{\mathbf{h}}_i, \mathcal{P}\tilde{\mathbf{h}}_i)$ for $i = 1, 2, \ldots$ are all real numbers and satisfy $\lambda_i \to 0$ as $i \to \infty$. In fact, the eigenvalues are all nonnegative since $\mathcal{P}$ is positive semidefinite, and therefore $\lambda_i = \|\mathcal{P}\tilde{\mathbf{h}}_i\|$ for $i = 1, 2, \ldots$. Further, it follows from Eq. (27) that

$$\lambda_i = (\tilde{\mathbf{h}}_i, \mathcal{P}\tilde{\mathbf{h}}_i) = \mathcal{E}(\tilde{\mathbf{h}}_i, \mathbf{s}')^2 = \sigma_i^2,$$

where σ_i^2 is the variance of the scalar random variable $(\tilde{\mathbf{h}}_i, \mathbf{s})$, for $i = 1, 2, \ldots$. By the definition of $\operatorname{tr}\mathcal{P}$,

$$\mathcal{E}\|\mathbf{s}'\|^2 = \operatorname{tr}\mathcal{P} = \sum_{i=1}^{\infty}(\tilde{\mathbf{h}}_i, \mathcal{P}\tilde{\mathbf{h}}_i) = \sum_{i=1}^{\infty}\lambda_i < \infty.$$

Thus the eigenvalues $\{\lambda_i\}_{i=1}^{\infty}$ of $\mathcal{P}$ are the variances $\{\sigma_i^2\}_{i=1}^{\infty}$ and have finite sum $\operatorname{tr}\mathcal{P}$. Equation (28) can then be rewritten as

$$\mathcal{E}\|\mathbf{s}\|^2 = \|\overline{\mathbf{s}}\|^2 + \sum_{i=1}^{\infty}\sigma_i^2, \tag{29}$$

which is another generalization of Eq. (42).

Since every $\mathbf{h} \in \mathcal{H}$ has the representation $\mathbf{h} = \sum_{i=1}^{\infty}(\mathbf{h}_i, \mathbf{h})\mathbf{h}_i$, and since $\mathcal{P}\mathbf{h} \in \mathcal{H}$ for every $\mathbf{h} \in \mathcal{H}$, taking $\mathbf{h}_i = \tilde{\mathbf{h}}_i$ and using the fact that

$$(\tilde{\mathbf{h}}_i, \mathcal{P}\mathbf{h}) = (\mathcal{P}\tilde{\mathbf{h}}_i, \mathbf{h}) = \lambda_i(\tilde{\mathbf{h}}_i, \mathbf{h}) = \sigma_i^2(\tilde{\mathbf{h}}_i, \mathbf{h})$$

for $i = 1, 2, \ldots$, gives the following representation for $\mathcal{P}$:

$$\mathcal{P}\mathbf{h} = \sum_{i=1}^{\infty}\sigma_i^2(\tilde{\mathbf{h}}_i, \mathbf{h})\tilde{\mathbf{h}}_i \tag{30}$$

for every $\mathbf{h} \in \mathcal{H}$. Thus the expectation $\mathcal{E}(\mathbf{g}, \mathbf{s}')(\mathbf{h}, \mathbf{s}')$ is given by the convergent series

$$\mathcal{E}(\mathbf{g}, \mathbf{s}')(\mathbf{h}, \mathbf{s}') = C[\mathbf{g}, \mathbf{h}] = (\mathbf{g}, \mathcal{P}\mathbf{h}) = \sum_{i=1}^{\infty}\sigma_i^2(\tilde{\mathbf{h}}_i, \mathbf{g})(\tilde{\mathbf{h}}_i, \mathbf{h}),$$

for every $\mathbf{g}, \mathbf{h} \in \mathcal{H}$.

Finally, since $\mathcal{P}$ is a positive semidefinite bounded linear operator from $\mathcal{H}$ into $\mathcal{H}$, there exists a unique positive semidefinite bounded linear operator $\mathcal{P}^{1/2}$:

$\mathcal{H} \to \mathcal{H}$, called the *square root* of $\mathcal{P}$, that satisfies $\left(\mathcal{P}^{1/2}\right)^2 = \mathcal{P}$ (e.g. Reed and Simon (1972, Theorem VI.9, p. 196)). Since $\mathcal{P}$ is also self-adjoint and trace class, $\mathcal{P}^{1/2}$ is self-adjoint and *Hilbert-Schmidt* (e.g. Reed and Simon (1972, p. 210)), with the same eigenvectors as $\mathcal{P}$ and with eigenvalues that are the nonnegative square roots of the corresponding eigenvalues of $\mathcal{P}$. That is,

$$\mathcal{P}^{1/2}\tilde{\mathbf{h}}_i = \sigma_i \tilde{\mathbf{h}}_i,$$

where $\sigma_i = \lambda_i^{1/2} = [\mathcal{E}(\tilde{\mathbf{h}}_i, \mathbf{s}')^2]^{1/2}$, for $i = 1, 2, \ldots$. Therefore $\sigma_i = (\tilde{\mathbf{h}}_i, \mathcal{P}^{1/2}\tilde{\mathbf{h}}_i) = \|\mathcal{P}^{1/2}\tilde{\mathbf{h}}_i\|$ for $i = 1, 2, \ldots$, and $\mathcal{P}^{1/2}$ has the representation

$$\mathcal{P}^{1/2}\mathbf{h} = \sum_{i=1}^{\infty} \sigma_i (\tilde{\mathbf{h}}_i, \mathbf{h}) \tilde{\mathbf{h}}_i \tag{31}$$

for every $\mathbf{h} \in \mathcal{H}$.

A.4 Construction of second-order $\mathcal{H}$-valued random variables

It will now be shown how essentially all second-order $\mathcal{H}$-valued random variables can be constructed. This will be accomplished by first reconsidering, in a suggestive notation, the defining properties of every second-order $\mathcal{H}$-valued random variable. The construction given here is by Itô's regularization theorem (Itô (1984, Theorem 2.3.3, p. 27), Kallianpur and Xiong (1995, Theorem 3.1.2, p. 87)) applied to $\mathcal{H}$, and amounts to formalizing on $\mathcal{H}$ the usual construction of infinite-dimensional random variables through random Fourier series.

For the moment, fix a second-order $\mathcal{H}$-valued random variable $\mathbf{s}$, and consider the behavior of

$$s[\mathbf{h}] = (\mathbf{h}, \mathbf{s})$$

as a functional of $\mathbf{h} \in \mathcal{H}$, that is, as $\mathbf{h}$ varies throughout $\mathcal{H}$. The functional $s[\cdot]$ has three important properties. First, on evaluation at any $\mathbf{h} \in \mathcal{H}$, it is a scalar random variable, with

$$s[\mathbf{h}](\omega) = (\mathbf{h}, \mathbf{s}(\omega))$$

for each $\omega \in \Omega$, since $\mathbf{s} : \Omega \to \mathcal{H}$ is an $\mathcal{H}$-valued random variable. Thus $s[\cdot]$ is a map from $\mathcal{H}$ into the set of scalar random variables on $(\Omega, \mathcal{F}, P)$. Second, this map is linear,

$$s[\alpha\mathbf{g} + \beta\mathbf{h}] = \alpha s[\mathbf{g}] + \beta s[\mathbf{h}]$$

for all $\mathbf{g}, \mathbf{h} \in \mathcal{H}$ and $\alpha, \beta \in \mathbb{R}$, by linearity of the inner product. Third, according to Eq. (25),

$$(\mathcal{E}s^2[\mathbf{h}])^{1/2} \leq \gamma \|\mathbf{h}\|, \tag{32}$$

where

$$\gamma = (\mathcal{E}\|\mathbf{s}\|^2)^{1/2} < \infty,$$

since the $\mathcal{H}$-valued random variable $\mathbf{s}$ is second-order. Thus $s[\cdot]$ is a linear map from $\mathcal{H}$ into the set of second-order scalar random variables on $(\Omega, \mathcal{F}, P)$.

Now recall the space $L^2(\Omega, \mathcal{F}, P)$, whose elements are the equivalence classes of second-order scalar random variables, where two scalar random variables are called equivalent if they are equal wp1. The space $L^2(\Omega, \mathcal{F}, P)$ is a Hilbert space, with the inner product of any two elements $\tilde{r}, \tilde{s} \in L^2(\Omega, \mathcal{F}, P)$ given by $\mathcal{E}\tilde{r}\tilde{s}$ and the corresponding norm of any element $\tilde{s} \in L^2(\Omega, \mathcal{F}, P)$ given by $(\mathcal{E}\tilde{s}^2)^{1/2}$. Inequality (32) states that the functional $s[\cdot]$ is bounded, when viewed as a map from $\mathcal{H}$ into $L^2(\Omega, \mathcal{F}, P)$.

A map $s[\cdot]$ from $\mathcal{H}$ into the set of scalar random variables on $(\Omega, \mathcal{F}, P)$, which is linear in the sense that if $\mathbf{g}, \mathbf{h} \in \mathcal{H}$ and $\alpha, \beta \in \mathbb{R}$ then

$$s[\alpha\mathbf{g} + \beta\mathbf{h}] = \alpha s[\mathbf{g}] + \beta s[\mathbf{h}] \text{ wp1},$$

is called a *random linear functional* (e.g. Itô (1984, p. 22), Omatu and Seinfeld (1989, p. 48)). Observe that the set of $\omega \in \Omega$ of probability measure zero where linearity fails to hold can depend on $\alpha, \beta, \mathbf{g}$ and $\mathbf{h}$. If linearity holds for all $\omega \in \Omega$, for all $\mathbf{g}, \mathbf{h} \in \mathcal{H}$ and $\alpha, \beta \in \mathbb{R}$, then the random linear functional is called *perfect*. If $s[\cdot]$ is a random linear functional and there is a constant $\gamma \in \mathbb{R}$ such that Eq. (32) holds for all $\mathbf{h} \in \mathcal{H}$, then the random linear functional is called *second-order*. Thus, given any particular $\mathcal{H}$-valued random variable $\mathbf{s}$, the map $s[\cdot]$ defined for all $\mathbf{h} \in \mathcal{H}$ by $s[\mathbf{h}] = (\mathbf{h}, \mathbf{s})$ is a perfect random linear functional, and if $\mathbf{s}$ is second-order then so is $s[\cdot]$.

