An Novel Algorithm to Solve the Nonlinear Filtering Problems in Real-Time

BY

XUE LUO
B.A. (East China Normal University) 2004
M.S. (University of Illinois at Chicago) 2011

THESIS

Submitted in partial fulfillment of the requirements
for the degree of Doctor of Philosophy in Mathematics
in the Graduate College of the
University of Illinois at Chicago, 2013

Chicago, Illinois

Defense Committee:

Stephen Yau, Chair and Advisor
Charles Knessl
Jan Verschelde
Jie Yang
Lixing Jia, Chicago State University

To my father Qicheng Luo, my mother Xueli Zhang,

my husband Dr. Huaiqing Zuo and my son Eric J. Zuo.

ACKNOWLEDGMENTS

Without constant support and encouragement, this work would be a "Mission Impossible".

I'm indebted to my supervisor Stephen S.-T. Yau for his guidance and support. He introduced me a brand new interesting field to working in. He was always available for help despite his busy schedule. I'm truely grateful.

My warmest thanks also go to Roman Shvydkoy, who always treat me as his own student and support me financially as his RA during my last year in graduate college so that I could concentrate on my research. I'm benefit a lot from his sharing of mathematical experience and his encouragement.

I would also like to express my gratitude to all the colleagues and the faculty in department of Mathematics, Statistics and Computer Sciences, University of Illinois at Chicago for their intriguing discussions. And I thanks all the staff for their diligent work on our daily routines.

Last but not the least, I owe much to my loving family: my husband, my son and my parents. Without their care and love, it is no way to bring my efforts to fruition.

TABLE OF CONTENTS

TABLE OF CONTENTS (Continued)

CHAPTER **PAGE**

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS

ASME	American Society of Mechanical Engineers
DMZ	Duncan-Mortensen-Zakai
EKF	Extended Kalman filter
FKE	Forward Kolmogorov equation
HC	Hyperbolic cross
HSM	Hermite spectral method
KF	Kalman filter
NLF	Nonlinear filtering
ODE	Ordinary differential equation
OHC	Optimized hyperbolic cross
PF	Particle filters
PDE	Partial differential equation
RHC	Regular hyperbolic cross

SUMMARY

This dissertation provides an affirmative answer to the well-known half century old engineering question raised by Office of Naval Research: How can one solve nonlinear filtering (NLF) problems in real time without memory, if enough computational resources are provided? Instead of the prestigious Kalman filter (KF) and its derivatives to estimate the mean and the covariance matrix of the states, we resort to solving the Duncan-Mortensen-Zakai (DMZ) equation, which is satisfied by the un-normalized probability density function of the states. In this dissertation, we develop a novel algorithm, which is applicable to the most general settings of the NLF problems and keeps two of the most important properties of KF: real-time and memory-less.

Briefly speaking, in our algorithm, we split the approximation of the conditional density function into two parts: one part could be pre-computed before any on-line experiments ran (so-called off-line computation); the other part has to be sychronized the real-time data with the pre-computed data (so-called on- line computation). More precisely, the off-line computation solves a forward Kolmogorov equation (FKE) with the initial conditions, which are chosen to be a complete base functions in square-integrable function space, while the on-line part computes the projection of the conditional density function at each time step onto the basis, and then synchronize them with the off-line data to obtain the conditional density function at the next time step.

First, we validate our algorithm theoretically, by estimating the convergence rate with respect to the sampling frequency. Second, we tackle some difficulties in the implementation of our algorithm and apply it to some 1-D benchmark NLF problems. Compared with the two most widely used methods nowadays, extended Kalman filter and particle filters, our algorithm surpasses both of them in the real-time manner with comparable accuracy. Last, when we investigate the application of our algorithm to the high-dimensional state NLF problems, we combine the sparse grid algorithm with the Hermite spectral method to serve as the off-line solver of FKE. The convergence rate is investigated both theoretically and numerically.

CHAPTER 1

INTRODUCTION

Tracing back to 1960s, two most influential mathematics papers (36), (34) have been published in ASME's Journal of Basic Engineering. These are so-called Kalman filter (KF) and Kalman-Bucy filter. They addressed a significant question: How does one get accurate estimate from noisy data? The applications of KF are endless, from seismology to bioengineering to econometrics. The KF surpasses the other existing filtering in, at least, the following two aspects:

- The KF uses each new obervation to update the state of the system without refering back to any earlier observations. This is so-called "memory-less" or "without memory".
- The KF makes the decisions of the state instantaneously, while the observation data keep coming in. This property is called "real-time".

Despite its success in many real applications, the limitations on the nonlinearity and Gaussian assumption of the initial probability density of the KF pushed mathematicians and engineers to seek the optimal nonlinear filtering (NLF). One direction is to adapt the KF to the nonlinearities. The researchers developed extended Kalman filter (EKF), unscented Kalman filter, ensemble Kalman filter, etc., which could be used to handle weak nonlinearities (say, the almost linear ones). But for serious nonlinearities, they may completely fail.

Another direction, and also the most popular method nowadays, are the particle filters (PF), refer to as (1), (2) and reference therein. They are developed from sequential Monte Carlo methods. On the one hand, the PF are applicable to the nonlinear, non-Gaussian scenarios. As the number of particles goes to infinity, the PF become asymptotically optimal. On the other hand, they are hard to implement as a real-time application, due to its essence of Monte Carlo simulations.

Besides the two widely used methods above, the partial differential equations (PDE) methods are introduced to the NLF in 1960s. These methods are based on the fact that the unnormalized conditional density of the states is the solution of Duncan-Mortensen-Zakai (DMZ) equation, refer to as (15), (41) and (60). Yet, the main drawback of the PDE methods are the intensive computation. It is almost impossible to achieve the "real-time" performance. To overcome this shortcoming, the splitting-up algorithm is introduced to move the heavy computation off-line. It is like the Trotter product formula from semigroup theory. This operator-splitting algorithm is proposed for the DMZ equation by Bensoussan, Glowinski, and Rascanu (10). More research articles which follow this direction are (28), (42) and (31), etc. In 1990s, Lototsky, Mikulevicius and Rozovskii (38) developed a new algorithm (so-called S^3-algorithm) based on the Cameron-Martin version of Wiener chaos expansion. Unfortunately, it is pointed out in (10) that the soundness of these algorithms are verified only to the filtering with bounded drift and observation terms (i.e., f and h in Equation 2.1). In 2008, Yau and Yau (59) developed a novel algorithm to solve the "pathwise-robust" DMZ equation (see Equation 2.6), where the boundedness conditions are weakened to some mild growth conditions on f and h. The two nice

properties of the KF have also been kept in this algorithm: "memory-less" and "real-time". But their algorithm has only been rigorously proved in theory, when the drift term, the observation term (f, h in Equation 2.1) are not explicitly time-dependent, the variance of the noises (G in Equation 2.1) is the identity matrix, and the noises are standard Brownian motion processes ($S = I_{r\times r}$, $Q = I_{m\times m}$ in Equation 2.1) . Let us refer to the case studied in (59) as the "time-invariant" one in this dissertation.

The first task of this dissertation is to extend the algorithm in (59) to the most general settings of NLF problems, in the sense that the drift term, the observation term could explicitly depend on time, the variance of the noises S, Q are also time-dependent, and G could be a matrix of functions of both time and the states. We shall validate our algorithm under very mild growth conditions on f, h and G, for example Equation 3.30, Equation 4.2 and Equation 3.45, etc. These are essentially time-dependent analogue of those in (59). First of all, this extension is absolutely necessary. Many real applications have explicit time-dependence in their models, say the target orientation angle estimation for target position/velocity in constant turn model, where the angular velocities are piecewise constant functions in time, see (46). Second, this extension is nontrivial from the mathematical point of view. A trickier analysis of the PDE is required. For instance, we need to take care of the more general elliptic operator D_w^2, see the definition in Equation 2.7, rather than the Laplacian.

The second task in this dissertation is to implement our algorithm with 1-D state to achieve the "real-time" and "memory-less" manner. In the implementation, we shall solve the forward Kolmogorov equation (FKE) Equation 2.13 with the Hermite spectral method (HSM). There

are two reasons that we choose HSM: on the one hand, HSM is particularly suitable for functions defined on the unbounded domain which decays exponentially at infinity; on the other hand, HSM could be easily patched with the numerical solution obtained in the previous time step while the moving-window technique is in use in the on-line experiments.

The HSM itself is also a rich research field, which could be traced back to 1970s. In (21), Gottlieb et. al. gave the example $\sin x$ to illustrate the poor resolution of Hermite polynomials. To resolve M wavelength of $\sin x$, it requires nearly M^2 Hermite polynomials. Due to this fact, they doubted the usefulness of Hermite polynomials as basis. The Hermite functions inherit the same deficiency from the polynomials. Moreover, it is lack of fast Hermite transform (some analogue of fast Fourier transform). Despite of all these drawbacks, the HSM has its inherent strength. Many physical models need to solve a differential equation on an unbounded domain, and the solution decays exponentially at infinity. From the computational point of view, it is hard to describe the rate of decay at infinity numerically or to impose some artificial boundary condition cleverly on some faraway "boundary". Therefore the Chebyshev or Fourier spectral methods are not so useful in this situation. As to the HSM, dealing with the behavior at infinity is not necessary. Recent applications of the HSM can be found in (17), (19), (27), (48), (57), etc.

To overcome the poor resolution, a scaling factor is necessary to be introduced into the Hermite functions, refer to (6), (7). It is shown in (7) that the scaling factor should be chosen according to the truncated modes N and the asymptotical behavior of the function $f(x)$, as $|x| \to \infty$. Some efforts have been made in seeking the suitable scaling factor, see (7), (55), (48),

etc. To optimize the scaling factor is still an open problem, even in the case that $f(x)$ is given explicitly, to say nothing of the exact solution to a differential equation, which is generally unknown a-priori. Although some investigations about the scaling factor have been made theoretically, as far as we know, there are no practical guidelines of choosing a suitable scaling factor. Nearly all the scaling factors in the papers with the application of HSM are obtained by the trial-and-error method. Thus, we believe it is necessary and useful to give a practical strategy to pick an appropriate scaling factor and the corresponding truncated mode for at least the most commonly used types of functions, i.e. the Gaussian type and the super-Gaussian type functions. The strategy we shall give in section 5.2.2 only depends on the asymptotic behavior of the function. In the scenario where the solution of some differential equation needs to be approximated (the exact solution is unknown), we could use asymptotical analysis to obtain its asymptotic behavior. Thus, our strategy of picking the suitable scaling factor is still applicable.

Let us draw our attention back to the implementation of our algorithm to NLF problems. Through our study of HSM to FKE, the off-line data could be well prepared. However, when synchronizing the off-line data with the on-line experiments, to be more specifically, updating the initial data on-line, another difficulty arises due to the drifting of the conditional density function. The untranslated Hermite functions with limited truncation modes could only resolve the function well, if it is concentrated in the neighborhood of the origin. Let us call this neighborhood as a "window". Unfortunately, the density function will probably drift out of the current "window". The numerical evidence is displayed in Figure 6. To efficiently solve this problem, we for the first time introduce the translating factor to the Hermite functions and

the moving-window technique for the on-line implementation. The translating factor helps the moving-window technique to be implemented more neatly and easily. The idea of the moving-window technique is to shift the windows back and forth according to the "support" of the density function, by tuning the translating factor. Three NLF problems are solved numerically by our algorithm and compared with either EKF or PF. It is verified numerically that our algorithm is superior to both of them in the "real-time" manner.

The last task of this dissertation is to investigate our algorithm with high-dimensional state. That is, we need to solve the FKE in $\mathbb{R}^d$, where $d > 1$ is the dimension of the state. Among the existing literature, the Hermite and Laguerre spectral methods are the commonly used approaches to solve the PDE based on orthogonal polynomials in infinite interval, referring to (19), (57). Although the Hermite spectral method (HSM) appears to be a natural choice, it is not as widely used as the Chebyshev and Fourier spectral methods, due to its poor resolution (21) and the lack of fast algorithm for the transformation (8). However, it is shown in (6) that an appropriately chosen scaling factor could greatly improve the resolution. Moreover, we present practical guidelines of choosing the suitable scaling factors for Gaussian/super-Gaussian functions (39).