Now it will be shown that a random linear functional $s[\cdot]$ is second-order if, and only if,

$$\sum_{i=1}^{\infty} \mathcal{E}s^2[\mathbf{h}_i] < \infty. \tag{33}$$

In particular, a collection $\{s_i\}_{i=1}^{\infty}$ of scalar random variables with $\sum_{i=1}^{\infty} \mathcal{E}s_i^2 < \infty$ can be used to define a second-order random linear functional, by setting $s[\mathbf{h}_i] = s_i$ for $i = 1, 2, \ldots$. It will then be shown how to construct, from any given second-order random linear functional $s[\cdot]$, a second-order $\mathcal{H}$-valued random variable $\mathbf{s}$ such that, for all $\mathbf{h} \in \mathcal{H}$,

$$(\mathbf{h}, \mathbf{s}) = s[\mathbf{h}] \text{ wp1}.$$

Such an $\mathcal{H}$-valued random variable $\mathbf{s}$ is called a *regularized version* of the random linear functional $s[\cdot]$ (Itô (1984, Definition 2.3.2, p. 23)).

Let $s[\cdot]$ be a second-order random linear functional. Given any $\mathbf{h} \in \mathcal{H}$ and positive integer n, it follows from the linearity of $s[\cdot]$ that

$$s\left[\sum_{i=1}^{n} (\mathbf{h}_i, \mathbf{h})\mathbf{h}_i\right] = \sum_{i=1}^{n} (\mathbf{h}_i, \mathbf{h}) s[\mathbf{h}_i] \text{ wp1},$$

where the set of probability measure zero on which equality does not hold may depend on $\mathbf{h}$ and on the orthonormal basis elements $\{\mathbf{h}_i\}_{i=1}^{n}$. By the boundedness of $s[\cdot]$ it follows that

$$\mathcal{E}\left(\sum_{i=1}^{n} (\mathbf{h}_i, \mathbf{h}) s[\mathbf{h}_i]\right)^2 \leq \gamma^2 \left\| \sum_{i=1}^{n} (\mathbf{h}_i, \mathbf{h})\mathbf{h}_i \right\|^2,$$

for some constant $\gamma \in \mathbb{R}$ which is independent of $\mathbf{h}$ and n. Taking the limit as $n \to \infty$ gives

$$\mathcal{E}\left(\sum_{i=1}^{\infty}(\mathbf{h}_i, \mathbf{h})s[\mathbf{h}_i]\right)^2 \leq \gamma^2\|\mathbf{h}\|^2 < \infty,$$

for all $\mathbf{h} \in \mathcal{H}$. Thus the series $\sum_{i=1}^{\infty}(\mathbf{h}_i, \mathbf{h})s[\mathbf{h}_i]$ converges in $L^2(\Omega, \mathcal{F}, P)$, i.e., there exists a unique element $\tilde{s}[\mathbf{h}] \in L^2(\Omega, \mathcal{F}, P)$ such that

$$\lim_{n\to\infty} \mathcal{E}\left(\tilde{s}[\mathbf{h}] - \sum_{i=1}^{n}(\mathbf{h}_i, \mathbf{h})s[\mathbf{h}_i]\right)^2 = 0,$$

for all $\mathbf{h} \in \mathcal{H}$. Equivalently, since a series converges in a Hilbert space if, and only if, it converges in norm,

$$\sum_{i=1}^{\infty}\left(\mathcal{E}\left\{(\mathbf{h}_i, \mathbf{h})s[\mathbf{h}_i]\right\}^2\right)^{1/2} = \sum_{i=1}^{\infty}|(\mathbf{h}_i, \mathbf{h})|\left(\mathcal{E}s^2[\mathbf{h}_i]\right)^{1/2} < \infty,$$

for all $\mathbf{h} \in \mathcal{H}$. By the Riesz representation theorem applied to the Hilbert space of square-summable sequences of real numbers, and since

$$\sum_{i=1}^{\infty}(\mathbf{h}_i, \mathbf{h})^2 = \|\mathbf{h}\|^2$$

by Parseval's relation, the series $\sum_{i=1}^{\infty}(\mathbf{h}_i, \mathbf{h})s[\mathbf{h}_i]$ therefore converges in $L^2(\Omega, \mathcal{F}, P)$, for all $\mathbf{h} \in \mathcal{H}$, if, and only if, Eq. (33) holds, in which case

$$\sum_{i=1}^{\infty}|(\mathbf{h}_i, \mathbf{h})|\left(\mathcal{E}s^2[\mathbf{h}_i]\right)^{1/2} \leq \|\mathbf{h}\|\left[\sum_{i=1}^{\infty}\mathcal{E}s^2[\mathbf{h}_i]\right]^{1/2} < \infty,$$

by the Schwarz inequality. Thus, if $s[\cdot]$ is a second-order random linear functional, then Eq. (33) holds, for every orthonormal basis $\{\mathbf{h}_i\}_{i=1}^{\infty}$ of $\mathcal{H}$.

Conversely, suppose that Eq. (33) holds for a random linear functional $s[\cdot]$, for some orthonormal basis $\{\mathbf{h}_i\}_{i=1}^{\infty}$ of $\mathcal{H}$. Since every $\mathbf{h} \in \mathcal{H}$ has the representation $\mathbf{h} = \sum_{i=1}^{\infty}(\mathbf{h}_i, \mathbf{h})\mathbf{h}_i$, it follows from the linearity of $s[\cdot]$ that if $\mathbf{h} \in \mathcal{H}$ then

$$s[\mathbf{h}] = \sum_{i=1}^{\infty}(\mathbf{h}_i, \mathbf{h})s[\mathbf{h}_i] \ \text{ wp1},$$

and therefore

$$s^2[\mathbf{h}] \leq \|\mathbf{h}\|^2\sum_{i=1}^{\infty}s^2[\mathbf{h}_i] \ \text{ wp1},$$

by the Schwarz inequality and Parseval's relation. Thus

$$\mathcal{E}s^2[\mathbf{h}] \leq \|\mathbf{h}\|^2\sum_{i=1}^{\infty}\mathcal{E}s^2[\mathbf{h}_i],$$

for every $\mathbf{h} \in \mathcal{H}$, i.e., Eq. (32) holds with

$$\gamma^2 = \sum_{i=1}^{\infty} \mathcal{E} s^2[\mathbf{h}_i] < \infty,$$

by Eq. (33), and therefore $s[\cdot]$ is a second-order random linear functional. Furthermore, since $s[\cdot]$ is a second-order random linear functional, Eq. (33) holds for every orthonormal basis $\{\mathbf{h}_i\}_{i=1}^{\infty}$ of $\mathcal{H}$.

Now let $s[\cdot]$ be a given second-order random linear functional. Since

$$\mathcal{E} \sum_{i=1}^{\infty} s^2[\mathbf{h}_i] = \sum_{i=1}^{\infty} \mathcal{E} s^2[\mathbf{h}_i] < \infty,$$

the sum $\sum_{i=1}^{\infty} s^2[\mathbf{h}_i]$ must be finite wp1, i.e., if

$$E = \{\omega \in \Omega : \sum_{i=1}^{\infty} s^2[\mathbf{h}_i](\omega) < \infty\}$$

then $E \in \mathcal{F}$ and $P(E) = 1$, where the set E may depend on $\{\mathbf{h}_i\}_{i=1}^{\infty}$. Define $\mathbf{s}(\omega)$ for each $\omega \in \Omega$ by

$$\mathbf{s}(\omega) = \begin{cases} \sum_{i=1}^{\infty} \mathbf{h}_i s[\mathbf{h}_i](\omega) & \text{if } \omega \in E \\ 0 & \text{if } \omega \notin E \end{cases}.$$

By Parseval's relation it follows that

$$\|\mathbf{s}(\omega)\|^2 = \begin{cases} \sum_{i=1}^{\infty} s^2[\mathbf{h}_i](\omega) & \text{if } \omega \in E \\ 0 & \text{if } \omega \notin E \end{cases},$$

and therefore $\|\mathbf{s}(\omega)\|^2 < \infty$ for all $\omega \in \Omega$. Thus $\mathbf{s}$ is a map from Ω into $\mathcal{H}$, and for any $\mathbf{h} \in \mathcal{H}$,

$$(\mathbf{h}, \mathbf{s}(\omega)) = \sum_{i=1}^{\infty} (\mathbf{h}_i, \mathbf{h}) s[\mathbf{h}_i](\omega) \tag{34}$$

for each $\omega \in E$. Now, if $\mathbf{h} \in \mathcal{H}$ then

$$s[\mathbf{h}] = \sum_{i=1}^{\infty} (\mathbf{h}_i, \mathbf{h}) s[\mathbf{h}_i] \text{ wp1},$$

and so there is a set $E_{\mathbf{h}} \in \mathcal{F}$ with $P(E_{\mathbf{h}}) = 1$, that may depend on $\{\mathbf{h}_i\}_{i=1}^{\infty}$ as well as on $\mathbf{h}$, such that

$$s[\mathbf{h}](\omega) = \sum_{i=1}^{\infty} (\mathbf{h}_i, \mathbf{h}) s[\mathbf{h}_i](\omega)$$

for each $\omega \in E_{\mathbf{h}}$. Therefore, for all $\mathbf{h} \in \mathcal{H}$,

$$(\mathbf{h}, \mathbf{s}(\omega)) = s[\mathbf{h}](\omega)$$

for each $\omega \in E \cap E_{\mathbf{h}}$, and $P(E \cap E_{\mathbf{h}}) = 1$. Since the probability space $(\Omega, \mathcal{F}, P)$ was assumed to be complete, and since $s[\mathbf{h}]$ is a scalar random variable for each $\mathbf{h} \in \mathcal{H}$, it follows that $(\mathbf{h}, \mathbf{s})$ is a scalar random variable for each $\mathbf{h} \in \mathcal{H}$. Therefore the map $\mathbf{s} : \Omega \to \mathcal{H}$ is an $\mathcal{H}$-valued random variable. Since

$$\mathcal{E}\|\mathbf{s}\|^2 = \mathcal{E}\sum_{i=1}^{\infty} s^2[\mathbf{h}_i] = \sum_{i=1}^{\infty} \mathcal{E}s^2[\mathbf{h}_i] < \infty,$$

$\mathbf{s}$ is a second-order $\mathcal{H}$-valued random variable. Since $s[\cdot]$ is bounded as a map from $\mathcal{H}$ into $L^2(\Omega, \mathcal{F}, P)$,

$$\lim_{n\to\infty} \mathcal{E}\left(s[\mathbf{h}] - \sum_{i=1}^{n}(\mathbf{h}_i, \mathbf{h})s[\mathbf{h}_i]\right)^2 = 0$$

for all $\mathbf{h} \in \mathcal{H}$, and since Eq. (34) holds for all $\mathbf{h} \in \mathcal{H}$ and $\omega \in E$, it follows that

$$\mathcal{E}\left(s[\mathbf{h}] - (\mathbf{h}, \mathbf{s})\right)^2 = 0$$

for all $\mathbf{h} \in \mathcal{H}$. Therefore, for all $\mathbf{h} \in \mathcal{H}$, $(\mathbf{h}, \mathbf{s}) = s[\mathbf{h}]$ wp1.