Nevertheless, the main difficulty comes from the so-called "curse of dimensionality". Take the target tracking problem in 3-dim as example, there are at least six states involved in this system (three for position, three for velocity). That is, we need to solve a linear parabolic PDE in $\mathbb{R}^6$. Naively, if we implement the spectral method with tensor product formulation and assume that N modes need to be computed in each direction, then the total amount of the computation

is N^6. Even with moderately small N, it is still not within the reasonable computing capacity. An efficient tool to reduce this effect is the *sparse grids approximations* from Smolyak's algorithm (54), which is based on a hierarchy of one-dimensional quadrature. It has a potential to obtain higher rates of convergence than many existing methods, under certain regularity conditions. For example, the convergence rate of Monte Carlo simulations are $\mathcal{O}(N^{-\frac{1}{2}})$ with N sample points, while the sparse grids from (54) achieves $\mathcal{O}(N^{-r}(\log N)^{(d-1)(r+1)})$, provided that the function has bounded mixed derivatives of order r. The studies of sparse grids start from the basis functions in the physical spaces: piecewise linear multiscale bases (13), wavelets (13), (49). In the recent decade, the hyperbolic cross (HC) approximation in the frequency space has also been investigated with various basis functions: Fourier series (22), (24), polynomial approximations generated from the Chebyshev-Gauss-Lobatto points (4), Jacobi polynomials (51).

Although the regular hyperbolic cross (RHC) approximation reduce the effect of the "curse of dimensionality" in some degree, the convergence rate still deteriorates slowly with the dimension increasing (noting the term $(\log N)^{(d-1)(r+1)}$ in the previous paragraph). To completely break the "curse of dimensionality", the optimized hyperbolic cross (OHC) approximation is introduced in (24). It has been shown in (36) that the convergence rate of the OHC approximation with $\gamma \in (0, 1)$ (see definition in Equation 6.40) with Fourier series is of $\mathcal{O}(N^{-r})$ in our notation, where the dimension enters the big-$\mathcal{O}$. We shall first establish the error estimate for the HC approximations with properly scaled Hermite functions in the weighted Korobov spaces $\mathcal{K}^m_{\boldsymbol{\alpha},\boldsymbol{\beta}}(\mathbb{R}^d)$, see Equation 6.28. Next, we shall study the application of the Galerkin-type HSM

with the HC approximation to high-dimensional linear parabolic PDEs. The error estimates in appropriate weighted Korobov spaces are investigated under various conditions (cf. conditions $(\boldsymbol{C_1})$-$(\boldsymbol{C_6})$ in Chapter 6). There also exists a rich literature on the applications of sparse grids algorithms. It has already been successfully applied to problems from the integral equations (26), to interpolation and approximation (35), to the stochastic differential equations (50), (43), to high dimensional integration problems from physics and finance (20), and to the solutions to elliptic PDEs, (61), (52). As to the parabolic PDEs, they are treated with a wavelet-based sparse grid discretization in (56). Besides the finite element approaches, they are also handled wth finite differences on sparse grids (23) and finite volume schemes (29). Griebel and Oeltz (25) proposed a space-time sparse grid technique, where the tensor product of one-dimensional multilevel basis in time and a proper multilevel basis in space have been employed. To our best knowledge, it is the first time in this paper that the Galerkin HSM with sparse grids algorithm is applied to parabolic PDEs, and the error estimates are obtained in the appropriate spaces.

CHAPTER 2

MODEL AND ALGORITHM

The model we are considering is the signal observation model with explicit time-dependence in the drift term, observation term and the variance of the noises:

$$\begin{cases} dx_t = f(x_t, t)dt + G(x_t, t)dv_t, \\ \\ dy_t = h(x_t, t)dt + dw_t, \end{cases} \tag{2.1}$$

where x_t and f are n-vectors, G is an $n \times r$ matrix, and v_t is an r-vector Brownian motion process with $E[dv_t dv_t^T] = Q(t)dt$, y_t and h are m-vectors and w_t is an m-vector Brownian motion process with $E[dw_t dw_t^T] = S(t)dt$ and $S(t) > 0$. We refer to x_t as the state of the system at time t with some initial state x_0 (not necessarily obeying Gaussian distribution) and y_t as the observation at time t with $y_0 = 0$. We assume that $\{v_t, t \geq 0\}$, $\{w_t, t \geq 0\}$ and x_0 are independent. For the sake of convenience, let us call this system with explicit time-dependence in f, h and G the "time-varying" case, while the one without explicit time-dependence the "time-invariant" as those studied in (59).

Let us assume throughout the dissertation that f, h and G are smooth, say C^2 in space and C^1 in time. However, some growth conditions on f, h and G will be specified later in the study of "pathwise-robust" DMZ equation.

The unnormalized density function $\sigma(x,t)$ of x_t conditioned on the observation history $Y_t = \{y_s : 0 \leq s \leq t\}$ satisfies the DMZ equation (for the detailed formulation, see (15), (41) and (60))

$$\begin{cases} d\sigma(x,t) = L\sigma(x,t)dt + \sigma(x,t)h^T(x,t)S^{-1}(t)dy_t \\ \\ \sigma(x,0) = \sigma_0(x), \end{cases} \tag{2.2}$$

where $\sigma_0(x)$ is the probability density of the initial state x_0, and

$$L(*) \equiv \frac{1}{2}\sum_{i,j=1}^{n} \frac{\partial^2}{\partial x_i \partial x_j}\left[\left(GQG^T\right)_{ij} *\right] - \sum_{i=1}^{n} \frac{\partial (f_i *)}{\partial x_i}, \tag{2.3}$$

where $(GQG^T)_{ij}$ is the $(i,j)^{\text{th}}$ entry of the matrix and f_i is the i^{th} component of f.

We won't solve the DMZ equation directly, due to the following two reasons. On the one hand, the DMZ equation, i.e. Equation 2.2, is a stochastic partial differential equation due to the term dy_t. There is no easy way to derive a recursive algorithm to solve this equation. On the other hand, in real applications, one may be more interested in constructing robust state estimators from each observation path, instead of having certain statistical data of thousands of repeated experiments. Here, the robustness means our state esitmator is not sensitive to the observation path. This property is important, since in most of the real applications, the observation arrives and is processed at discrete moments in time. The state estimator is still expected to perform well based on the linear interpolation of the discrete observations, without the knowledge of the real continuous observation path.

For each "given" observation, making an invertible exponential transformation as in (47)

$$\sigma(x,t) = \exp\left[h^T(x,t)S^{-1}(t)y_t\right]\rho(x,t), \tag{2.4}$$

the DMZ equation is transformed into a deterministic partial differential equation (PDE) with stochastic coefficients, which we will refer as the "pathwise-robust" DMZ equation

$$\begin{cases} \dfrac{\partial\rho}{\partial t}(x,t) + \dfrac{\partial}{\partial t}(h^TS^{-1})^Ty_t\rho(x,t) = \exp\left(-h^TS^{-1}y_t\right)\left[L - \dfrac{1}{2}h^TS^{-1}h\right]\cdot\left[\exp\left(h^TS^{-1}y_t\right)\rho(x,t)\right] \\ \\ \qquad\qquad\qquad\qquad \rho(x,0) = \sigma_0(x). \end{cases} \tag{2.5}$$

Or equivalently,

$$\begin{cases} \dfrac{\partial\rho}{\partial t}(x,t) = \dfrac{1}{2}D_w^2\rho(x,t) + F(x,t)\cdot\nabla\rho(x,t) + J(x,t)\rho(x,t) \\ \\ \rho(x,0) = \sigma_0(x), \end{cases} \tag{2.6}$$

where

$$D_w^2 = \sum_{i,j=1}^{n}(GQG^T)_{ij}\frac{\partial^2}{\partial x_i\partial x_j}, \tag{2.7}$$

$$F(x,t) = \left[\sum_{j=1}^{n}\frac{\partial}{\partial x_j}\left(GQG^T\right)_{ij} + \sum_{j=1}^{n}(GQG^T)_{ij}\frac{\partial K}{\partial x_j} - f_i\right]_{i=1}^{n}, \tag{2.8}$$

$$\begin{aligned} J(x,t) = &-\frac{\partial}{\partial t}(h^TS^{-1})^Ty_t + \frac{1}{2}\sum_{i,j=1}^{n}\frac{\partial^2}{\partial x_i\partial x_j}\left(GQG^T\right)_{ij} + \sum_{i,j=1}^{n}\frac{\partial}{\partial x_i}\left(GQG^T\right)_{ij}\frac{\partial K}{\partial x_j} \\ &+ \frac{1}{2}\sum_{i,j=1}^{n}(GQG^T)_{ij}\left[\frac{\partial^2K}{\partial x_i\partial x_j} + \frac{\partial K}{\partial x_i}\frac{\partial K}{\partial x_j}\right] - \sum_{i=1}^{n}\frac{\partial f_i}{\partial x_i} - \sum_{i=1}^{n}f_i\frac{\partial K}{\partial x_i} - \frac{1}{2}(h^TS^{-1}h), \end{aligned} \tag{2.9}$$

in which

$$K(x,t) = h^T(x,t)S^{-1}(t)y_t. \tag{2.10}$$

The existence and uniqueness of the "pathwise-robust" DMZ equation under certain conditions has been investigated by Pardoux (44), (45), Fleming-Mitter (16), Baras-Blankenship-Hopkins (3) and Yau-Yau (58), (59). The well-posedness is shown in (44), when the drift term $f \in C^1$ and the observation term $h \in C^2$ are bounded. Later, in (45) Pardoux extended his result to where f, h have at most linear growth. Fleming and Mitter treated the case where f and ∇f are bounded. Baras, Blankenship and Hopkins (3) obtained the well-posedness result on the "pathwise-robust" DMZ equation with a class of unbounded coefficients only when the state is of 1-D. In the appendices of (59), Yau and Yau obtained the existence and uniqueness results in the weighted Sobolev space, where f and h satisfy some mild growth condition. However, there is a gap in their proof of existence (Theorem A.4, (59)). In Chapter 3 of this dissertation, we shall circumvent the gap by a more delicate analysis to give a time-dependent analogous well-posedness result to the "pathwise-robust" DMZ equation under some mild growth conditions on f, h and G.

2.1 Approximation in our algorithm

Let us assume that we know apriori the observation time sequence $\mathcal{P}_k := \{\tau_0 < \tau_1 < \cdots < \tau_k = T\}$. But the observation data $\{y_{\tau_i}\}$ at each sampling time τ_i, $i = 0, \cdots, k$ are unknown until the on-line experiment runs. We call the computation "off-line", if it can be performed

without any on-line experimental data (or say pre-computed); otherwise, it is called "on-line" computations. One only concerns the computational complexity of the on-line computations, since the "real-time" property hinges on it.

Let ρ_i be the solution of the "pathwise-robust" DMZ equation with $y_t = y_{\tau_{i-1}}$ on the interval $\tau_{i-1} \le t \le \tau_i$, $i = 1, 2, \cdots, k$,

$$\begin{cases} \dfrac{\partial \rho_i}{\partial t}(x,t) + \dfrac{\partial}{\partial t}\left(h^T S^{-1}\right)^T y_{\tau_{i-1}} \rho_i(x,t) \\ \qquad = \exp\left(-h^T S^{-1} y_{\tau_{i-1}}\right)\left[L - \dfrac{1}{2} h^T S^{-1} h\right] \cdot \left[\exp\left(h^T S^{-1} y_{\tau_{i-1}}\right) \rho_i(x,t)\right], \quad x \in \mathbb{R}^n \\ \rho_1(x,0) = \sigma_0(x), \\ \text{and} \\ \rho_i(x,\tau_{i-1}) = \rho_{i-1}(x,\tau_{i-1}), \quad \text{for } i = 2,3,\cdots,k. \end{cases} \tag{2.11}$$

That is, we freeze the observation data on the time interval $\tau_{i-1} \le t < \tau_i$ to be $y_{\tau_{i-1}}$. Define the norm of $\mathcal{P}_k$ to be $|\mathcal{P}_k| = \sup_{1 \le i \le k}(\tau_i - \tau_{i-1})$. Intuitively, as $|\mathcal{P}_k| \to 0$, we have

$$\sum_{i=1}^{k} \chi_{[\tau_{i-1},\tau_i]}(t) \rho_i(x,t) \to \rho(x,t)$$

in some sense, for all $0 \le t \le T$, where $\rho(x,t)$ is the exact solution of Equation 2.5. That is to say, intuitively, the denser the sampling time sequence is, the more accurate the approximate solution should be. This is the only approximation in our algorithm. Its convergence will be shown rigorously, and its convergence rate will be estimated, in Chapter 4.