B The Hilbert spaces Φ_p

Let $\mathcal{H}$ be a real, separable Hilbert space, with inner product and corresponding norm denoted by $(\cdot, \cdot)$ and $\|\cdot\|$, respectively. Denote by $\mathcal{B}(\mathcal{H})$ the Borel field generated by the open sets in $\mathcal{H}$. For convenience it will be assumed in this Appendix that $\mathcal{H}$ is infinite-dimensional.

Appendix B.1 uses a self-adjoint linear operator on $\mathcal{H}$ to construct a special family of Hilbert spaces $\{\Phi_p, p \geq 0\}$. The inner product and corresponding norm on Φ_p are denoted by $(\cdot, \cdot)_p$ and $\|\cdot\|_p$, respectively, for each $p \geq 0$. These Hilbert spaces have the following properties: (i) $\Phi_0 = \mathcal{H}$; (ii) for each $p > 0$, $\Phi_p \subset \mathcal{H}$, and therefore Φ_p is real and separable; (iii) for each $p > 0$, Φ_p is dense in $\mathcal{H}$, and therefore Φ_p is infinite-dimensional; and (iv) if $0 \leq q \leq r$, then $\|\mathbf{h}\| = \|\mathbf{h}\|_0 \leq \|\mathbf{h}\|_q \leq \|\mathbf{h}\|_r$ for all $\mathbf{h} \in \Phi_r$, and therefore $\mathcal{H} = \Phi_0 \supset \Phi_q \supset \Phi_r$. In view of property (iv), the family $\{\Phi_p, p \geq 0\}$ is called a *decreasing family* of Hilbert spaces. The construction given here follows closely that of Kallianpur and Xiong (1995, Example 1.3.2, pp. 40–42). For various concrete examples and classical applications of decreasing families of Hilbert spaces constructed in this way, see Reed and Simon (1972, pp. 141–145), Itô (1984, pp. 1–12), Kallianpur and Xiong (1995, pp. 29–40), and Lax (2006, pp. 61–67).

Appendix B.2 discusses the spaces Φ_p in case $\mathcal{H} = L^2(S)$, the space of square-integrable vector or scalar fields on the sphere S, when the operator $\mathbf{L}$ used in the construction of the spaces Φ_p is taken to be $\mathbf{L} = -\Delta$, where Δ is the Laplacian operator on the sphere.

B.1 Construction of the Hilbert spaces Φ_p

Let $\mathbf{L}$ be a densely defined, positive semidefinite, self-adjoint linear operator on $\mathcal{H}$, and let $\mathbf{I}$ denote the identity operator on $\mathcal{H}$. It follows from elementary arguments (e.g. Riesz and Sz.-Nagy (1955, p. 324)) that the inverse operator $(\mathbf{I}+\mathbf{L})^{-1}$ is a *bounded*, positive semidefinite, self-adjoint linear operator defined on *all* of $\mathcal{H}$, in fact with

$$\|(\mathbf{I}+\mathbf{L})^{-1}\mathbf{h}\| \le \|\mathbf{h}\|$$

for all $\mathbf{h} \in \mathcal{H}$. Assume that some power $p_1 > 0$ of $(\mathbf{I}+\mathbf{L})^{-1}$ is a compact operator on $\mathcal{H}$. Then it follows from the Hilbert-Schmidt theorem (e.g. Reed and Simon (1972, Theorem VI.16, p. 203)) that there exists a countable orthonormal basis for $\mathcal{H}$ which consists of eigenvectors $\{\mathbf{g}_i\}_{i=1}^{\infty}$ of $(\mathbf{I}+\mathbf{L})^{-p_1}$,

$$(\mathbf{I}+\mathbf{L})^{-p_1}\mathbf{g}_i = \mu_i \mathbf{g}_i$$

for $i = 1, 2, \ldots$, where the corresponding eigenvalues $\{\mu_i\}_{i=1}^{\infty}$ satisfy $1 \ge \mu_1 \ge \mu_2 \ge \cdots$, with $\mu_i \to 0$ as $i \to \infty$. Moreover, $\mu_i > 0$ for $i = 1, 2, \ldots$, for suppose otherwise. Then there is a first zero eigenvalue, call it μ_{M+1}, since the eigenvalues decrease monotonically toward zero. Therefore $(\mathbf{I}+\mathbf{L})^{-p_1}$ has finite rank M, hence $\mathbf{I}+\mathbf{L}$ is defined everywhere in $\mathcal{H}$ and also has rank M. But $\operatorname{rank}(\mathbf{I}+\mathbf{L}) \ge \operatorname{rank}\mathbf{I} = \infty$ since $\mathbf{L}$ is positive semidefinite and $\mathcal{H}$ was assumed infinite-dimensional, a contradiction.

Now define $\{\lambda_i\}_{i=1}^{\infty}$ by $(1+\lambda_i)^{-p_1} = \mu_i$. Then $0 \le \lambda_1 \le \lambda_2 \le \cdots$, with $\lambda_i \to \infty$ as $i \to \infty$, and $\lambda_i < \infty$ for $i = 1, 2, \ldots$ since $\mu_i > 0$ for $i = 1, 2, \ldots$. Since the function $\lambda(\mu) = \mu^{-1/p_1} - 1$ is measurable and finite for $\mu \in (0, 1]$, it follows from the functional calculus for self-adjoint operators (e.g. Riesz and Sz.-Nagy (1955, pp. 343–346), Reed and Simon (1972, pp. 259–264)) that

$$\mathbf{L}\mathbf{g}_i = \lambda_i \mathbf{g}_i$$

for $i = 1, 2, \ldots$, and similarly for all $p \ge 0$ that

$$(\mathbf{I}+\mathbf{L})^p \mathbf{g}_i = (1+\lambda_i)^p \mathbf{g}_i$$

for $i = 1, 2, \ldots$, with $(\mathbf{I}+\mathbf{L})^p$ densely defined and self-adjoint in $\mathcal{H}$ for all $p \ge 0$.

For each $p \ge 0$, denote by Φ_p the domain of definition of $(\mathbf{I}+\mathbf{L})^p$, i.e.,

$$\Phi_p = \{\mathbf{h} \in \mathcal{H} : \|(\mathbf{I}+\mathbf{L})^p \mathbf{h}\| < \infty\}.$$

In particular, $\Phi_0 = \mathcal{H}$. Now

$$\begin{aligned} \|(\mathbf{I}+\mathbf{L})^p \mathbf{h}\|^2 &= \sum_{i=1}^{\infty} ((\mathbf{I}+\mathbf{L})^p \mathbf{h}, \mathbf{g}_i)^2 = \sum_{i=1}^{\infty} (\mathbf{h}, (\mathbf{I}+\mathbf{L})^p \mathbf{g}_i)^2 \\ &= \sum_{i=1}^{\infty} (\mathbf{h}, (1+\lambda_i)^p \mathbf{g}_i)^2 = \sum_{i=1}^{\infty} (1+\lambda_i)^{2p} (\mathbf{h}, \mathbf{g}_i)^2 \end{aligned}$$

for each $p \geq 0$, where the first equality is Parseval's relation and the second one is due to the fact that $(\mathbf{I}+\mathbf{L})^p$ is self-adjoint. Thus for each $p \geq 0$, Φ_p is given explicitly by

$$\Phi_p = \left\{ \mathbf{h} \in \mathcal{H} : \sum_{i=1}^{\infty} (1+\lambda_i)^{2p} (\mathbf{h}, \mathbf{g}_i)^2 < \infty \right\}.$$

Using this formula, it can be checked that for each $p \geq 0$, Φ_p is an inner product space, with inner product $(\cdot,\cdot)_p$ defined by

$$(\mathbf{g}, \mathbf{h})_p = \sum_{i=1}^{\infty} (1+\lambda_i)^{2p} (\mathbf{g}, \mathbf{g}_i)(\mathbf{h}, \mathbf{g}_i) = ((\mathbf{I}+\mathbf{L})^p \mathbf{g}, (\mathbf{I}+\mathbf{L})^p \mathbf{h})$$

for all $\mathbf{g}, \mathbf{h} \in \Phi_p$, and corresponding norm $\|\cdot\|_p$ defined by

$$\|\mathbf{h}\|_p^2 = (\mathbf{h}, \mathbf{h})_p = \|(\mathbf{I}+\mathbf{L})^p \mathbf{h}\|^2$$

for all $\mathbf{h} \in \Phi_p$. It follows also that if $0 \leq q \leq r$, then $\|\mathbf{h}\| = \|\mathbf{h}\|_0 \leq \|\mathbf{h}\|_q \leq \|\mathbf{h}\|_r$ for all $\mathbf{h} \in \Phi_r$, and therefore that $\Phi_r \subset \Phi_q \subset \mathcal{H}$.

Each inner product space Φ_p, $p > 0$, is in fact a Hilbert space, i.e., is already complete in the norm $\|\cdot\|_p$. To see this, suppose that $\{\mathbf{h}_n\}_{n=1}^{\infty}$ is a Cauchy sequence in Φ_p for some fixed $p > 0$, i.e. that $\|\mathbf{h}_n - \mathbf{h}_m\|_p \to 0$ as $n, m \to \infty$. Since $\|\mathbf{h}_n - \mathbf{h}_m\| \leq \|\mathbf{h}_n - \mathbf{h}_m\|_p$ for all $n, m \geq 1$, it follows that $\{\mathbf{h}_n\}_{n=1}^{\infty}$ is also a Cauchy sequence in $\mathcal{H}$, and since $\mathcal{H}$ is complete, the sequence converges to a unique element $\mathbf{h}_\infty \in \mathcal{H}$. It remains to show that in fact $\mathbf{h}_n \to \mathbf{h}_\infty \in \Phi_p$ as $n \to \infty$.