2.2 Design of the algorithm

Remember that our aim is to develop a practical algorithm to solve NLF problems. Even though the convergence in our algorithm is shown rigorously, it is impractical to solve Equation 2.11 in the "real-time" manner, since the "on-line" data $\{y_{\tau_i}\}$, $i = 1, \cdots, k$, are contained in the coefficients of Equation 2.11. Therefore, we have to numerically solve the time-consuming PDE on-line, every time after the new observation data coming in. Yet, the proposition below helps to move the heavy computations off-line. This is the key ingredient of the algorithm in (59), and in ours.

Proposition 2.1. *For each $\tau_{i-1} \leq t < \tau_i$, $i = 1, 2, \cdots, k$, $\rho_i(x,t)$ satisfies Equation 2.11 if and only if*

$$u_i(x,t) = \exp\left[h^T(x,t)S^{-1}(t)y_{\tau_{i-1}}\right]\rho_i(x,t), \tag{2.12}$$

satisfies the forward Kolmogorov equation (FKE)

$$\frac{\partial u_i}{\partial t}(x,t) = \left(L - \frac{1}{2}h^T S^{-1} h\right) u_i(x,t), \tag{2.13}$$

where L is defined in Equation 2.3.

It is clear that Equation 2.13 is independent of the observation path $\{y_{\tau_i}\}_{i=0}^k$. So Equation 2.13 could be numerically solved beforehand. Let us denote the operator $\left(L - \frac{1}{2}h^T S^{-1} h\right)$ as $U(t)$ for short. As to the "time-varying" case, our algorithm still maintain the "real-time" property. $\{U(t)\}_{t\in[0,T]}$ forms a family of strong elliptic operators. Furthermore, the operator $U(t) : D(U(t)) \subset L^2(\mathbb{R}^n) \to L^2(\mathbb{R}^n)$ is the infinitesimal generator of the two-parameter semi-

groups $\mathcal{U}(t,\tau)$, for $t \geq \tau$. In particular, with the known observation time sequence $\{\tau_i\}_{i=1}^k$, we obtain a sequence of two-parameter semigroup $\{\mathcal{U}(t,\tau_{i-1})\}_{i=1}^k$, for $\tau_{i-1} \leq t < \tau_i$. Let us take the initial conditions of FKE Equation 2.13 at $t = \tau_i$ to be one member of a set of complete orthonormal base in $L^2(\mathbb{R}^n)$, say $\{\phi_l(x)\}_{l=1}^\infty$. We pre-compute the solutions of Equation 2.13 at time $t = \tau_{i+1}$, denoted as $\{\mathcal{U}(\tau_{i+1},\tau_i)\phi_l\}_{l=1}^\infty$. These data should be stored in preparation of the on-line computations. The only price to pay is that the "time-varying" case requires more storage capacity, since $\{\mathcal{U}(\tau_{i+1},\tau_i)\phi_l\}_{l=1}^\infty$ differs from each τ_i, $i = 1, \cdots, k$, and all of them need to be stored. In general, the longer the simulation time is, the more storage it requires in the "time-varying" case. While the storage of the data is independent of the simulation time in the "time-invariant" case. Nevertheless, it won't affect the off-line virture of our algorithm.

The on-line computation in our algorithm consists of two parts at each time step τ_{i-1}, $i = 1, \cdots, k$, as described below.

- Project the initial condition $u_i(x,\tau_{i-1}) \in L^2(\mathbb{R}^n)$ at $t = \tau_{i-1}$ onto the base $\{\phi_l(x)\}_{l=1}^\infty$, i.e., $u_i(x,\tau_{i-1}) = \sum_{l=1}^\infty \hat{u}_{i,l}\phi_l(x)$. Hence, the solution to Equation 2.13 at $t = \tau_i$ can be expressed as

$$u_i(x,\tau_i) = \mathcal{U}(\tau_i,\tau_{i-1})u_i(x,\tau_{i-1}) = \sum_{l=1}^{\infty} \hat{u}_{i,l}\left[\mathcal{U}(\tau_i,\tau_{i-1})\phi_l(x)\right], \tag{2.14}$$

  where $\{\mathcal{U}(\tau_i,\tau_{i-1})\phi_l(x)\}_{l=1}^\infty$ have already been computed off-line.

- Update the initial condition of Equation 2.13 at τ_i with the new observation y_{τ_i}. Let us specify the observation updates (the initial condition of Equation 2.13) for each time

step. For $0 \leq t \leq \tau_1$, the initial condition is $u_1(x,0) = \sigma_0(x)$. At time $t = \tau_1$, when the observation y_{τ_1} is available,

$$u_2(x,\tau_1) = \exp\left[h^T(x,\tau_1)S^{-1}(\tau_1)y_{\tau_1}\right]\rho_2(x,\tau_1) = \exp\left[h^T(x,\tau_1)S^{-1}(\tau_1)y_{\tau_1}\right]u_1(x,\tau_1),$$

in view of the fact that $y_0 = 0$, by Equation 2.12 and Equation 2.11. Here, $u_1(x,\tau_1) = \sum_{l=1}^{\infty} \hat{u}_{1,l}\left[\mathcal{U}(\tau_1,0)\phi_l(x)\right]$, where $\{\hat{u}_{1,l}\}_{l=1}^{\infty}$ is computed in the previous step, and $\{\mathcal{U}(\tau_1,0)\phi_l(x)\}_{l=1}^{\infty}$ are prepared by off-line computations. Hence, we obtain the initial condition $u_2(x,\tau_1)$ of Equation 2.13 for the next time interval $\tau_1 \leq t \leq \tau_2$. Recursively, the initial condition of Equation 2.13 for $\tau_{i-1} \leq t \leq \tau_i$ is

$$u_i(x,\tau_{i-1}) = \exp\left[h^T(x,\tau_{i-1})S^{-1}(\tau_{i-1})(y_{\tau_{i-1}} - y_{\tau_{i-2}})\right] \cdot u_{i-1}(x,\tau_{i-1}), \tag{2.15}$$

for $i = 2, 3, \cdots, k$, where $u_{i-1}(x,\tau_{i-1}) = \sum_{l=1}^{\infty} \hat{u}_{i-2,l}\left[\mathcal{U}(\tau_{i-1},\tau_{i-2})\phi_l(x)\right]$.

The approximation of $\rho(x,t)$, denoted as $\hat{\rho}(x,t)$, is obtained

$$\hat{\rho}(x,t) = \sum_{i=1}^{k} \chi_{[\tau_{i-1},\tau_i]}(t)\rho_i(x,t), \tag{2.16}$$

where $\rho_i(x,t)$ is obtained from $u_i(x,t)$ by Equation 2.12. Then $\sigma(x,t)$ can be recovered from Equation 2.4.

CHAPTER 3

STUDY OF THE "PATHWISE-ROBUST" DMZ EQUATION

3.1 Notations

Let $\mathbb{Q}_T = \mathbb{R}^n \times [0,T]$. Let $H^1(\mathbb{R}^n)$ be the Sobolev space, equipped with the norm

$$||u(x)||_1^2 = \int_{\mathbb{R}^n} (u^2 + |\nabla_x u|^2) dx.$$

And let $H^{1;1}(\mathbb{Q}_T)$ be the functional space of both t and x, with the norm

$$||v(x,t)||_{1;1}^2 = \int_{\mathbb{Q}_T} (v^2 + |\nabla_x v|^2 + |\partial_t v|^2) dx dt.$$

Let $H_0^{1;1}(\mathbb{Q}_T)$ denote the subspace of $H^{1;1}(\mathbb{Q}_T)$ consisting of functions $v(x,t)$ which have compact support in $\mathbb{R}^n$ for any t.

Definition 3.1. *The function $u(x,t)$ in $H_0^{1;1}(\mathbb{Q}_T)$ is called a weak solution of the initial value problem*

$$\begin{cases} \displaystyle\sum_{i,j=1}^{n} \frac{\partial}{\partial x_i}\left(A_{ij}(x,t)\frac{\partial u}{\partial x_j}\right) + \sum_{i=1}^{n} B_i(x,t)\frac{\partial u}{\partial x_i} + C(x,t)u = \frac{\partial u}{\partial t}, \\ u(x,0) = u_0(x) \end{cases}$$

if for any function $\Phi(x,t) \in H_0^{1;1}(\mathbb{Q}_T)$ *the following relation holds:*

$$\int\int_{\mathbb{Q}_T} \left[\sum_{i,j=1}^{n} A_{ij} \frac{\partial u}{\partial x_i} \frac{\partial \Phi}{\partial x_j} - \left(\sum_{i=1}^{n} B_i \frac{\partial u}{\partial x_i} + Cu + \frac{\partial u}{\partial t} \right) \Phi \right] dxdt = 0$$

and $u(x,0) = u_0(x)$.

We assume that the following conditions hold throughout the dissertation:

Condition 1. *The operator* L *defined in Equation 2.3 is a strong elliptic operator and it is bounded from above on* $\mathbb{Q}_T$. *That is, there exists a constant* $\lambda > 0$ *such that*

$$\lambda|\xi|^2 \leq \sum_{i,j=1}^{n} (GQG^T)_{ij} \xi_i \xi_j,$$

for any $(x,t) \in \mathbb{Q}_T$, *for any* $\xi = (\xi_1, \xi_2, \cdots, \xi_n) \in \mathbb{R}^n$. *And*

$$||GQG^T||_\infty = \sup_{(x,t)\in\mathbb{Q}_T} |GQG^T|_\infty < \infty,$$

where $|\cdot|_\infty$ *is the sup-norm of the matrix.*

Condition 2. *The initial density function* $\sigma_0(x) \in H^1(\mathbb{R}^n)$ *decays fast enough. To be more specific, we require that*

$$\int_{\mathbb{R}^n} e^{\sqrt{1+|x|^2}} \sigma_0(x) dx < \infty.$$

3.2 Well-posedness of the "pathwise-robust" DMZ equation

Before we show the existence of the weak solution, we shall give a priori estimations of up to the first order derivative of the solution to the "pathwise-robust" DMZ equation on $\mathbb{Q}_R := B_R \times [0,T]$, where $B_R = \{x \in \mathbb{R}^n : |x| \leq R\}$ is a ball of radius R.