Now

$$\|\mathbf{h}_n - \mathbf{h}_m\|_p^2 = \|(\mathbf{I}+\mathbf{L})^p (\mathbf{h}_n - \mathbf{h}_m)\|^2 = \sum_{i=1}^{\infty} (1+\lambda_i)^{2p} (\mathbf{h}_n - \mathbf{h}_m, \mathbf{g}_i)^2.$$

Thus, that $\{\mathbf{h}_n\}_{n=1}^{\infty}$ is a Cauchy sequence in Φ_p means that, given any $\epsilon > 0$, there exists an $M = M(\epsilon)$ such that, for all $n, m \geq M$,

$$\sum_{i=1}^{I} (1+\lambda_i)^{2p} (\mathbf{h}_n - \mathbf{h}_m, \mathbf{g}_i)^2 < \epsilon$$

for any $I \geq 1$. But for each $i = 1, 2, \ldots,$

$$|(\mathbf{h}_m - \mathbf{h}_\infty, \mathbf{g}_i)| \leq \|\mathbf{h}_m - \mathbf{h}_\infty\| \, \|\mathbf{g}_i\| = \|\mathbf{h}_m - \mathbf{h}_\infty\| \to 0 \text{ as } m \to \infty,$$

hence $(\mathbf{h}_m, \mathbf{g}_i) \to (\mathbf{h}_\infty, \mathbf{g}_i)$ as $m \to \infty$, and therefore

$$\sum_{i=1}^{I} (1+\lambda_i)^{2p} (\mathbf{h}_n - \mathbf{h}_\infty, \mathbf{g}_i)^2 < \epsilon$$

for all $n \geq M$ and $I \geq 1$. Letting $I \to \infty$ then gives

$$\|\mathbf{h}_n - \mathbf{h}_\infty\|_p^2 = \sum_{i=1}^{\infty} (1+\lambda_i)^{2p} (\mathbf{h}_n - \mathbf{h}_\infty, \mathbf{g}_i)^2 < \epsilon$$

for all $n \geq M$, and therefore $\mathbf{h}_n \to \mathbf{h}_\infty \in \Phi_p$ as $n \to \infty$.

Thus, for each $p > 0$, Φ_p is a Hilbert space, with inner product $(\cdot, \cdot)_p$ and corresponding norm $\|\cdot\|_p$. It can be checked that $\{(1+\lambda_i)^{-p}\mathbf{g}_i\}_{i=1}^{\infty}$ is an orthonormal basis for Φ_p, for each $p > 0$.[2]

B.2 The case $\mathcal{H} = L^2(S)$ with $\mathbf{L} = -\Delta$

Now let $\mathcal{H} = L^2(S)$, the Hilbert space of real, Lebesgue square-integrable scalars on the unit 2-sphere S, with inner product

$$(\phi, \psi) = \int_S \phi(\mathbf{x})\psi(\mathbf{x})\, d\mathbf{x}$$

for all $\phi, \psi \in L^2(S)$, where $\mathbf{x} = (x_1, x_2)$ denotes spherical coordinates on S and $d\mathbf{x}$ denotes the surface area element, and with corresponding norm $\|\phi\| = (\phi, \phi)^{1/2}$ for all $\phi \in L^2(S)$. Let $\mathbf{L} = -\Delta$, where Δ is the Laplacian operator on $L^2(S)$. Thus $\mathbf{L}$ is a densely defined, positive semidefinite, self-adjoint linear operator on $L^2(S)$. Denote by I the identity operator on $L^2(S)$.

It will be shown first that for all $p_1 > 1/2$, $(I-\Delta)^{-p_1}$ is a Hilbert-Schmidt operator on $L^2(S)$, hence a compact operator on $L^2(S)$. By Appendix B.1, this allows construction of the decreasing family of Hilbert spaces $\{\Phi_p = \Phi_p(S), p \geq 0\}$,

$$\Phi_p = \{\phi \in L^2(S) : \|(I-\Delta)^p \phi\| < \infty\},$$

with inner product

$$(\phi, \psi)_p = ((I-\Delta)^p \phi, (I-\Delta)^p \psi)$$

for all $\phi, \psi \in \Phi_p$, and corresponding norm $\|\phi\|_p = (\phi, \phi)_p^{1/2}$ for all $\phi \in \Phi_p$. Thus if $\phi \in \Phi_p$ and p is a positive integer or half-integer, then all partial (directional) derivatives of ϕ up to order $2p$ are Lebesgue square-integrable.

Second, a Sobolev-type lemma for the sphere will be established, showing that if $\phi \in \Phi_{1/2+q}$ with $q > 0$, then ϕ is a bounded function on S, with bound

$$\max_{\mathbf{x} \in D} |\phi(\mathbf{x})|^2 < \tfrac{1}{4\pi}(1 + \tfrac{1}{2q})\|\phi\|_{1/2+q}^2. \tag{35}$$

It follows that if $\phi \in \Phi_{1+q}$ with $q > 0$, then the first partial derivatives of ϕ are bounded functions on the sphere, and in particular that $\Phi_{1+q} \subset C^0(S)$, the space of continuous functions on the sphere. It will be shown that, in fact, if $\phi \in \Phi_{1+q}$ with $q > 0$, then ϕ is Lipschitz continuous on S. Thus, for any

[2] It follows that, for any sequence $\{r_n\}_{n=0}^{\infty}$ with $0 \leq r_0 < r_1 < r_2 < \cdots \to \infty$, $\Phi = \cap_{n=0}^{\infty} \Phi_{r_n}$ is a separable Frèchet space, and since the norms $\|\cdot\|_{r_n}$ are Hilbertian seminorms on Φ, also a countably Hilbertian space. If $(\mathbf{I}+\mathbf{L})^{-p_1}$ is not just compact but in fact Hilbert-Schmidt, and if, for instance, $p_n = np_1$, then $\Phi = \cap_{n=0}^{\infty} \Phi_{p_n}$ is a countably Hilbertian nuclear space, and it is possible to define Φ'-valued random variables, where Φ' is the dual space of Φ. Such random variables are useful for stochastic differential equations in infinite-dimensional spaces (see the books of Itô (1984) and Kallianpur and Xiong (1995)), but are not immediately important for the principle of energetic consistency developed in this chapter.

$q > 0$ and any nonnegative integer l, $\Phi_{1+l/2+q} \subset C^l(S)$, the space of functions with l continuous partial derivatives on the sphere, and in fact all of the partial derivatives up to order l of a function $\phi \in \Phi_{1+l/2+q}$ are Lipschitz continuous.

These results carry over to vectors in the usual way. Thus denoting by $L^2(S)$ also the Hilbert space of real, Lebesgue square-integrable n-vectors on S, the inner product is

$$(\mathbf{g}, \mathbf{h}) = \int_S \mathbf{g}^T(\mathbf{x})\mathbf{h}(\mathbf{x})\, d\mathbf{x} = \sum_{i=1}^n (g_i, h_i)$$

for all $\mathbf{g}, \mathbf{h} \in L^2(S)$, and the corresponding norm is $\|\mathbf{h}\| = (\mathbf{h}, \mathbf{h})^{1/2}$ for all $\mathbf{h} \in L^2(S)$. Thus for n-vectors on S, the Hilbert spaces Φ_p, $p \geq 0$, are defined by

$$\Phi_p = \{\mathbf{h} \in L^2(S) : \|(I - \Delta)^p \mathbf{h}\| < \infty\},$$

with inner product

$$(\mathbf{g}, \mathbf{h})_p = ((I - \Delta)^p \mathbf{g}, (I - \Delta)^p \mathbf{h}) = \sum_{i=1}^n (g_i, h_i)_p$$

for all $\mathbf{g}, \mathbf{h} \in \Phi_p$, and corresponding norm $\|\mathbf{h}\|_p = (\mathbf{h}, \mathbf{h})_p^{1/2}$ for all $\mathbf{h} \in \Phi_p$.

To establish that $(I - \Delta)^{-p}$ is a Hilbert-Schmidt operator on $L^2(S)$ if $p > 1/2$, note first that

$$\sum_{l=0}^{\infty} \frac{2l+1}{[1 + l(l+1)]^{1+2\epsilon}} < 1 + \frac{1}{2\epsilon} \tag{36}$$

if $\epsilon > 0$. To obtain this inequality, let

$$f(x) = \frac{2x+1}{[1 + x(x+1)]^{1+2\epsilon}}$$

for $x \geq 0$ and $\epsilon > 0$. Then f is monotone decreasing for $x \geq 1/2$, and $f(0) > f(1)$, and so

$$\sum_{l=0}^{\infty} \frac{2l+1}{[1 + l(l+1)]^{1+2\epsilon}} = f(0) + \sum_{l=1}^{\infty} f(l) < f(0) + \int_0^{\infty} f(x)\, dx = 1 + \frac{1}{2\epsilon}.$$

The sum in Eq. (36) diverges logarithmically for $\epsilon = 0$.

Now let $C = (I - \Delta)^{-p}$ with $p > 0$. Thus C is a bounded operator from $L^2(S)$ into $L^2(S)$, with $\|C\phi\| \leq \|\phi\|$ for all $\phi \in L^2(S)$. The real and imaginary parts of the spherical harmonics Y_l^m form an orthonormal basis for $L^2(S)$, and

$$\Delta Y_l^m = -l(l+1) Y_l^m$$

for $l \geq 0$ and $|m| \leq l$. Thus

$$C Y_l^m = \lambda_l^m Y_l^m,$$

with eigenvalues $\lambda_l^m = (Y_l^m, CY_l^m) = [1 + l(l+1)]^{-p}$ for $l \geq 0$ and $|m| \leq l$. But

$$\sum_{l=0}^{\infty} \sum_{m=-l}^{l} (\lambda_l^m)^2 = \sum_{l=0}^{\infty} \frac{2l+1}{[1+l(l+1)]^{2p}},$$

and so this sum is finite for $p > 1/2$ by Eq. (36). Hence C is Hilbert-Schmidt for $p > 1/2$.

To establish the bound of Eq. (35), suppose that $\phi \in \Phi_{1/2+q}$ with $q > 0$. Thus $(I - \Delta)^{1/2+q}\phi \in L^2(S)$ and has a spherical harmonic expansion

$$(I - \Delta)^{1/2+q}\phi = \sum_{l=0}^{\infty} \sum_{m=-l}^{l} \beta_l^m Y_l^m,$$

where the convergence is in $L^2(S)$, with

$$\|\phi\|_{1/2+q}^2 = \|(I - \Delta)^{1/2+q}\phi\|^2 = \sum_{l=0}^{\infty} \sum_{m=-l}^{l} |\beta_l^m|^2 < \infty. \tag{37}$$

Therefore

$$\phi = \sum_{l=0}^{\infty} \sum_{m=-l}^{l} [1 + l(l+1)]^{-1/2-q} \beta_l^m Y_l^m, \tag{38}$$

where the convergence is in $\Phi_{1/2+q}$. It will be shown that this series converges absolutely, hence pointwise, so that

$$\phi(\mathbf{x}) = \sum_{l=0}^{\infty} \sum_{m=-l}^{l} [1 + l(l+1)]^{-1/2-q} \beta_l^m Y_l^m(\mathbf{x})$$

for each $\mathbf{x} \in S$. This will also give Eq. (35).

Now,

$$|\phi| \leq \sum_{l=0}^{\infty} [1 + l(l+1)]^{-1/2-q} \sum_{m=-l}^{l} |\beta_l^m| \, |Y_l^m|,$$

and so

$$|\phi| \leq \sum_{l=0}^{\infty} [1 + l(l+1)]^{-1/2-q} \left\{ \sum_{m=-l}^{l} |\beta_l^m|^2 \right\}^{1/2} \left\{ \sum_{m=-l}^{l} |Y_l^m|^2 \right\}^{1/2}$$

by the Schwarz inequality. The spherical harmonic addition theorem says that

$$P_l(\cos\gamma) = \frac{4\pi}{2l+1} \sum_{m=-l}^{l} Y_l^m(\mathbf{x}) \overline{Y}_l^m(\mathbf{y})$$

for $l \geq 0$, where P_l is the l^{th} Legendre polynomial and γ is the angle between $\mathbf{x}$ and $\mathbf{y}$. This implies that

$$\sum_{m=-l}^{l} |Y_l^m(\mathbf{x})|^2 = \frac{2l+1}{4\pi}$$

for all $\mathbf{x} \in S$, and so

$$|\phi| \leq \frac{1}{\sqrt{4\pi}} \sum_{l=0}^{\infty} [2l+1]^{1/2} [1 + l(l+1)]^{-1/2-q} \left\{ \sum_{m=-l}^{l} |\beta_l^m|^2 \right\}^{1/2}.$$

Another application of the Schwarz inequality then gives

$$|\phi| \leq \frac{1}{\sqrt{4\pi}} \left\{ \sum_{l=0}^{\infty} [2l+1][1 + l(l+1)]^{-1-2q} \right\}^{1/2} \left\{ \sum_{l=0}^{\infty} \sum_{m=-l}^{l} |\beta_l^m|^2 \right\}^{1/2},$$

or, using Eq. (37),

$$|\phi|^2 \leq \frac{1}{4\pi} \|\phi\|_{1/2+q}^2 \sum_{l=0}^{\infty} \frac{2l+1}{[1 + l(l+1)]^{1+2q}}.$$

Therefore, by Eq. (36), the sum in Eq. (38) converges absolutely, and Eq. (35) holds.

Now suppose that $\phi \in \Phi_{1+q}$ with $q > 0$. To establish that ϕ is Lipschitz continuous on S, note first that by the previous result,

$$\phi(\mathbf{x}) = \sum_{l=0}^{\infty} \sum_{m=-l}^{l} [1 + l(l+1)]^{-1-q} \beta_l^m Y_l^m(\mathbf{x})$$

for each $\mathbf{x} \in S$, where

$$\|\phi\|_{1+q}^2 = \|(I - \Delta)^{1+q} \phi\|^2 = \sum_{l=0}^{\infty} \sum_{m=-l}^{l} |\beta_l^m|^2 < \infty. \tag{39}$$

Therefore,

$$|\phi(\mathbf{x}) - \phi(\mathbf{y})| \leq \sum_{l=0}^{\infty} \sum_{m=-l}^{l} [1 + l(l+1)]^{-1-q} |\beta_l^m| \, |Y_l^m(\mathbf{x}) - Y_l^m(\mathbf{y})|$$

for each $\mathbf{x}, \mathbf{y} \in S$, and so by the Schwarz inequality,

$$|\phi(\mathbf{x}) - \phi(\mathbf{y})| \leq \sum_{l=0}^{\infty} [1 + l(l+1)]^{-1-q} \left\{ \sum_{m=-l}^{l} |\beta_l^m|^2 \right\}^{1/2} \left\{ \sum_{m=-l}^{l} |Y_l^m(\mathbf{x}) - Y_l^m(\mathbf{y})|^2 \right\}^{1/2}.$$

By the spherical harmonic addition theorem,

$$\begin{aligned}\sum_{m=-l}^{l}|Y_l^m(\mathbf{x})-Y_l^m(\mathbf{y})|^2 &= \sum_{m=-l}^{l}\left[|Y_l^m(\mathbf{x})|^2-2\mathrm{Re}Y_l^m(\mathbf{x})\overline{Y}_l^m(\mathbf{y})+|Y_l^m(\mathbf{y})|^2\right]\\ &= \frac{2l+1}{2\pi}\left[1-P_l(\cos\gamma)\right],\end{aligned}$$

where $\gamma=\gamma(\mathbf{x},\mathbf{y})$ is the angle between $\mathbf{x}$ and $\mathbf{y}$. Therefore,

$$|\phi(\mathbf{x})-\phi(\mathbf{y})|\le\sum_{l=0}^{\infty}[1+l(l+1)]^{-1-q}\left(\frac{2l+1}{2\pi}\right)^{1/2}[1-P_l(\cos\gamma)]^{1/2}\left\{\sum_{m=-l}^{l}|\beta_l^m|^2\right\}^{1/2},$$

and so by Eq. (39) and the Schwarz inequality,

$$|\phi(\mathbf{x})-\phi(\mathbf{y})|\le\frac{1}{\sqrt{2\pi}}\left\{\sum_{l=0}^{\infty}[1+l(l+1)]^{-2-2q}(2l+1)[1-P_l(\cos\gamma)]\right\}^{1/2}\|\phi\|_{1+q}.$$

Now, $P_l(1)=1$, $P_l'(1)=l(l+1)/2$, and $P_l''(1)=[l(l+1)-2]P_l'(1)/4\ge 0$ for $l\ge 0$. It follows that for γ sufficiently small,

$$1-P_l(\cos\gamma)\le(1-\cos\gamma)P_l'(1)=l(l+1)\sin^2\tfrac{\gamma}{2},$$

and so

$$|\phi(\mathbf{x})-\phi(\mathbf{y})|\le\frac{K}{\sqrt{2\pi}}\|\phi\|_{1+q}\left|\sin\frac{\gamma(\mathbf{x},\mathbf{y})}{2}\right|, \tag{40}$$

where

$$K^2=\sum_{l=0}^{\infty}[1+l(l+1)]^{-2-2q}(2l+1)l(l+1).$$

This series converges for $q>0$ since the terms decay like l^{-1-4q}, and Eq. (40) shows that ϕ is Lipschitz continuous.

C Some basic concepts and definitions

This appendix summarizes background material used elsewhere in this article. For further treatment see, for instance, Doob (1953), Royden (1968), and Reed and Simon (1972).

C.1 Measure spaces

Let X be a set. A collection $\mathcal{C}$ of subsets of X is called a *σ-algebra*, or *Borel field*, if (i) the empty set $\emptyset$ is in $\mathcal{C}$, (ii) for every set $A\in\mathcal{C}$, the complement $\tilde{A}=\{x\in X: x\notin A\}$ of A is in $\mathcal{C}$, and (iii) for every countable collection $\{A_i\}_{i=1}^{\infty}$ of sets $A_i\in\mathcal{C}$, the union $\cup_{i=1}^{\infty}A_i$ of the sets is in $\mathcal{C}$. Given any collection $\mathcal{A}$ of subsets of X, there is a smallest σ-algebra which contains $\mathcal{A}$, i.e., there is

a σ-algebra $\mathcal{C}$ such that (i) $\mathcal{A} \subset \mathcal{C}$, and (ii) if $\mathcal{B}$ is a σ-algebra and $\mathcal{A} \subset \mathcal{B}$ then $\mathcal{C} \subset \mathcal{B}$. The smallest σ-algebra containing a given collection $\mathcal{A}$ of subsets of X is called the *Borel field of X generated by $\mathcal{A}$*. A *measurable space* is a couple $(X, \mathcal{C})$ consisting of a set X and a σ-algebra $\mathcal{C}$ of subsets of X. If $(X, \mathcal{C})$ is a measurable space and $Y \in \mathcal{C}$, then $(Y, \mathcal{C}_Y)$ is a measurable space, where

$$\mathcal{C}_Y = \{A \in \mathcal{C} : A \subset Y\},$$

i.e., $\mathcal{C}_Y$ consists of all the sets in $\mathcal{C}$ that are subsets of Y.

The set $\mathbb{R}^e$ of *extended real numbers* is the union of the set $\mathbb{R}$ of real numbers and the sets $\{\infty\}$ and $\{-\infty\}$. Multiplication of any two extended real numbers is defined as usual, with the convention that $0 \cdot \infty = 0$. Addition and subtraction of any two extended real numbers is also defined, except that $\infty - \infty$ is undefined, as usual.

Let Y and Z be two sets. A function g is called a *map* from Y into Z, written $g : Y \to Z$, if $g(y)$ is defined for all $y \in Y$ and $g(y) \in Z$ for all $y \in Y$. Thus a map $g : \mathbb{R} \to \mathbb{R}$ is a real-valued function defined on all of the real line, a map $g : Y \to \mathbb{R}$ is a real-valued function defined on all of Y, and a map $g : Y \to \mathbb{R}^e$ is an extended real-valued function defined on all of Y.

Let $(X, \mathcal{C})$ be a measurable space. A subset A of X is called *measurable* if $A \in \mathcal{C}$. A map $g : X \to \mathbb{R}^e$ is called *measurable* (with respect to $\mathcal{C}$) if

$$\{x \in X : g(x) \leq \alpha\} \in \mathcal{C},$$

for every $\alpha \in \mathbb{R}$. If $g : X \to \mathbb{R}^e$ is measurable then $|g|$ is measurable, and if $h : X \to \mathbb{R}^e$ is another measurable map then gh is measurable. A *measure* μ on $(X, \mathcal{C})$ is a map $\mu : \mathcal{C} \to \mathbb{R}^e$ that satisfies (i) $\mu(A) \geq 0$ for every measurable set A, (ii) $\mu(\emptyset) = 0$, and (iii)

$$\mu\left(\bigcup_{i=1}^{\infty} E_i\right) = \sum_{i=1}^{\infty} \mu(E_i),$$

for every countable collection $\{E_i\}_{i=1}^{\infty}$ of disjoint measurable sets, i.e., for every countable collection of sets $E_i \in \mathcal{C}$ with $\cap_{i=1}^{\infty} E_i = \emptyset$. A *measure space* $(X, \mathcal{C}, \mu)$ is a measurable space $(X, \mathcal{C})$ together with a measure μ on $(X, \mathcal{C})$.

Let $(X, \mathcal{C}, \mu)$ be a measure space. A condition $C(x)$ defined for all $x \in X$ is said to hold *almost everywhere* (a.e.) (with respect to μ) if the set $E = \{x \in X : C(x) \text{ is false}\}$ on which it fails to hold is a measurable set of measure zero, i.e., $E \in \mathcal{C}$ and $\mu(E) = 0$. In particular, two maps $g : X \to \mathbb{R}^e$ and $h : X \to \mathbb{R}^e$ are said to be equal almost everywhere, written $g = h$ a.e., if the subset of X on which they are not equal is a measurable set of measure zero.