Theorem 3.1. *Let ρ_R be the solution to the "pathwise-robust" DMZ equation on $\mathbb{Q}_R$, i.e.*

$$\begin{cases} \dfrac{\partial \rho_R}{\partial t}(x,t) = \dfrac{1}{2} D_w^2 \rho_R(x,t) + F(x,t)\nabla \rho_R(x,t) + J(x,t)\rho_R(x,t), \quad x \in B_R \\ \rho_R(x,0) = \sigma_{0,B_R}(x) \\ \rho_R(x,t) = 0 \qquad \text{for } (x,t) \in \partial B_R \times [0,T], \end{cases} \tag{3.1}$$

where D_w^2, $F(x,t)$ and $J(x,t)$ are defined in Equation 2.7, Equation 2.8 and Equation 2.9, respectively, and $\sigma_{0,\Omega}$ is defined as

$$\sigma_{0,\Omega}(x) = \begin{cases} \sigma_0(x), & x \in \Omega_\epsilon \\ \text{smooth}, & x \in \Omega \setminus \Omega_\epsilon \\ 0, & x \in \mathbb{R}^n \setminus \Omega, \end{cases} \tag{3.2}$$

in which $\Omega_\epsilon = \{x \in \Omega : \ \text{dist}(x, \partial\Omega) > \epsilon\}$ and here $\Omega = B_R$. Assume that

$$\left\| \frac{\partial}{\partial t}(GQG^T) \right\|_\infty < \infty, \tag{3.3}$$

for all $t \in [0,T]$. Suppose there exists a positive function $g(x)$ on $\mathbb{R}^n$ such that for all $t \in [0,T]$, g and $\tilde{g} \triangleq g + \log|D_w J|$ satisfy

1. $\left|D_w g + \frac{1}{2}\nabla(GQG^T) - F\right|^2 + 2\lambda_1 J \leq C,$ *(3)*

2. $D_w^2 g + 2D_w g \cdot \nabla g + 2[\nabla(GQG^T) - F] \cdot \nabla g + \frac{1}{2}\nabla^2(GQG^T) - \mathrm{div}F + J \leq C,$ *(4)*

3. $D_w^2 \tilde{g} + 2D_w \tilde{g} \cdot \nabla \tilde{g} + 2[\nabla(GQG^T) - F] \cdot \nabla \tilde{g} + \frac{1}{2}\nabla^2(GQG^T) - \mathrm{div}F + J \leq C,$ *(5)*

4. $\int_{\mathbb{R}^n} e^{2\tilde{g}}\sigma_0^2(x) \leq C \quad \text{and} \quad \int_{\mathbb{R}^n} e^{2g} D_w\sigma_0 \cdot \nabla\sigma_0 \leq C,$ *(6)*

where C is a generic constant, which may differ from line to line, and $\nabla(*) = \left[\sum_{i=1}^n \frac{\partial(*)_{ij}}{\partial x_i}\right]_{j=1}^n$, $\nabla^2(*) = \sum_{i,j=1}^n \frac{\partial^2(*)_{ij}}{\partial x_i \partial x_j}$. Then, for $0 \leq t \leq T$,

$$\int_{B_R} e^{2g}\rho_R^2(x,t)dx \leq e^{Ct}\int_{B_R} e^{2g}\sigma_0^2(x)dx, \tag{3.4}$$

$$\int_{B_R} e^{2g} D_w\rho_R(x,t) \cdot \nabla\rho_R(x,t)dx \leq e^{Ct}\int_{B_R} e^{2g} D_w\sigma_0(x) \cdot \nabla\sigma_0(x)dx + Ce^{Ct}\int_{B_R} e^{2\tilde{g}}\sigma_0^2(x)dx, \tag{3.5}$$

where

$$D_w * = \left[\sum_{j=1}^n \left(GQG^T\right)_{ij}(x,t)\frac{\partial(*)}{\partial x_j}\right]_{i=1}^n, \tag{3.6}$$

$F(x,t)$ and $J(x,t)$ are defined in Equation 2.8 and Equation 2.9, respectively.

Remark 3.1. *The conditions in Theorem 3.1 could be easily checked, if the drift terms $h(x)$ and $f(x)$ are at most polynomial growth in $r = |x|$ and $Q = I_{r\times r}$. However, in general, the choice of such g in Theorem 3.1 is not clear.*

Proof. Let g be some positive function on $\mathbb{R}^n$.

$$\frac{d}{dt}\int_{B_R} e^{2g}\rho_R^2 = \int_{B_R} e^{2g}\rho_R D_w^2 \rho_R + 2\int_{B_R} e^{2g}\rho_R (F\cdot\rho_R) + 2\int_{B_R} e^{2g} J\rho_R^2 \triangleq \mathrm{I} + \mathrm{II} + \mathrm{III}. \tag{3.7}$$

Applying integration by parts to I and II in Equation 3.7 yields

$$\begin{aligned}\mathrm{I} =& -2\int_{B_R} \rho_R e^{2g} D_w g\cdot\nabla\rho_R - \int_{B_R} e^{2g} D_w\rho_R\cdot\nabla\rho_R - \int_{B_R} e^{2g}\rho_R\nabla(GQG^T)\cdot\nabla\rho_R \\ \leq& -2\int_{B_R} \rho_R e^{2g} D_w g\cdot\nabla\rho_R - \int_{B_R} e^{2g}\rho_R\nabla(GQG^T)\cdot\nabla\rho_R \triangleq \mathrm{I}_1 + \mathrm{I}_2.\end{aligned}$$

Integrating by parts further, we have

$$\begin{aligned}\mathrm{I}_1 =& 4\int_{B_R} e^{2g}\rho_R^2 D_w g\cdot\nabla g + 2\int_{B_R} e^{2g}\rho_R D_w g\cdot\nabla\rho_R \\ &+ 2\int_{B_R} e^{2g}\rho_R^2\nabla(GQG^T)\cdot\nabla g + 2\int_{B_R} e^{2g}\rho_R^2 D_w^2 g.\end{aligned} \tag{3.8}$$

Noting that the second term of the right-hand side of Equation 3.8 is $-\mathrm{I}_1$, we have

$$\mathrm{I}_1 = 2\int_{B_R} e^{2g}\rho_R^2 D_w g\cdot\nabla g + \int_{B_R} e^{2g}\rho_R^2[\nabla(GQG^T)\cdot\nabla g + D_w^2 g]. \tag{3.9}$$

The similar argument applied to I_2 gives us

$$\mathrm{I}_2 = \int_{B_R} e^{2g}\rho_R^2\nabla(GQG^T)\cdot\nabla g + \frac{1}{2}\int_{B_R} e^{2g}\rho_R^2\nabla^2(GQG^T). \tag{3.10}$$

Thus,

$$\mathrm{I} \leq \int_{B_R} e^{2g}\rho_R^2 \left[D_w^2 g + 2D_w g \cdot \nabla g + 2\nabla(GQG^T)\cdot\nabla g + \frac{1}{2}\nabla^2(GQG^T)\right]. \tag{3.11}$$

The same trick of I_1 applies to II in Equation 3.7, we obtain

$$\mathrm{II} = -\int_{B_R} e^{2g}\rho_R^2[2F\cdot\nabla g + \mathrm{div}F]. \tag{3.12}$$

Substitute Equation 3.11 and Equation 3.12 back to Equation 3.7, one gets

$$\begin{aligned}\frac{d}{dt}\int_{B_R} e^{2g}\rho_R^2 \leq & \int_{B_R} e^{2g}\rho_R^2 \left\{ D_w^2 g + 2D_w g\cdot\nabla g + 2[\nabla(GQG^T) - F]\cdot\nabla g \right. \\ & \left. + \frac{1}{2}\nabla^2(GQG^T) - \mathrm{div}F + J \right\} \\ \leq & C\int_{B_R} e^{2g}\rho_R^2,\end{aligned}$$

by Condition (4). Equation 3.4 follows directly from Gronwall's inequality. To show Equation 3.5, we consider

$$\begin{aligned}&\frac{d}{dt}\int_{B_R} e^{2g} D_w\rho_R\cdot\nabla\rho_R \\ =&\int_{B_R} e^{2g}\sum_{i,j=1}^{n}\frac{\partial}{\partial t}(GQG^T)_{ij}\frac{\partial\rho_R}{\partial x_i}\frac{\partial\rho_R}{\partial x_j} + 2\int_{B_R} e^{2g}\sum_{i,j=1}^{n}(GQG^T)_{ij}\frac{\partial}{\partial x_i}\left(\frac{\partial\rho_R}{\partial t}\right)\frac{\partial\rho_R}{\partial x_j} \\ \triangleq&\mathrm{IV} + \mathrm{V}.\end{aligned} \tag{3.13}$$

Due to Equation 3.3, IV in Equation 3.13 turns out to be

$$\begin{aligned}\text{IV} \leq & \frac{1}{2}\left\|\frac{\partial}{\partial t}(GQG^T)\right\|_\infty \int_{B_R} e^{2g} \sum_{i,j=1}^{n}\left[\left(\frac{\partial \rho_R}{\partial x_i}\right)^2+\left(\frac{\partial \rho_R}{\partial x_j}\right)^2\right] = n\left\|\frac{\partial}{\partial t}(GQG^T)\right\|_\infty \int_{B_R} e^{2g}|\nabla \rho_R|^2 \\ \leq & \frac{n}{\lambda_1}\left\|\frac{\partial}{\partial t}(GQG^T)\right\|_\infty \int_{B_R} e^{2g} D_w\rho_R \cdot \nabla \rho_R, \end{aligned} \tag{3.14}$$

since $D_w\rho_R \cdot \nabla \rho_R \geq \lambda_1 |\nabla \rho_R|^2$. Next, V in Equation 3.13 is

$$\begin{aligned}\text{V} = & -2\int_{B_R} e^{2g}[(2D_w g + \nabla(GQG^T)) \cdot \nabla \rho_R + D_w^2 \rho_R] \cdot \left(\frac{1}{2}D_w^2\rho_R + F \cdot \nabla \rho_R + J\rho_R\right) \\ = & -\int_{B_R} e^{2g}\left\{D_w^2\rho_R + \left[D_w g + \frac{1}{2}\nabla(GQG^T) + F\right] \cdot \nabla \rho_R\right\}^2 \\ & + \int_{B_R} e^{2g}\left[D_w g + \frac{1}{2}\nabla(GQG^T) - F\right]^2 |\nabla \rho_R|^2 \\ & - 2\int_{B_R} e^{2g}[D_w^2\rho_R + (2D_w g + \nabla(GQG^T)) \cdot \nabla \rho_R] J\rho_R \\ \leq & \int_{B_R} e^{2g}\left[D_w g + \frac{1}{2}\nabla(GQG^T) - F\right]^2 |\nabla \rho_R|^2 \\ & - 2\int_{B_R} e^{2g}[D_w^2\rho_R + (2D_w g + \nabla(GQG^T)) \cdot \nabla \rho_R] J\rho_R. \end{aligned} \tag{3.15}$$

Notice that

$$\begin{aligned}\int_{B_R} e^{2g} D_w^2\rho_R J\rho_R = & -\int_{B_R} e^{2g}\left[2(D_w g \cdot \nabla \rho_R)J\rho_R + JD_w\rho_R \cdot \nabla \rho_R\right. \\ & \left. + (D_w\rho_R \cdot \nabla J)\rho_R + \nabla(GQG^T) \cdot \nabla \rho_R J\rho_R\right]. \end{aligned} \tag{3.16}$$

Taking Equation 3.16 into account, V becomes

$$\mathrm{V} \leq \int_{B_R} e^{2g} \cdot \left\{ \frac{1}{\lambda_1} \left\{ \left[D_w g + \frac{1}{2}\nabla(GQG^T) - F \right]^2 + 1 \right\} + 2J \right\} D_w\rho_R \cdot \nabla\rho_R + \int_{B_R} e^{2g} |D_w J|^2 \rho_R^2. \tag{3.17}$$

Combining Equation 3.14 and Equation 3.17, we have

$$\begin{aligned} &\frac{d}{dt}\int_{B_R} e^{2g} D_w\rho_R \cdot \nabla\rho_R \\ &\leq \int_{B_R} e^{2g} \left\{ \frac{1}{\lambda_1} \left\{ n \left\| \frac{\partial}{\partial t}(GQG^T) \right\|_\infty + \left[D_w g + \frac{1}{2}\nabla(GQG^T) - F \right]^2 + 1 \right\} + 2J \right\} D_w\rho_R \cdot \nabla\rho_R \\ &\quad + \int_{B_R} e^{2g} |D_w J|^2 \rho_R^2. \end{aligned} \tag{3.18}$$

By Conditions (3) - (6), Equation 3.5 follows immediately. □

Theorem 3.2 (Existence)**.** *Under the conditions **1 - 2**, Equation 3.3 and Condition* (3) - (6) *in Theorem 3.1, the "pathwise-robust" DMZ equation on $\mathbb{Q}_T$ with the initial value $\sigma_0 \in H^1(\mathbb{R}^n)$ admits a non-negative weak solution $\rho \in H^{1;1}(\mathbb{Q}_T)$.*