A measure space $(X, \mathcal{C}, \mu)$ is called *complete* if $\mathcal{C}$ contains all subsets of measurable sets of measure zero, i.e., if $B \in \mathcal{C}$, $\mu(B) = 0$, and $A \subset B$ together imply that $A \in \mathcal{C}$. If $(X, \mathcal{C}, \mu)$ is a complete measure space and A is a subset of a measurable set of measure zero, then $\mu(A) = 0$. If $(X, \mathcal{C}, \mu)$ is a measure space then there is a complete measure space $(X, \mathcal{C}_0, \mu_0)$, called the *completion* of $(X, \mathcal{C}, \mu)$, which is determined uniquely by the conditions that (i) $\mathcal{C} \subset \mathcal{C}_0$, (ii)

if $D \in \mathcal{C}$ then $\mu(D) = \mu_0(D)$, and (iii) $D \in \mathcal{C}_0$ if and only if $D = A \cup B$ where $B \in \mathcal{C}$ and $A \subset C \in \mathcal{C}$ with $\mu(C) = 0$. Thus a measure space can always be completed by enlarging its σ-algebra to include the subsets of measurable sets of measure zero and extending its measure so that the domain of definition of the extended measure includes the enlarged σ-algebra.

An *open interval* on the real number line $\mathbb{R}$ is a set $(\alpha, \beta) = \{x \in \mathbb{R} : \alpha < x < \beta\}$ with $\alpha, \beta \in \mathbb{R}^e$ and $\alpha < \beta$. Denote by $\mathcal{B}(\mathbb{R})$ the Borel field of $\mathbb{R}$ generated by the open intervals, and denote by $\mathcal{I}(\mathbb{R}) \subset \mathcal{B}(\mathbb{R})$ the sets that are countable unions of disjoint open intervals. For each set $I = \cup_{i=1}^{\infty}(\alpha_i, \beta_i) \in \mathcal{I}(\mathbb{R})$, define

$$m^*(I) = \sum_{i=1}^{\infty}(\beta_i - \alpha_i),$$

and for each set $B \in \mathcal{B}(\mathbb{R})$ define

$$m^*(B) = \inf m^*(I),$$

where the infimum (greatest lower bound) is taken over all those $I \in \mathcal{I}(\mathbb{R})$ such that $B \subset I$. Then m^* is a measure on the measurable space $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$. The completion of the measure space $(\mathbb{R}, \mathcal{B}(\mathbb{R}), m^*)$ is denoted by $(\mathbb{R}, \mathcal{M}, m)$. The sets in $\mathcal{M}$ are called the *Lebesgue measurable sets* on $\mathbb{R}$, and m is called *Lebesgue measure* on $\mathbb{R}$.

Let $(X, \mathcal{C}, \mu)$ be a complete measure space, and let $g : X \to \mathbb{R}^e$ and $h : X \to \mathbb{R}^e$ be two maps. If g is measurable and $g = h$ a.e., then h is measurable.

C.2 Integration

In this subsection let $(X, \mathcal{C}, \mu)$ be a measure space. The *characteristic function* χ_A of a subset A of X is the map $\chi_A : X \to \{0, 1\}$ defined for each $x \in X$ by

$$\chi_A(x) = \begin{cases} 1 & \text{if } x \in A \\ 0 & \text{if } x \notin A \end{cases}.$$

A characteristic function χ_A is a measurable map if, and only if, A is a measurable set. A map $\phi : X \to \mathbb{R}^e$ is called *simple* if it is measurable and takes on only a finite number of values. Thus the characteristic function of a measurable set is simple, and if ϕ is simple and takes on the values $\alpha_1, \ldots, \alpha_n$ then $\phi = \sum_{i=1}^{n} \alpha_i \chi_{E_i}$, where $E_i = \{x \in X : \phi(x) = \alpha_i\} \in \mathcal{C}$ for $i = 1, \ldots, n$. If ϕ is simple and the values $\alpha_1, \ldots, \alpha_n$ it takes on are all nonnegative, the integral of ϕ over a measurable set E with respect to measure μ is defined as

$$\int_E \phi \, d\mu = \sum_{i=1}^{n} \alpha_i \mu(E_i \cap E),$$

where $E_i = \{x \in X : \phi(x) = \alpha_i\}$ for $i = 1, \ldots, n$. It is possible that $\int_E \phi \, d\mu = \infty$, for instance if $\alpha_1 \neq 0$ and $\mu(E_1 \cap E) = \infty$, or if $\alpha_1 = \infty$ and $\mu(E_1 \cap E) \neq 0$.

Let E be a measurable set and let $g : X \to \mathbb{R}^e$ be a map which is nonnegative, i.e., $g(x) \geq 0$ for all $x \in X$. If g is measurable, the integral of g over E with respect to μ is defined as

$$\int_E g \, d\mu = \sup \int_E \phi \, d\mu,$$

where the supremum (least upper bound) is taken over all simple maps ϕ with $0 \leq \phi \leq g$. Function g is called *integrable* (over E, with respect to μ) if g is measurable and

$$\int_E g \, d\mu < \infty.$$

If $\{h_i\}_{i=1}^{\infty}$ is a collection of nonnegative measurable maps from X into $\mathbb{R}^e$, then $h = \sum_{i=1}^{\infty} h_i$ is a nonnegative measurable map from X into $\mathbb{R}^e$ and

$$\int_E h \, d\mu = \sum_{i=1}^{\infty} \int_E h_i \, d\mu,$$

and in particular, h is integrable if and only if $\sum_{i=1}^{\infty} \int_E h_i \, d\mu < \infty$.

Let E be a measurable set and let $g : X \to \mathbb{R}^e$ be a map. The *positive part* g^+ of g is the nonnegative map $g^+ = g \vee 0$, i.e., $g^+(x) = \max\{g(x), 0\}$ for each $x \in X$, and the *negative part* g^- is the nonnegative map $g^- = (-g) \vee 0$. Thus $g = g^+ - g^-$ and $|g| = g^+ + g^-$. If g is measurable, so are g^+ and g^-, as well as $|g|$. Function g is called *integrable* (over E, with respect to μ) if both g^+ and g^- are integrable, in which case the integral of g is defined as

$$\int_E g \, d\mu = \int_E g^+ \, d\mu - \int_E g^- \, d\mu.$$

Thus g is integrable over E if, and only if, $|g|$ is integrable over E, in which case

$$\left| \int_E g \, d\mu \right| \leq \int_E |g| \, d\mu < \infty.$$

If g is integrable over X, then $|g| < \infty$ a.e., g is integrable over E, and

$$\int_E |g| \, d\mu \leq \int_X |g| \, d\mu < \infty.$$

If g is measurable, then

$$\int_X |g| \, d\mu = 0$$

if, and only if, $g = 0$ a.e.

Let E be a measurable set and let $g : X \to \mathbb{R}^e$ and $h : X \to \mathbb{R}^e$ be two maps. If g^2 and h^2 are integrable over E then gh is integrable over E, and

$$\left| \int_E gh \, d\mu \right| \leq \int_E |gh| \, d\mu \leq \left(\int_E g^2 \, d\mu \right)^{1/2} \left(\int_E h^2 \, d\mu \right)^{1/2} < \infty. \qquad (41)$$

If g and h are integrable over E and $g = h$ a.e., then

$$\int_E g\,d\mu = \int_E h\,d\mu.$$

If the measure space is complete, and if g is integrable over E and $g = h$ a.e., then h is integrable over E and

$$\int_E g\,d\mu = \int_E h\,d\mu.$$

Now consider the complete measure space $(\mathbb{R}, \mathcal{M}, m)$, where $\mathcal{M}$ is the σ-algebra of Lebesgue measurable sets on $\mathbb{R}$ and m is Lebesgue measure on $\mathbb{R}$. If $g : \mathbb{R} \to \mathbb{R}^e$ is measurable with respect to $\mathcal{M}$, and is either nonnegative or integrable over $\mathbb{R}$ with respect to m, the integral of g over a Lebesgue measurable set E is called the *Lebesgue integral* of g over E, and is often written as

$$\int_E g\,dm = \int_E g(x)\,dx.$$

A *Borel measure* on $\mathbb{R}$ is a measure defined on the Lebesgue measurable sets $\mathcal{M}$ that is finite for bounded sets. If F is a monotone increasing function on $\mathbb{R}$ that is continuous on the right, i.e., if $F(\beta) \geq F(\alpha)$ and $\lim_{\beta \to \alpha} F(\beta) = F(\alpha)$ for all $\alpha, \beta \in \mathbb{R}$ with $\alpha < \beta$, then there exists a unique Borel measure μ on $\mathbb{R}$ such that

$$\mu((\alpha, \beta]) = F(\beta) - F(\alpha)$$

for all $\alpha, \beta \in \mathbb{R}$ with $\alpha < \beta$, where $(\alpha, \beta] = \{x \in \mathbb{R} : \alpha < x \leq \beta\}$. Let F be a monotone increasing function that is continuous on the right, and let μ be the corresponding Borel measure. If $g : \mathbb{R} \to \mathbb{R}^e$ is measurable with respect to $\mathcal{M}$, and is either nonnegative or integrable over $\mathbb{R}$ with respect to the Borel measure μ, the *Lebesgue-Stieltjes integral* of g over a Lebesgue measurable set E is defined as

$$\int_E g(x)\,dF(x) = \int_E g\,d\mu.$$

C.3 Probability

A *probability space* is a measure space $(\Omega, \mathcal{F}, P)$ with $P(\Omega) = 1$. The set Ω is called the *sample space*, the σ-algebra $\mathcal{F}$ of measurable sets is called the *event space*, a measurable set is called an *event*, and P is called the *probability measure*. For the rest of this subsection, let $(\Omega, \mathcal{F}, P)$ be a probability space.

A measurable map from Ω into $\mathbb{R}^e$ is called a (scalar) *random variable*. Thus a map $s : \Omega \to \mathbb{R}^e$ is a random variable if, and only if,

$$\{\omega \in \Omega : s(\omega) \leq x\} \in \mathcal{F}$$

for every $x \in \mathbb{R}$. In particular, if s is a random variable then the function

$$F_s(x) = P(\{\omega \in \Omega : s(\omega) \leq x\}),$$

called the probability *distribution function* of s, is defined for all $x \in \mathbb{R}$. The distribution function of a random variable is monotone increasing and continuous on the right. If the distribution function F_s of a random variable s is an indefinite integral, i.e., if

$$F_s(x) = \int_{-\infty}^{x} f_s(y)\, dy$$

for all $x \in \mathbb{R}$ and some Lebesgue integrable function f_s, then f_s is called the probabilty *density function* of s, and $dF_s/dx = f_s$ a.e. (with respect to Lebesgue measure) in $\mathbb{R}$.