Proof. Let R_k be a sequence of positive numbers such that $\lim_{k\to\infty} R_k = \infty$. Let $\rho_k(x,t)$ be the solution of the "pathwise-robust" DMZ equation on $\mathbb{Q}_{R_k}$, i.e. Equation 3.1 with $R = R_k$. In view of Theorem 3.1, the sequence $\{\rho_k\}$ is a bounded set in $H_0^{1;1}(\mathbb{Q}_{R_k})$. Thus, there exists a subsequence $\{\rho_{k'}\}$ which is weakly convergent to ρ. Moreover, ρ has the weak derivative

$\frac{\partial \rho}{\partial x_i} \in L^2(\mathbb{Q}_{R_k})$, and $\frac{\partial \rho_{k'}}{\partial x_i}$ weakly tends to it. Now we claim that the weak derivative $\frac{\partial \rho}{\partial t}$ exists. To see this, let $\Phi(x,t) \in H_0^{1;1}(\mathbb{Q}_{R_k})$, then

$$\begin{aligned}
&\iint_{\mathbb{Q}_{R_k}} \frac{1}{2}\sum_{i,j=1}^{n}(GQG^T)_{ij}\frac{\partial \Phi}{\partial x_j}\frac{\partial \rho}{\partial x_i} + \left[\sum_{i=1}^{n}\left(\sum_{j=1}^{n}\frac{\partial (GQG^T)_{ij}}{\partial x_j} - F_i\right)\frac{\partial \rho}{\partial x_i} - J\rho\right]\Phi \\
&= \lim_{k'\to\infty}\iint_{\mathbb{Q}_{R_k}} \frac{1}{2}\sum_{i,j=1}^{n}(GQG^T)_{ij}\frac{\partial \Phi}{\partial x_j}\frac{\partial \rho_{k'}}{\partial x_i} + \left[\sum_{i=1}^{n}\left(\sum_{j=1}^{n}\frac{\partial (GQG^T)_{ij}}{\partial x_j} - F_i\right)\frac{\partial \rho_{k'}}{\partial x_i} - J\rho_{k'}\right]\Phi \\
&= -\lim_{k'\to\infty}\iint_{\mathbb{Q}_{R_k}} \frac{\partial \rho_{k'}}{\partial t}\Phi = \lim_{k'\to\infty}\iint_{\mathbb{Q}_{R_k}} \rho_{k'}\frac{\partial \Phi}{\partial t} = \iint_{\mathbb{Q}_{R_k}} \rho\frac{\partial \Phi}{\partial t}.
\end{aligned}$$

Clearly, $\rho(x,0) = \lim_{k'\to\infty}\rho_{k'}(x,0) = \sigma_0(x)$. □

Theorem 3.3 (Uniqueness)**.** *Assume further that for some $c > 0$,*

$$\sup_{0\le t\le T}\int_{\mathbb{R}^n} e^{cr}\rho^2(x,t)dx < \infty, \tag{3.19}$$

and

$$\int_{\mathbb{Q}_T} |\nabla \rho(x,t)|^2 dxdt < \infty, \tag{3.20}$$

where $r = |x|$. Suppose that there exists a finite number $\alpha > 0$ such that

$$2J(x,t) - \frac{1}{4\lambda_1}[cD_w r - (F(x,t) + \tilde{F}(x,t))]^2 \le \alpha, \tag{3.21}$$

for all $(x,t) \in \mathbb{Q}_T$*, where* λ_1 *is the smallest eigenvalue of the matrix* (GQG^T)*,*

$$\tilde{F}(x,t) = \left[\frac{1}{2} \sum_{j=1}^{n} (GQG^T)_{ij} + \sum_{j=1}^{n} (GQG^T)_{ij} \frac{\partial K}{\partial x_j} - f_i \right]_{i=1}^{n}, \tag{3.22}$$

$F(x,t)$*,* $J(x,t)$ *and* K *are defined as in Equation 2.8, Equation 2.9 and Equation 2.10, respectively. Then the non-negative weak solution* $\rho(x,t)$ *of the "pathwise-robust" DMZ equation on* $\mathbb{Q}_T$ *is unique.*

Proof. To show the uniqueness of the solution, we only need to show that $\rho(x,t) = 0$ on $\mathbb{Q}_T$ if $\rho(x,0) = 0$. Let $\alpha T < 1$. For any test function $\psi(x,t) = e^{cr}\Phi(x,t)$, where $r = |x|$, c is some constant and $\Phi(x,t) \in H_0^{1;1}(\mathbb{Q}_T)$, $\rho(x,t)$ satisfies

$$\begin{aligned}
&\int_{\mathbb{R}^n} \rho(x,T)\Phi(x,T)e^{cr}dx - \int_0^T \int_{\mathbb{R}^n} \rho(x,t)\frac{\partial \Phi}{\partial t}(x,t)e^{cr}dxdt \\
=&\int_{\mathbb{Q}_T} -\frac{1}{2}e^{cr}\nabla\Phi(x,t)\cdot D_w\rho(x,t) - \frac{c}{2}e^{cr}\Phi(x,t)\nabla r \cdot D_w\rho(x,t) + \tilde{F}(x,t)\cdot\nabla\rho(x,t)\Phi(x,t)e^{cr} \\
&\quad + J(x,t)\rho(x,t)\Phi(x,t)e^{cr}dxdt.
\end{aligned} \tag{3.23}$$

where $\tilde{F}$ is defined in Equation 3.22. Approximating $\rho(x,t)$ by $\Phi(x,t)$ in the $H^{1;1}(\mathbb{Q}_T)$-norm, we get

$$
\begin{aligned}
&\int_{\mathbb{R}^n} \rho^2(x,T)e^{cr}dx \\
=&\int_{\mathbb{Q}_T} e^{cr}\Big[-D_w\rho(x,t)\cdot\nabla\rho(x,t) - c\rho(x,t)\nabla r\cdot D_w\rho(x,t) + (\tilde{F}(x,t)+F(x,t))\cdot\nabla\rho(x,t)\rho(x,t) \\
&\qquad +2J(x,t)\rho^2(x,t)\Big]\,dxdt \\
\leq&\int_{\mathbb{Q}_T} e^{cr}[-\lambda_1|\nabla\rho(x,t)|^2 - c\rho(x,t)D_w r\cdot\nabla\rho(x,t) + (F(x,t)+\tilde{F}(x,t))\cdot\nabla\rho(x,t)\rho(x,t) \\
&\qquad + 2J(x,t)\rho^2(x,t)]dxdt. \\
=&-\lambda_1\int_{\mathbb{Q}_T} e^{cr}\left\{\frac{1}{2\lambda_1}[cD_w r-(F(x,t)+\tilde{F}(x,t))]\rho(x,t)+|\nabla\rho(x,t)|\right\}^2 dxdt \\
&+\int_{\mathbb{Q}_T} e^{cr}\left\{2J(x,t)-\frac{1}{4\lambda_1}[cD_w r-(F(x,t)+\tilde{F}(x,t))]^2\right\}\rho^2(x,t)dxdt \\
\leq&\int_{\mathbb{Q}_T} e^{cr}\left\{2J(x,t)-\frac{1}{4\lambda_1}[cD_w r-(F(x,t)+\tilde{F}(x,t))]^2\right\}\rho^2(x,t)dxdt, \qquad (3.24)
\end{aligned}
$$

due to the positive definiteness of (GQG^T). By Equation 3.21, we have

$$\int_{\mathbb{R}^n} e^{cr}\rho^2(x,T)dx \leq \alpha\int_{\mathbb{Q}_T} e^{cr}\rho^2(x,t)dxdt. \qquad (3.25)$$

According to the mean value theorem, there exists $T_1\in(0,T)$ such that

$$\int_{\mathbb{Q}_T} e^{cr}\rho^2(x,t)dxdt = \int_0^T\int_{\mathbb{R}^n} e^{cr}\rho^2(x,t)dxdt = T\int_{\mathbb{R}^n} e^{cr}\rho^2(x,T_1)dx. \qquad (3.26)$$

Apply Equation 3.25 and Equation 3.26 recursively, there exists $T_m \in (0,T)$ such that

$$\int_{\mathbb{R}^n} e^{cr}\rho^2(x,T)dx \leq (\alpha T)^m \int_{\mathbb{R}^n} e^{cr}\rho^2(x,T_m)dx.$$

Since $\alpha T < 1$, we conclude that $\rho(x,t) \equiv 0$ for a.e $(x,t) \in \mathbb{Q}_T$. □

3.3 Properties of the solution

Let us first state a very useful lemma, which will be used repeatedly in this section and Chapter 4.

Lemma 3.1. *Assume that ρ_Ω satisfies the "pathwise-robust" DMZ equation on $\mathbb{Q}_\Omega := \Omega \times [0,T]$, where $\Omega \subset \mathbb{R}^n$ is some bounded domain. Then, for any test function $\psi(x) \in C^\infty(\Omega)$, we have*

$$\begin{aligned}\frac{d}{dt}\int_\Omega \psi\rho_\Omega =& \frac{1}{2}\int_\Omega D_w^2\psi\rho_\Omega + \int_\Omega (f - D_wK)\cdot\nabla\psi\rho_\Omega + \int_\Omega \psi\rho_\Omega N + \frac{1}{2}\int_{\partial\Omega}\psi\left(D_w\rho_\Omega\cdot\nu\right)\\ &- \frac{1}{2}\int_{\partial\Omega}\rho_\Omega\left(D_w\psi\cdot\nu\right) + \frac{1}{2}\int_{\partial\Omega}\psi\rho_\Omega\sum_{i,j=1}^n \frac{\partial}{\partial x_i}\left(GQG^T\right)_{ij}\nu_j + \int_{\partial\Omega}\psi\rho_\Omega\left(D_wK\cdot\nu\right)\\ &- \int_{\partial\Omega}\psi\rho_\Omega(f\cdot\nu), \end{aligned} \tag{3.27}$$

where $\nu = (\nu_1, \nu_2, \cdots, \nu_n)$ is the exterior normal vector of Ω,

$$N(x,t) \equiv -\frac{\partial}{\partial t}\left(h^T S^{-1}\right) y_t - \frac{1}{2}D_w^2K + \frac{1}{2}D_wK\cdot\nabla K - f\cdot\nabla K - \frac{1}{2}\left(h^T S^{-1}h\right), \tag{3.28}$$

and D_w^2, K and D_w are defined in Equation 2.7, Equation 2.10 and Equation 3.6, respectively.

Sketch of the proof. Multiply $\psi(x)$ on both sides of Equation 2.6 and integrate over the domain Ω, which yields

$$\frac{d}{dt}\int_{\Omega}\psi\rho_{\Omega} = \int_{\Omega}\psi\left[\frac{1}{2}D_w^2\rho_{\Omega} + F(x,t)\cdot\nabla\rho_{\Omega} + J(x,t)\rho_{\Omega}\right], \tag{3.29}$$

where $F(x,t)$ and $J(x,t)$ are defined in Equation 2.8 and Equation 2.9, respectively. After applying integration by parts to the first two terms on the right-hand side of Equation 3.29, Equation 3.27 is obtained by written in short notations. □

3.3.1 Density function in a large ball

We show an interesting proposition, which reflects how the density function in the large ball changes with respect to time. It will also be an important ingredient of the error estimate in Theorem 4.1.

Proposition 3.1. *For any $T > 0$, let $\rho_R(x,t)$ be a solution of the "pathwise-robust" DMZ equation on $\mathbb{Q}_R$, i.e. Equation 3.1. Assume that*

$$N(x,t) + \frac{3}{2}n\left|\left|GQG^T\right|\right|_{\infty} + |f - D_wK| \leq C, \tag{3.30}$$

Then

$$\int_{B_R} e^{\sqrt{1+|x|^2}}\rho_R(x,t) \leq e^{Ct}\int_{\mathbb{R}^n} e^{\sqrt{1+|x|^2}}\sigma_0(x). \tag{3.31}$$

Proof. Letting the test function ψ in Lemma 3.1 be $\psi = e^{\phi_1}$, where $\phi_1 \in C^\infty(B_R)$, gives

$$\frac{d}{dt}\int_{B_R} e^{\phi_1}\rho_R = \int_{B_R} e^{\phi_1}\rho_R \left[\frac{1}{2}\left(D_w^2\phi_1 + D_w\phi_1\cdot\nabla\phi_1\right) + (f - D_wK)\cdot\nabla\phi_1 + N\right] + \frac{1}{2}\int_{\partial B_R} e^{\phi_1}(D_w\rho_R\cdot\nu). \tag{3.32}$$

All the boundary integrals in Equation 3.27 vanish, except the first term, since $\rho_R|_{\partial\Omega} = 0$. Moreover, recall that $\rho_R \geq 0$ in B_R and vanishes on ∂B_R which implies that $\frac{\partial\rho_R}{\partial\nu}|_{\partial B_R} \leq 0$. Hence, on ∂B_R,

$$(D_w\rho_R\cdot\nu) = \sum_{i=1}^n \left[\sum_{j=1}^n (GQG^T)_{ij}\frac{\partial\rho_R}{\partial r}\frac{\partial r}{\partial x_j}\right]\nu_i = \frac{\partial\rho_R}{\partial r}\left[\sum_{i,j=1}^n (GQG^T)_{ij}\frac{x_j}{r}\frac{x_i}{r}\right] \leq 0,$$

by the positive definite assumption of (GQG^T). Thus, Equation 3.32 can be reduced further to

$$\frac{d}{dt}\int_{B_R} e^{\phi_1}\rho_R \leq \int_{B_R} e^{\phi_1}\rho_R \left[\frac{1}{2}\left(D_w^2\phi_1 + D_w\phi_1\cdot\nabla\phi_1\right) + (f - D_wK)\cdot\nabla\phi_1 + N\right]. \tag{3.33}$$

Choose $\phi_1(x) = \sqrt{1+|x|^2}$ and estimate the terms containing ϕ_1 on the right-hand side of Equation 3.33 one by one:

$$\begin{aligned} D_w^2\phi_1 &= \sum_{i=1}^n \left(GQG^T\right)_{ii}\frac{1}{\sqrt{1+|x|^2}} - \sum_{i,j=1}^n \left(GQG^T\right)_{ij}\frac{x_ix_j}{(1+|x|^2)^{\frac{3}{2}}} \\ &\leq \left|\left|GQG^T\right|\right|_\infty \left[\frac{n}{\sqrt{1+|x|^2}} + \frac{n|x|^2}{(1+|x|^2)^{\frac{3}{2}}}\right] \leq 2n\left|\left|GQG^T\right|\right|_\infty, \end{aligned} \tag{3.34}$$

$$D_w\phi_1\cdot\nabla\phi_1 = \sum_{i,j=1}^n \left(GQG^T\right)_{ij}\frac{x_ix_j}{1+|x|^2} \leq \left|\left|GQG^T\right|\right|_\infty \frac{\sum_{i,j=1}^n x_ix_j}{1+|x|^2} \leq n\left|\left|GQG^T\right|\right|_\infty, \tag{3.35}$$

and

$$|(f - D_w K) \cdot \nabla \phi_1| \leq |f - D_w K| \cdot \frac{|x|}{\sqrt{1 + |x|^2}} \leq |f - D_w K|, \tag{3.36}$$

where $|\cdot|$ is the Euclidean norm. Substituting the estimates in Equation 3.34 - Equation 3.36 back into Equation 3.33, we get

$$\frac{d}{dt} \int_{B_R} e^{\phi_1} \rho_R \leq \int_{B_R} e^{\phi_1} \rho_R \left[\frac{3}{2} n \left|\left| GQG^T \right|\right|_\infty + |f - D_w K| + N \right] \leq C \int_{B_R} e^{\phi_1} \rho_R,$$

by Equation 3.30. Hence,

$$\int_{B_R} e^{\phi_1} \rho_R(x,t) \leq e^{Ct} \int_{B_R} e^{\phi_1} \rho_R(x,0) \leq e^{Ct} \int_{\mathbb{R}^n} e^{\phi_1} \rho(x,0) = e^{Ct} \int_{\mathbb{R}^n} e^{\phi_1} \sigma_0(x),$$

for $0 \leq t \leq T$. □

3.3.2 Concentration of the density function

The following theorem asserts that ρ, the solution to the "pathwise-robust" DMZ equation in $\mathbb{Q}_T$, captures almost all the density in a large ball. And we give a precise estimate of the density outside the large ball.

Theorem 3.4. *Let $\rho(x,t)$ be a solution of the "pathwise-robust" DMZ equation, i.e. Equation 2.6, on $\mathbb{Q}_T$. Assume that Equation 3.30 and*

$$e^{-\frac{1}{2}\sqrt{1+|x|^2}} \left[16n \left|\left| GQG^T \right|\right|_\infty + 4 \left| f - D_w K \right| \right] \leq C \tag{3.37}$$

are satisfied for all $(x,t) \in \mathbb{Q}_T$. *Then*

$$\int_{|x|\geq R} \rho(x,T) \leq Ce^{-\frac{1}{2}\sqrt{1+R^2}} \int_{\mathbb{R}^n} e^{\sqrt{1+|x|^2}} \sigma_0(x), \tag{3.38}$$

where C *is a generic constant, which depends on* T.

Proof. Let $v = \rho - \rho_R$. By the maximum principle, we have that $v \geq 0$ for all $(x,t) \in \mathbb{Q}_R$. Choose the test function ψ in Lemma 3.1 as

$$\Phi(x) = \gamma(x)\varrho(x),$$

where $\gamma(x) = e^{\frac{1}{2}\phi_1(x)}$, $\phi_1(x) = \sqrt{1+|x|^2}$ is defined in the proof of Proposition 3.1, $\varrho(x) = e^{-\phi_2(x)} - e^{-R}$, ϕ_2 is a radially symmetric function such that $\phi_2(x)|_{\partial B_R} = R$, $\nabla\phi_2(x)|_{\partial B_R} = 0$ and $\phi_2(x)$ is increasing in $|x|$. It follows directly that $\Phi|_{\partial B_R} = \nabla_x\Phi|_{\partial B_R} = 0$, by the fact that $\varrho|_{\partial B_R} = \nabla\varrho|_{\partial B_R} = 0$. Applying Lemma 3.1, with v taking the place of ρ_Ω and with the test function Φ, we have

$$\begin{aligned}
\frac{d}{dt}\int_{B_R} \Phi v =& \frac{1}{2}\int_{B_R} D_w^2\Phi v + \int_{B_R} (f - D_wK)\cdot\Phi v + \int_{B_R} \Phi N v \\
=& \frac{1}{2}\int_{B_R} (D_w^2\gamma\varrho + 2D_w\gamma\cdot\nabla\varrho + \gamma D_w^2\varrho)v + \int_{B_R} (f - D_wK)\cdot(\nabla\gamma\varrho + \gamma\nabla\varrho)v \\
&+ \int_{B_R} \gamma\varrho N v.
\end{aligned} \tag{3.39}$$

Substituting $\gamma(x) = e^{\frac{1}{2}\phi_1(x)}$ and $\varrho(x) = e^{-\phi_2(x)} - e^{-R}$ into Equation 3.39 yields

$$\begin{aligned}
\frac{d}{dt}\int_{B_R}\Phi v =& \frac{1}{2}\int_{B_R}\left[\frac{1}{2}e^{\frac{1}{2}\phi_1}\left(D_w^2\phi_1 + \frac{1}{2}D_w\phi_1\cdot\nabla\phi_1\right)\varrho - e^{\frac{1}{2}\phi_1}D_w\phi_1\cdot e^{-\phi_2}\nabla\phi_2\right.\\
&\left.+\gamma e^{-\phi_2}\left(D_w\phi_2\cdot\nabla\phi_2 - D_w^2\phi_2\right)\right]v\\
&+\int_{B_R}(f - D_wK)\cdot\left(\frac{1}{2}e^{\frac{1}{2}\phi_1}\nabla\phi_1\varrho - \gamma e^{-\phi_2}\nabla\phi_2\right)v + \int_{B_R}\gamma\varrho Nv\\
=&\int_{B_R}\Phi v\left[\frac{1}{4}\left(D_w^2\phi_1 + \frac{1}{2}D_w\phi_1\cdot\nabla\phi_1\right) - \frac{1}{2}D_w\phi_1\cdot\nabla\phi_2\right.\\
&\left.+\frac{1}{2}\left(D_w\phi_2\cdot\nabla\phi_2 - D_w^2\phi_2\right) + (f - D_wK)\cdot(\frac{1}{2}\nabla\phi_1 - \nabla\phi_2) + N\right]\\
&+e^{-R}\int_{B_R}\gamma v\left[-\frac{1}{2}D_w\phi_1\cdot\nabla\phi_2 + \frac{1}{2}\left(D_w\phi_2\cdot\nabla\phi_2 - D_w^2\phi_2\right) - (f - D_wK)\cdot\nabla\phi_2\right]\\
\triangleq&\int_{B_R}\Phi v[\mathrm{VI}] + e^{-R}\int_{B_R}\gamma v[\mathrm{VII}],
\end{aligned}$$

Let us choose $\phi_2(x)$ in $\varrho(x)$ to be $\phi_2(x) = R\vartheta(\frac{|x|^2}{R^2})$, where $\vartheta(x) = 1 - (1-x)^2$. It is easy to check that $\phi_2(x)$ satisfies all the conditions we mentioned before. Direct computations yield, for any $x \in B_R$, $R >> 1$,

$$\begin{aligned}
|D_w^2\phi_2| &= \left|\sum_{i,j=1}^{n}\left(GQG^T\right)_{ij}\left(-\frac{8x_ix_j}{R^3}\right) + \sum_{i=1}^{n}\left(GQG^T\right)_{ii}\frac{4}{R}\left(1 - \frac{|x|^2}{R^2}\right)\right|\\
&\le \left|\left|GQG^T\right|\right|_\infty\left(\frac{8n|x|^2}{R^3} + \frac{4n}{R}\right) \le 12n\left|\left|GQG^T\right|\right|_\infty, \qquad (3.40)
\end{aligned}$$

$$|D_w\phi_2\cdot\nabla\phi_2| = \left|\left(1 - \frac{|x|^2}{R^2}\right)^2\sum_{i,j=1}^{n}\left(GQG^T\right)_{ij}\frac{4x_i}{R}\frac{4x_j}{R}\right| \le 16n\left|\left|GQG^T\right|\right|_\infty, \qquad (3.41)$$

and

$$|(f - D_w K) \cdot \nabla \phi_2| = \left| (f - D_w K) \frac{4x}{R} \left(1 - \frac{|x|^2}{R^2} \right) \right| \leq 4 \left| f - D_w K \right|. \tag{3.42}$$

It follows that

$$\sup_{B_R} |\mathrm{VI}| \leq 17n \left|\left| GQG^T \right|\right|_\infty + 5|f - D_w K| + N,$$

$$\sup_{B_R} |\mathrm{VII}| \leq 16n \left|\left| GQG^T \right|\right|_\infty + 4|f - D_w K|.$$

by Equation 3.34 - Equation 3.36 and Equation 3.40 - Equation 3.42. Hence,

$$\begin{aligned} \frac{d}{dt} \int_{B_R} \Phi v \leq & C \int_{B_R} \Phi v + e^{-R} \tilde{C} \int_{B_R} e^{\phi_1} v \leq C \int_{B_R} \Phi v + e^{-R} \tilde{C} \int_{B_R} e^{\phi_1} \rho \\ \leq & C \int_{B_R} \Phi v + \tilde{C} e^{-R+Ct} \int_{B_R} e^{\phi_1} \sigma_0(x) \leq C \int_{B_R} \Phi v + \tilde{C} e^{-R+Ct} \int_{\mathbb{R}^n} e^{\phi_1} \sigma_0(x), \end{aligned}$$

by Equation 3.30, Equation 3.37 and Equation 3.31. By Gronwall's inequality, we have

$$\int_{B_R} \Phi v(x, T) \leq C e^{-R} \int_{\mathbb{R}^n} e^{\sqrt{1+|x|^2}} \sigma_0(x), \tag{3.43}$$

where C is a generic constant, which depends on T. Recall that $\Phi(x) = \gamma(x)\varrho(x)$ and $\varrho(x) = e^{-R[-(|x|^2/R^2-1)^2+1]} - e^{-R}$, which implies that

$$\int_{B_R} \Phi v(x, T) \geq \frac{1}{2} e^{-\frac{7}{16}R} \int_{B_{\frac{R}{2}}} \gamma v(x, T). \tag{3.44}$$