The *expectation operator* $\mathcal{E}$ is the integration operator over Ω with respect to probability measure. Thus if s is a random variable then $\mathcal{E}|s|$ is defined, since $|s|$ is a random variable and $|s| \geq 0$, and

$$\mathcal{E}|s| = \int_{\Omega} |s|\, dP \leq \infty,$$

while a random variable s is integrable over Ω if, and only if, $\mathcal{E}|s| < \infty$, in which case

$$\mathcal{E}s = \int_{\Omega} s\, dP$$

and $|\mathcal{E}s| \leq \mathcal{E}|s| < \infty$. If s is a random variable with $\mathcal{E}|s| < \infty$, then $\overline{s} = \mathcal{E}s$ is called the *mean* of s, and the mean can be evaluated equivalently as the Lebesgue-Stieltjes integral

$$\mathcal{E}s = \int_{-\infty}^{\infty} x\, dF_s(x),$$

where F_s is the distribution function of s, hence

$$\mathcal{E}s = \int_{-\infty}^{\infty} x f_s(x)\, dx$$

if also s has a density function f_s, where the integral is the Lebesgue integral.

If s is a random variable then $\mathcal{E}s^2$ is defined, since $s^2 \geq 0$ is a random variable, and either $\mathcal{E}s^2 = \infty$ or $\mathcal{E}s^2 < \infty$. A random variable s is called *second-order* if $\mathcal{E}s^2 < \infty$. If r and s are random variables then $\mathcal{E}|rs|$ is defined since rs is a random variable, and $\mathcal{E}|rs| \leq \infty$. If r and s are second-order random variables, then

$$\mathcal{E}\,|rs| \leq \left(\mathcal{E}r^2\right)^{1/2} \left(\mathcal{E}s^2\right)^{1/2} < \infty$$

by Eq. (41), hence $\mathcal{E}rs$ is defined and $|\mathcal{E}rs| \leq \mathcal{E}\,|rs| < \infty$. In particular, on taking $r = 1$ and using the fact that

$$\mathcal{E}1 = \int_{\Omega} 1\, dP = P(\Omega) = 1,$$

it follows that if s is a second-order random variable then its mean $\overline{s} = \mathcal{E}s$ is defined, with

$$0 \le |\overline{s}| = |\mathcal{E}s| \le \mathcal{E}\,|s| \le \left(\mathcal{E}s^2\right)^{1/2} < \infty.$$

The *variance* $\sigma^2 = \mathcal{E}(s - \overline{s})^2$ of a second-order random variable s is therefore also defined, and finite, with

$$0 \le \sigma^2 = \mathcal{E}\left(s^2 - 2\overline{s}s + \overline{s}^2\right) = \mathcal{E}s^2 - \overline{s}^2 < \infty,$$

and

$$\mathcal{E}s^2 = \overline{s}^2 + \sigma^2. \tag{42}$$

A condition $C(\omega)$ defined for all $\omega \in \Omega$ is said to hold *with probability one* (wp1), or *almost surely* (a.s.), if it holds a.e. with respect to probability measure. Thus if s is a random variable, then $s = 0$ wp1 if, and only if, $\mathcal{E}|s| = 0$. If s is a random variable with $\mathcal{E}|s| < \infty$, i.e., if the mean of s is defined, then $|s| < \infty$ wp1. If r and s are two random variables with $\mathcal{E}|r| < \infty$ and $\mathcal{E}|s| < \infty$, and if $r = s$ wp1, then r and s have the same distribution function and, in particular, $\mathcal{E}r = \mathcal{E}s$. If the probability space is complete, and if s is a random variable, $r : \Omega \to \mathbb{R}^e$ and $r = s$ wp1, then r is a random variable and has the same distribution function as s, and if, in addition, $\mathcal{E}|s| < \infty$, then $\mathcal{E}|r| < \infty$ and $\mathcal{E}r = \mathcal{E}s$.

C.4 Hilbert space

A nonempty set V is called a *linear space* or *vector space* (over the reals) if $\alpha\mathbf{g} + \beta\mathbf{h} \in V$ for all $\mathbf{g}, \mathbf{h} \in V$ and $\alpha, \beta \in \mathbb{R}$. A *norm* on a linear space V is a real-valued function $\|\cdot\|$ such that, for all $\mathbf{g}, \mathbf{h} \in V$ and $\alpha \in \mathbb{R}$, (i) $\|\mathbf{h}\| \ge 0$, (ii) $\|\mathbf{h}\| = 0$ if, and only if, $\mathbf{h} = \mathbf{0}$, (iii) $\|\alpha\mathbf{h}\| = |\alpha|\,\|\mathbf{h}\|$, and (iv) $\|\mathbf{g}+\mathbf{h}\| \le \|\mathbf{g}\|+\|\mathbf{h}\|$. An *inner product* on a linear space V is a real-valued function $(\cdot,\cdot)$ such that, for all $\mathbf{f}, \mathbf{g}, \mathbf{h} \in V$ and $\alpha \in \mathbb{R}$, (i) $(\mathbf{h}, \mathbf{h}) \ge 0$, (ii) $(\mathbf{h}, \mathbf{h}) = 0$ if, and only if, $\mathbf{h} = \mathbf{0}$, (iii) $(\mathbf{g}, \alpha\mathbf{h}) = \alpha(\mathbf{g}, \mathbf{h})$, (iv) $(\mathbf{f}, \mathbf{g}+\mathbf{h}) = (\mathbf{f}, \mathbf{g})+(\mathbf{f}, \mathbf{h})$, and (v) $(\mathbf{g}, \mathbf{h}) = (\mathbf{h}, \mathbf{g})$. A *normed linear space* is a linear space equipped with a norm, and an *inner product space* is a linear space equipped with an inner product. Every inner product space V is a normed linear space, with norm $\|\cdot\|$ given by $\|\mathbf{h}\| = (\mathbf{h}, \mathbf{h})^{1/2}$ for all $\mathbf{h} \in V$, where $(\cdot,\cdot)$ is the inner product on V. A normed linear space V is an inner product space if, and only if, its norm $\|\cdot\|$ satisfies the *parallelogram law*

$$\|\mathbf{g} + \mathbf{h}\|^2 + \|\mathbf{g} - \mathbf{h}\|^2 = 2(\|\mathbf{g}\|^2 + \|\mathbf{h}\|^2),$$

for all $\mathbf{g}, \mathbf{h} \in V$. On every inner product space V, the inner product $(\cdot,\cdot)$ is given by the *polarization identity*

$$(\mathbf{g}, \mathbf{h}) = \tfrac{1}{4}(\|\mathbf{g} + \mathbf{h}\|^2 - \|\mathbf{g} - \mathbf{h}\|^2),$$

for all $\mathbf{g}, \mathbf{h} \in V$, where $\|\cdot\|$ is the norm corresponding to the inner product, i.e., $\|\mathbf{h}\| = (\mathbf{h}, \mathbf{h})^{1/2}$ for all $\mathbf{h} \in V$. The *Schwarz inequality*

$$|(\mathbf{g}, \mathbf{h})| \le \|\mathbf{g}\|\,\|\mathbf{h}\| < \infty,$$

for all $\mathbf{g}, \mathbf{h} \in V$, holds on every inner product space V, where $(\cdot, \cdot)$ is the inner product on V and $\| \cdot \|$ is the corresponding norm.

A subset O of a normed linear space V is called *open* in V if for every $\mathbf{g} \in O$, there exists an $\epsilon > 0$ such that if $\mathbf{h} \in V$ and $\|\mathbf{g} - \mathbf{h}\| < \epsilon$ then $\mathbf{h} \in O$. A subset B of a normed linear space V is called *dense* in V if for every $\mathbf{h} \in V$ and $\epsilon > 0$, there exists an element $\mathbf{g} \in B$ such that $\|\mathbf{g} - \mathbf{h}\| < \epsilon$. A normed linear space is called *separable* if it has a dense subset that contains countably many elements.

A sequence of elements $\mathbf{h}_1, \mathbf{h}_2, \ldots$ in a normed linear space V is called a *Cauchy sequence* if $\|\mathbf{h}_m - \mathbf{h}_n\| \to 0$ as $m, n \to \infty$. A sequence of elements $\mathbf{h}_1, \mathbf{h}_2, \ldots$ in a normed linear space V is said to *converge* in V if there exists an element $\mathbf{h} \in V$ such that $\|\mathbf{h} - \mathbf{h}_n\| \to 0$ as $n \to \infty$, in which case one writes $\mathbf{h} = \lim_{n \to \infty} \mathbf{h}_n$. A normed linear space V is called *complete* if every Cauchy sequence of elements in V converges in V. A complete normed linear space is called a *Banach space*. A Banach space on which the norm is defined by an inner product is called a *Hilbert space*. That is, a Hilbert space is an inner product space which is complete in the norm defined by the inner product.

Let $\mathcal{H}$ be a Hilbert space, with inner product $(\cdot, \cdot)$ and corresponding norm $\| \cdot \|$. A subset S of $\mathcal{H}$ is called an *orthogonal system* if $\mathbf{g} \neq \mathbf{0}$, $\mathbf{h} \neq \mathbf{0}$ and $(\mathbf{g}, \mathbf{h}) = 0$, for every $\mathbf{g}, \mathbf{h} \in \mathcal{H}$. An orthogonal system S is called an *orthogonal basis* (or *complete orthogonal system*) if no other orthogonal system contains S as a proper subset. An orthogonal basis S is called an *orthonormal basis* if $\|\mathbf{h}\| = 1$ for every $\mathbf{h} \in S$. There exists an orthonormal basis which has countably many elements if, and only if, $\mathcal{H}$ is separable. If $\mathcal{H}$ is a separable Hilbert space then every orthonormal basis for $\mathcal{H}$ has the same number of elements $N \leq \infty$, and N is called the *dimension* of $\mathcal{H}$.