Combining Equation 3.43 and Equation 3.44, we obtain

$$\int_{B_{\frac{R}{2}}} \gamma v(x,T) \leq Ce^{-\frac{9}{16}R} \int_{\mathbb{R}^n} e^{\sqrt{1+|x|^2}} \sigma_0(x).$$

This implies that

$$\int_{B_{\frac{R}{2}}} \gamma\rho(x,T) \leq \int_{B_{\frac{R}{2}}} \gamma\rho_R(x,T) + Ce^{-\frac{9}{16}R} \int_{\mathbb{R}^n} e^{\sqrt{1+|x|^2}} \sigma_0(x) \leq C(1+e^{-\frac{9}{16}R}) \int_{\mathbb{R}^n} e^{\sqrt{1+|x|^2}} \sigma_0(x),$$

by Equation 3.31. Letting $R \to \infty$, yields

$$\int_{\mathbb{R}^n} \gamma\rho(x,T) \leq C \int_{\mathbb{R}^n} e^{\sqrt{1+|x|^2}} \sigma_0(x).$$

Consider the integration outside the large ball B_R,

$$e^{\frac{1}{2}\sqrt{1+R^2}} \int_{|x|\geq R} \rho(x,T) \leq \int_{|x|\geq R} \gamma\rho(x,T) \leq C \int_{\mathbb{R}^n} e^{\sqrt{1+|x|^2}} \sigma_0(x).$$

Therefore, we reach the conclusion that

$$\int_{|x|\geq R} \rho(x,T) \leq Ce^{-\frac{1}{2}\sqrt{1+|R|^2}} \int_{\mathbb{R}^n} e^{\sqrt{1+|x|^2}} \sigma_0(x).$$

□

3.3.3 Lower bound estimate of density function

It is well-known that solving the "pathwise-robust" DMZ equation numerically is not easy because it is easily vanishing. We are also interested in whether a lower bound for the density function could be derived in the case where the drift term f and the observation term h are of at most polynomial growth. The theorem below gives this lower bound:

Theorem 3.5. *Let ρ_R be the solution of the "pathwise-robust" DMZ equation on $\mathbb{Q}_R$, i.e. Equation 3.1. Assume that*

$$N(x,t) \leq C, \tag{3.45}$$

and

1. *$f(x,t)$ and $h(x,t)$ have at most polynomial growth in $|x|$, for all $t \in [0,T]$;*
2. *For any $0 \leq t \leq T$, there exists positive integer m and positive constants C' and C'' independent of R such that the following two conditions hold:*

$$\begin{aligned}
\text{(a)}\quad & \frac{|x|^{m-2}}{2}\left[nm(m-2)\left|\left|GQG^T\right|\right|_\infty + m\,\mathrm{Tr}\left(GQG^T\right)\right] - m|x|^{m-2}(f - D_w K)\cdot x \\
& + N(x,t) \geq -C'; \qquad (3.46)\\
\text{(b)}\quad & \left| n\left|\left|GQG^T\right|\right|_\infty \left(\frac{1}{2}m^2|x|^{2m-2} - m\left(\frac{1}{2}m - 1\right)|x|^{m-2}\right) - \frac{1}{2}m\,\mathrm{Tr}\left(GQG^T\right)|x|^{m-2}\right. \\
& \left. -m(f - D_w K)\cdot x|x|^{m-2}\right| \\
\leq & \frac{1}{2}nm(m+1)\left|\left|GQG^T\right|\right|_\infty |x|^{2m-2} + C'', \qquad (3.47)
\end{aligned}$$

where $\mathrm{Tr}(*)$ *is the trace of* $*$.

Then for any $R_0 < R$,

$$\int_{B_{R_0}} \zeta\rho_R(x,T) \geq \frac{e^{(C-C')T-R_0^m}}{C'}\left(\frac{1}{2}nm(m+1)\left|\left|GQG^T\right|\right|_\infty R_0^{2m-2} + C''\right) \cdot \left(1-e^{C'T}\right)\int_{B_R}\sigma_{0,R}(x) + e^{-C'T}\int_{B_{R_0}}\zeta\sigma_{0,R}(x),$$

where $\zeta(x) = e^{-\xi(x)} - e^{-\xi(R_0)}$, $\xi(x) = |x|^m$.

In particular, the solution ρ *of the "pathwise-robust" DMZ equation, i.e. Equation 2.5, on* $\mathbb{Q}_T$ *has the estimate*

$$\int_{\mathbb{R}^n} e^{-|x|^m}\rho(x,T) \geq e^{-C'T}\int_{\mathbb{R}^n} e^{-|x|^m}\sigma_0(x).$$

Proof. Apply Lemma 3.1 to ρ_R with the test function ψ to be $\zeta = e^{-\xi(x)} - e^{-\xi(R_0)}$, where $\xi(x)$ is an increasing function in $|x|$. Then we have

$$\frac{d}{dt}\int_{B_{R_0}}\zeta\rho_R = \int_{B_{R_0}}\rho_R\left[\frac{1}{2}D_w^2\zeta + (f - D_wK)\cdot\nabla\zeta + \zeta N\right].$$

All the boundary integrals vanish, since $\zeta|_{\partial B_R} = \rho_R|_{\partial B_R} = 0$. Direct computations yield

$$\begin{aligned}
&\frac{d}{dt}\int_{B_{R_0}} \zeta\rho_R \\
=&\int_{B_{R_0}} \rho_R e^{-\xi(R_0)} \left\{ \frac{1}{2}\frac{\xi'^2(r)}{r^2}\sum_{i,j=1}^{n}\left(GQG^T\right)_{ij} x_i x_j - \frac{\xi'(r)}{r}(f - D_w K)\cdot x \right. \\
&\qquad \left. -\frac{1}{2}\sum_{i,j=1}^{n}\left(GQG^T\right)_{ij}\left[\left(\xi''(r) - \frac{\xi'(r)}{r}\right)\frac{x_i x_j}{r^2}\right] - \frac{1}{2}\mathrm{Tr}\left(GQG^T\right)\frac{\xi'(r)}{r} \right\} \\
&+\int_{B_{R_0}} \zeta\rho_R\left[\frac{1}{2}D_w\xi\cdot\nabla\xi - \frac{1}{2}D_w^2\xi - (f - D_w K)\cdot\nabla\xi + N\right] \\
\triangleq&\mathrm{VIII} + \int_{B_{R_0}} \zeta\rho_R[\mathrm{IX}].
\end{aligned} \tag{3.48}$$

Let $\xi(r) = r^m$, where $r = |x|$, m is some positive integer sufficiently large. Through elementary computations, we get

$$\begin{aligned}
\mathrm{IX} =&\frac{1}{2}\frac{\xi'^2(r)}{r^2}\sum_{i,j=1}^{n}\left(GQG^T\right)_{ij} x_i x_j \\
&-\frac{1}{2}\left[m(m-2)r^{m-4}\sum_{i,j=1}^{n}\left(GQG^T\right)_{ij} x_i x_j + mr^{m-2}\mathrm{Tr}\left(GQG^T\right)\right] \\
&-mr^{m-2}(f - D_w K)\cdot x + N \\
\geq&-\frac{1}{2}\left[nm(m-2)\left|\left|GQG^T\right|\right|_\infty + m\,\mathrm{Tr}\left(GQG^T\right)\right] r^{m-2} - mr^{m-2}(f - D_w K)\cdot x + N \geq C',
\end{aligned} \tag{3.49}$$

where C' is a positive constant independent of R_0, by Equation 3.46. For large enough m, we have

$$\begin{aligned}
|\text{VIII}| &\le e^{-R_0^m} \int_{B_R} \Big| n \left|\left|GQG^T\right|\right|_\infty \left[\frac{1}{2}m^2 r^{2m-2} - m\left(\frac{1}{2}m - 1\right) r^{m-2}\right] - \frac{1}{2}m \operatorname{Tr}\left(GQG^T\right) r^{m-2} \\
&\qquad -m(f - D_w K)\cdot x r^{m-2}\Big| \rho_R \\
&\le e^{-R_0^m} \left(\frac{1}{2}nm(m+1)\left|\left|GQG^T\right|\right|_\infty R_0^{2m-2} + C''\right) \int_{B_R} \rho_R \\
&\le \left(\frac{1}{2}nm(m+1)\left|\left|GQG^T\right|\right|_\infty R_0^{2m-2} + C''\right) e^{CT - R_0^m} \int_{B_R} \sigma_{0,R} \triangleq \eta(R_0). \qquad (3.50)
\end{aligned}$$

The last inequality follows from the fact that $\frac{d}{dt}\int_{B_R}\rho \le \int_{B_R}\rho N \le C\int_{B_R}\rho$, where Lemma 3.1 with $\psi = 1$ is applied to Equation 2.6. Hence, combining Equation 3.48 - Equation 3.50, we get

$$\frac{d}{dt}\int_{B_{R_0}} \zeta\rho_R \ge -\eta(R_0) - C'\int_{B_{R_0}} \zeta\rho_R.$$

This implies that

$$\begin{aligned}
&\int_{B_{R_0}} \zeta\rho_R(x,T) \ge e^{-C'T}\int_{B_{R_0}} \zeta\sigma_{0,R}(x) + \frac{\gamma(R_0)}{C'}\left(e^{-C'T} - 1\right) \\
\ge& e^{-C'T}\int_{B_{R_0}} \zeta\sigma_{0,R}(x) \\
&+ \left(\frac{1}{2}nm(m+1)\left|\left|GQG^T\right|\right|_\infty R_0^{2m-2} + C''\right)\cdot \frac{e^{(C-C')T - R_0^m}}{C'}\left(1 - e^{C'T}\right)\int_{B_R}\sigma_{0,R}(x). \qquad (3.51)
\end{aligned}$$

In particular, letting $R_0 \to \infty$, we have

$$\int_{\mathbb{R}^n} e^{-|x|^m} \rho(x,T) \geq e^{-C'T} \int_{\mathbb{R}^n} e^{-|x|^m} \sigma_0(x).$$

□

CHAPTER 4

CONVERGENCE ANALYSIS OF OUR ALGORITHM

In Chapter 2, we described our algorithm in detail, where y_t in Equation 2.11 is approximated by $y_{\tau_{i-1}}$ on $[\tau_{i-1}, \tau_i)$. This is the only approximation in our algorithm. In this chapter, we shall show the convergence of our algorithm rigorously.