Let $\mathcal{H}$ be a separable Hilbert space, with inner product $(\cdot, \cdot)$, corresponding norm $\| \cdot \|$, and orthonormal basis $S = \{\mathbf{h}_i\}_{i=1}^N$, $N \leq \infty$. If $\mathbf{h} \in \mathcal{H}$ then the sequence of partial sums $\sum_{i=1}^n (\mathbf{h}_i, \mathbf{h}) \mathbf{h}_i$ converges to $\mathbf{h}$, i.e.,

$$\lim_{n \to N} \|\mathbf{h} - \sum_{i=1}^{n} (\mathbf{h}_i, \mathbf{h}) \mathbf{h}_i\| = 0,$$

and so every $\mathbf{h} \in \mathcal{H}$ has the representation

$$\mathbf{h} = \sum_{i=1}^{N} (\mathbf{h}_i, \mathbf{h}) \mathbf{h}_i.$$

Furthermore,

$$(\mathbf{g}, \mathbf{h}) = \sum_{i=1}^{N} (\mathbf{h}_i, \mathbf{g})(\mathbf{h}_i, \mathbf{h}),$$

for every $\mathbf{g}, \mathbf{h} \in \mathcal{H}$. Therefore, for every $\mathbf{h} \in \mathcal{H}$,

$$\|\mathbf{h}\|^2 = \sum_{i=1}^{N} (\mathbf{h}_i, \mathbf{h})^2,$$

which is called *Parseval's relation*.

An example of a separable Hilbert space of dimension $N \leq \infty$ is the space ℓ_N^2 of square-summable sequences of N real numbers, with inner product $(\mathbf{g}, \mathbf{h}) = \sum_{i=1}^N g_i h_i$, where g_i and h_i denote element i of $\mathbf{g} \in \ell_N^2$ and $\mathbf{h} \in \ell_N^2$, respectively. An orthonormal basis for ℓ_N^2 is the set of unit vectors $\{\mathbf{e}_j\}_{j=1}^N$, where element i of $\mathbf{e}_j$ is 1 if $i = j$ and 0 if $i \neq j$. In case $N < \infty$, the elements of ℓ_N^2 are usually written as (column) N-vectors $\mathbf{g} = (g_1, \ldots, g_N)^T$, the inner product is then $(\mathbf{g}, \mathbf{h}) = \mathbf{g}^T \mathbf{h}$, and the columns of the $N \times N$ identity matrix constitute an orthonormal basis.

Let $(X, \mathcal{C}, \mu)$ be a measure space. Denote by $\mathcal{L}^1(X, \mathcal{C}, \mu)$ the set of integrable maps from X into $\mathbb{R}^e$, and consider the function $\|\cdot\|$ defined for all $g \in \mathcal{L}^1(X, \mathcal{C}, \mu)$ by

$$\|g\| = \int_X |g| \, d\mu.$$

The set $\mathcal{L}^1(X, \mathcal{C}, \mu)$ is a linear space, and the function $\|\cdot\|$ is by definition real-valued, i.e., $\|g\| < \infty$ for all $g \in \mathcal{L}^1(X, \mathcal{C}, \mu)$. The function $\|\cdot\|$ also satisfies all of the properties of a norm, except that $\|g\| = 0$ does not imply $g = 0$. However, $\|g\| = 0$ does imply that $g = 0$ a.e., and $g = 0$ a.e. implies that $\|g\| = 0$, for all $g \in \mathcal{L}^1(X, \mathcal{C}, \mu)$. Two maps g and h from X into $\mathbb{R}^e$ are called *equivalent*, or are said to belong to the same *equivalence class*, if $g = h$ a.e. If g and h are equivalent, and if $g, h \in \mathcal{L}^1(X, \mathcal{C}, \mu)$, then $\|g\| = \|h\|$. That is, $\|\cdot\|$ assigns the same real number to each member of a given equivalence class of elements of $\mathcal{L}^1(X, \mathcal{C}, \mu)$, and thereby the domain of definition of the function $\|\cdot\|$ is extended from the elements of $\mathcal{L}^1(X, \mathcal{C}, \mu)$ to the equivalence classes of elements of $\mathcal{L}^1(X, \mathcal{C}, \mu)$. The set $L^1(X, \mathcal{C}, \mu)$ of equivalence classes of elements of $\mathcal{L}^1(X, \mathcal{C}, \mu)$ is a linear space, and $\|\cdot\|$ is a norm on this space. The Riesz-Fischer theorem states that $L^1(X, \mathcal{C}, \mu)$ is complete in this norm, i.e., that $L^1(X, \mathcal{C}, \mu)$ is a Banach space under the norm $\|\cdot\|$. The elements of $L^1(X, \mathcal{C}, \mu)$, unlike those of $\mathcal{L}^1(X, \mathcal{C}, \mu)$, are not defined pointwise in X, and therefore are not maps.

Denote by $\mathcal{L}^2(X, \mathcal{C}, \mu)$ the set of square-integrable maps from X into $\mathbb{R}^e$, and consider the function $\|\cdot\|$ defined for all $g \in \mathcal{L}^2(X, \mathcal{C}, \mu)$ by

$$\|g\| = \left(\int_X g^2 \, d\mu \right)^{1/2}.$$

Again, the function $\|\cdot\|$ assigns the same real number to each member of any given equivalence class of elements of $\mathcal{L}^2(X, \mathcal{C}, \mu)$, i.e., to each $g, h \in \mathcal{L}^2(X, \mathcal{C}, \mu)$ such that $g = h$ a.e., and in particular, $\|g\| = 0$ if and only if $g = 0$ a.e. Thus the domain of definition of the function $\|\cdot\|$ can be extended to the equivalence classes. The set $L^2(X, \mathcal{C}, \mu)$ of equivalence classes of elements of $\mathcal{L}^2(X, \mathcal{C}, \mu)$ is a linear space, $\|\cdot\|$ is a norm on this space, and $L^2(X, \mathcal{C}, \mu)$ is complete in this norm. Therefore $L^2(X, \mathcal{C}, \mu)$ is a Banach space under the norm $\|\cdot\|$. Moreover, this norm satisfies the parallelogram law, and therefore $L^2(X, \mathcal{C}, \mu)$ is a Hilbert space. The polarization identity yields the inner product $(\cdot, \cdot)$ on $L^2(X, \mathcal{C}, \mu)$,

viz.,

$$(g, h) = \int_X gh \, d\mu,$$

for all $g, h \in L^2(X, \mathcal{C}, \mu)$. Again, the elements of $L^2(X, \mathcal{C}, \mu)$ are not defined pointwise and are not maps. The Schwarz inequality holds on $L^2(X, \mathcal{C}, \mu)$ since $L^2(X, \mathcal{C}, \mu)$ is an inner product space, and gives Eq. (41) when restricted to the elements of $\mathcal{L}^2(X, \mathcal{C}, \mu)$.

Let V_1 and V_2 be two normed linear spaces, with inner products $\|\cdot\|_1$ and $\|\cdot\|_2$, respectively, and let $\mathcal{H}$ be a Hilbert space, with inner product $(\cdot, \cdot)$. A *bounded linear operator* from V_1 into V_2 is a map $\mathcal{T} : V_1 \to V_2$ such that (i) $\mathcal{T}(\alpha\mathbf{g} + \beta\mathbf{h}) = \alpha\mathcal{T}\mathbf{g} + \beta\mathcal{T}\mathbf{h}$ for all $\mathbf{g}, \mathbf{h} \in V_1$ and $\alpha, \beta \in \mathbb{R}$, and (ii) there exists a constant $\gamma \in \mathbb{R}$ such that $\|\mathcal{T}\mathbf{h}\|_2 \leq \gamma\|\mathbf{h}\|_1$ for all $\mathbf{h} \in V_1$. A bounded linear operator $\mathcal{T} : \mathcal{H} \to \mathcal{H}$ is called *self-adjoint* if $(\mathcal{T}\mathbf{g}, \mathbf{h}) = (\mathbf{g}, \mathcal{T}\mathbf{h})$ for all $\mathbf{g}, \mathbf{h} \in \mathcal{H}$, and is called *positive semidefinite* if $(\mathbf{h}, \mathcal{T}\mathbf{h}) \geq 0$ for all $\mathbf{h} \in \mathcal{H}$.

At the beginning of this subsection, the field of scalars for linear spaces V was taken to be the real numbers, and inner products were therefore defined to be real-valued. Thus the Hilbert spaces defined here are real Hilbert spaces. It is also possible, of course, to define complex Hilbert spaces. One property that is lost by restricting attention in this chapter to real Hilbert spaces is that, while every positive semidefinite operator on a complex Hilbert space is self-adjoint, a positive semidefinite operator on a real Hilbert space need not be self-adjoint (e.g. Reed and Simon (1972, p. 195)). Covariance operators on a real Hilbert space are necessarily self-adjoint as well as positive semidefinite, however, as discussed in Appendix A.3.

References

[1] Coddington, E. A., and N. Levinson, 1955: *Theory of Ordinary Differential Equations*, McGraw-Hill.

[2] Cohn, S. E., 2008: Energetic consistency and coupling of the mean and covariance dynamics. Pp. 443–478 in *Handbook of Numerical Analysis, Vol. XIV* (P. G. Ciarlet, ed.), *Special Volume: Computational Methods for the Atmosphere and the Oceans*, R. M. Temam and J. J. Tribbia (guest eds.), Elsevier.

[3] Courant, R., and D. Hilbert, 1962: *Methods of Mathematical Physics, Vol. II: Partial Differential Equations*, Wiley-Interscience.

[4] Doob, J. L., 1953: *Stochastic Processes*, Wiley.

[5] Itô, K., 1984: *Foundations of Stochastic Differential Equations in Infinite Dimensional Spaces*, CBMS-NSF Regional Conference Series in Applied Mathematics, Vol. 47, Society for Industrial and Applied Mathematics.

[6] Kallianpur, G., and J. Xiong, 1995: *Stochastic Differential Equations in Infinite Dimensional Spaces*, Lecture Notes-Monograph Series, Vol. 26, Institute of Mathematical Statistics.

[7] Kreiss, H.-O., and J. Lorenz, 1989: *Initial-Boundary Value Problems and the Navier-Stokes Equations*, Academic Press.

[8] Lax, P. D., 1973: *Hyperbolic Systems of Conservation Laws and the Mathematical Theory of Shock Waves*, CBMS-NSF Regional Conference Series in Applied Mathematics, Vol. 11, Society for Industrial and Applied Mathematics.

[9] Lax, P. D., 2006: *Hyperbolic Partial Differential Equations*, Courant Lecture Notes in Mathematics, Vol. 14, American Mathematical Society.

[10] Omatu, S., and J. H. Seinfeld, 1989: *Distributed Parameter Systems: Theory and Applications*, Oxford University Press.

[11] Reed, M., and B. Simon, 1972: *Methods of Modern Mathematical Physics, Vol. I: Functional Analysis*, Academic Press.

[12] Riesz, F., and B. Sz.-Nagy, 1955: *Functional Analysis*, Frederick Ungar.

[13] Royden, H. L., 1968: *Real Analysis*, 2nd ed., Macmillan.

[14] Rudin, W., 1991: *Functional Analysis*, 2nd ed., McGraw-Hill.