We first show that the solution ρ to the "pathwise-robust" DMZ equation, i.e. Equation 2.5, is well approximated by ρ_R as $R \to \infty$, for any $t \in [0, T]$, where ρ_R is the solution to the "pathwise-robust" DMZ equation on $\mathbb{Q}_R$, i.e. Equation 3.1. Next, we shall show that $\rho_{i,R} \to \rho_R$ in some sense, as $|\mathcal{P}_k| \to 0$, where $\rho_{i,R}$ is the solution to Equation 2.11 on $\mathbb{Q}_R$, which we rewrite below

$$
\begin{cases}
\dfrac{\partial \rho_{i,R}}{\partial t}(x,t) + \dfrac{\partial}{\partial t}\left(h^T S^{-1}\right)^T y_{\tau_{i-1}} \rho_{i,R}(x,t) \\
\qquad = \exp\left(-h^T S^{-1} y_{\tau_{i-1}}\right)\left[L - \dfrac{1}{2} h^T S^{-1} h\right] \cdot \left[\exp\left(h^T S^{-1} y_{\tau_{i-1}}\right) \rho_{i,R}(x,t)\right], \quad x \in B_R \\
\rho_{i,R}(x,t) = 0, \quad (x,t) \in \partial B_R \times [0,T], \\
\rho_{1,R}(x,0) = \sigma_{0,R}(x), \\
\text{and} \\
\rho_{i,R}(x,\tau_{i-1}) = \rho_{i-1,R}(x,\tau_{i-1}), \quad \text{for } i = 2, 3, \cdots, k.
\end{cases}
\tag{4.1}
$$

4.1 Reduction to the bounded domain case

Theorem 4.1. *For any $T > 0$, let $\rho(x,t)$ be a solution of the "pathwise-robust" DMZ equation Equation 2.6 in $\mathbb{R}^n \times [0,T]$. Let $R \gg 1$ and ρ_R be the solution to Equation 3.1. Assume that Equation 3.30 and the bound*

$$e^{-\sqrt{1+|x|^2}} \left[14n \left|\left|GQG^T\right|\right|_\infty + 4\left|f - D_w K\right|\right] \leq \tilde{C}, \tag{4.2}$$

are satisfied for all $(x,t) \in \mathbb{Q}_R$, where N, D_w and K are defined in Equation 3.28, Equation 3.6 and Equation 2.10, respectively, and C, $\tilde{C}$ are constants possibly depending on T. Let $v = \rho - \rho_R$, then $v \geq 0$ for all $(x,t) \in \mathbb{Q}_R$ and

$$\int_{B_{\frac{R}{2}}} v(x,T) \leq \bar{C} e^{-\frac{9}{16}R} \int_{\mathbb{R}^n} e^{\sqrt{1+|x|^2}} \sigma_0(x), \tag{4.3}$$

where $\bar{C}$ is some constant, which may depend on T.

Proof of Theorem 4.1. By the maximum principle (cf. Theorem 1, (18)), we have $v = \rho - \rho_R \geq 0$ for $(x,t) \in \mathbb{Q}_R$, since $v|_{\partial B_R} \geq 0$ for $0 \leq t \leq T$. Let us choose ψ in Lemma 3.1 to be $\varrho(x)$

$$\varrho(x) = e^{-\phi_2(x)} - e^{-R},$$

where ϕ_2 is a radial symmetric function such that $\phi_2(x)|_{\partial B_R} = R$, $\nabla\phi_2|_{\partial B_R} = 0$ and ϕ_2 is increasing in $|x|$. Hence, $\varrho|_{\partial B_R} = 0$ and $\nabla\varrho|_{\partial B_R} = 0$. Apply Lemma 3.1 to v, not ρ_Ω in Equation 3.27, with the test function $\psi = \varrho$, we have

$$\begin{aligned}
\frac{d}{dt}\int_{B_R}\varrho v &= \int_{B_R} v\left[\frac{1}{2}D_w^2\varrho + (f - D_wK)\cdot\nabla\varrho + \varrho N\right] \\
&= \int_{B_R} v\left\{\frac{1}{2}e^{-\phi_2}\left(D_w\phi_2\cdot\nabla\phi_2 - D_w^2\phi_2\right) - e^{-\phi_2}\left(f - D_wK\right)\cdot\nabla\phi_2 + \varrho N\right\} \\
&= \int_{B_R} v\varrho\left[-\frac{1}{2}D_w^2\phi_2 + \frac{1}{2}D_w\phi_2\cdot\nabla\phi_2 - (f - D_wK)\cdot\nabla\phi_2 + N\right] \\
&\quad + e^{-R}\int_{B_R} e^{\sqrt{1+|x|^2}}v\left[e^{-\sqrt{1+|x|^2}}\left(-\frac{1}{2}D_w^2\phi_2 + \frac{1}{2}D_w\phi_2\cdot\nabla\phi_2 - (f - D_wK)\cdot\nabla\phi_2\right)\right] \\
&\triangleq \int_{B_R} v\varrho\mathrm{X} + e^{-R}\int_{B_R} e^{\sqrt{1+|x|^2}}v\mathrm{XI}.
\end{aligned}$$

Estimating X and XI as in the proof of Theorem 6.3, we have

$$\sup_{B_R}|\mathrm{X}| \le 14n\left|\left|GQG^T\right|\right|_\infty + 4\left|f - D_wK\right| + N \le C,$$

by Equation 3.30. Similarly,

$$\sup_{B_R}|\mathrm{XI}| \le \sup_{B_R}\left[e^{-\sqrt{1+|x|^2}}\left(14n\left|\left|GQG^T\right|\right|_\infty + 4\left|f - D_wK\right|\right)\right] \le \tilde{C},$$

by Equation 4.2. In the view of Proposition 3.1, one gets

$$\frac{d}{dt}\int_{B_R}\varrho v \le C\int_{B_R}\varrho v + e^{-R}\tilde{C}\int_{B_R} e^{\sqrt{1+|x|^2}}\rho \le C\int_{B_R}\varrho v + e^{-R+\hat{C}T}\tilde{C}\int_{\mathbb{R}^n} e^{\sqrt{1+|x|^2}}\sigma_0(x). \quad (4.4)$$

Multiplying both sides of Equation 4.4 by e^{-Ct} yields

$$\frac{d}{dt}\left[e^{-Ct}\int_{B_R}\varrho v\right] \leq e^{-R+\hat{C}T-Ct}\tilde{C}\int_{\mathbb{R}^n} e^{\sqrt{1+|x|^2}}\sigma_0(x).$$

Integrate from 0 to T and multiply e^{CT} on both sides gives us

$$\int_{B_R}\varrho v(x,T) \leq ||v(x,0)||_\infty e^{CT}\int_{B_R}\varrho dx + \frac{e^{CT}-1}{C}e^{-R+\hat{C}T}\tilde{C}\int_{\mathbb{R}^n} e^{\sqrt{1+|x|^2}}\sigma_0(x),$$

where $v(x,0) = \sigma_0 - \sigma_{0,R}$. Recalling that $\varrho(x) = e^{-R\left[-\left(|x|^2/R^2-1\right)^2+1\right]} - e^{-R}$, $|x| \leq R$, we arrive the following estimates:

$$\int_{B_R}\varrho \leq \int_{B_R}\left(1-e^{-R}\right) \leq CR^n$$

and

$$\int_{B_R}\varrho v(x,T) \geq \int_{B_{\frac{R}{2}}}\left(e^{-R\left[-\left(|x|^2/R^2-1\right)^2+1\right]} - e^{-R}\right)v(x,T) \geq \frac{1}{2}e^{-\frac{7}{16}R}\int_{B_{\frac{R}{2}}} v(x,T).$$

It is easy to see that $||v(x,0)||_\infty \int_{B_R}\varrho \leq C(n)\epsilon R^n$ is arbitrarily small, since ϵ is independent of R. It follows that

$$\int_{B_{\frac{R}{2}}} v(x,T) \leq Ce^{-\frac{9}{16}R}\int_{\mathbb{R}^n} e^{\sqrt{1+|x|^2}}\sigma_0(x), \tag{1.28}$$

where C is a generic constant, depending on T. □

4.2 L^1 convergence of $\rho_{i,R}$

For any $0 < \tau \leq T$, let us denote the partition $\mathcal{P}_k^\tau = \{0 = \tau_0 < \tau_1 < \cdots < \tau_k = \tau\}$. We shall show $\rho_{k,R}(x,\tau) \to \rho_R(x,t)$ in L^1 sense, as $|\mathcal{P}_k^\tau| \to 0$, where $\rho_{k,R}$ is the solution of Equation 4.1 (or equivalently, Equation 4.7 with $\Omega = B_R$) and ρ_R is the solution to Equation 3.1.

Theorem 4.2. *Let Ω be a bounded domain in $\mathbb{R}^n$. Assume that Equation 3.45 is satisfied and there exists some $\alpha \in (0,1)$, such that*

$$|N(x,t) - N(x,t;\bar{t})| \leq \tilde{C}|t-\bar{t}|^\alpha, \tag{4.5}$$

for all $(x,t) \in \Omega \times [0,T]$, $\bar{t} \in [0,T]$, where $N(x,t)$ is in Equation 3.28, and $N(x,t;\bar{t})$ denotes $N(x,t)$ with the observation $y_t = y_{\bar{t}}$. Let $\rho_\Omega(x,t)$ be the solution of Equation 2.6 on $\Omega \times [0,T]$ with $0-$Dirichlet boundary condition:

$$\begin{cases} \dfrac{\partial \rho_\Omega}{\partial t}(x,t) = \dfrac{1}{2} D_w^2 \rho_\Omega(x,t) + F(x,t) \cdot \nabla \rho_\Omega(x,t) + J(x,t)\rho_\Omega(x,t) \\ \rho_\Omega(x,0) = \sigma_{0,\Omega}(x) \\ \rho_\Omega(x,t)|_{\partial\Omega} = 0, \end{cases} \tag{4.6}$$

where D_w^2, $F(x,t)$ and $J(x,t)$ are defined in Equation 2.7, Equation 2.8 and Equation 2.9, and $\sigma_{0,\Omega}$ is defined in Equation 3.2. For any $0 \leq \tau \leq T$, let $\mathcal{P}_k^\tau = \{0 = \tau_0 < \tau_1 < \tau_2 < \cdots <$

$\tau_k = \tau\}$ be a partition of $[0, \tau]$, where $\tau_i = \frac{i\tau}{k}$. Let $\rho_{i,\Omega}(x,t)$ be the solution to Equation 4.1 on $\Omega \times [\tau_{i-1}, \tau_i]$. Equivalently, $\rho_{i,\Omega}$ is the solution on $\Omega \times [\tau_{i-1}, \tau_i]$ of the equation

$$\begin{cases} \dfrac{\partial \rho_{i,\Omega}}{\partial t}(x,t) = \dfrac{1}{2} D_w^2 \rho_{i,\Omega}(x,t) + F(x,t;\tau_{i-1}) \cdot \nabla \rho_{i,\Omega}(x,t) + J(x,t;\tau_{i-1}) \rho_{i,\Omega}(x,t) \\ \rho_{i,\Omega}(x,\tau_{i-1}) = \rho_{i-1,\Omega}(x,\tau_{i-1}) \\ \rho_{i,\Omega}(x,t)|_{\partial\Omega} = 0, \end{cases} \tag{4.7}$$

for $i = 1, 2, \cdots, k$, with the convention that $\rho_{1,\Omega}(x,0) = \sigma_{0,\Omega}(x)$. Here, $F(x,t;\tau_{i-1})$ and $J(x,t;\tau_{i-1})$ denote $F(x,t)$ and $J(x,t)$ with the observation $y_t = y_{\tau_{i-1}}$, respectively. Then

$$\rho_\Omega(x,\tau) = \lim_{k\to\infty} \rho_{k,\Omega}(x,\tau),$$

in the L^1 sense in space and the following estimate holds:

$$\int_\Omega |\rho_\Omega - \rho_{k,\Omega}|(x,\tau) \le \frac{\bar{C}}{k^\alpha}, \tag{4.8}$$

where $\bar{C}$ is a generic constant, depending on T and $\int_\Omega \sigma_{0,\Omega}$. The right-hand side of Equation 4.8 tends to zero as $k \to \infty$.

For clarity, we state the technique will be used in the proof of Theorem 4.2 as a lemma below.

Lemma 4.1. (Lemma 4.1, (59)) *Let Ω be a bounded domain in $\mathbb{R}^n$ and let $v : \overline{\Omega} \times [0,T] \to \mathbb{R}$ be a C^1 function. Assume that $v(x,t) = 0$ for $(x,t) \in \partial\Omega \times [0,T]$. Let $\Omega_t^+ = \{x \in \Omega : v(x,t) \geq 0\}$. Then*

$$\frac{d}{dt}\int_{\Omega_t^+} v(x,t) = \int_{\Omega_t^+} \frac{\partial v}{\partial t}(x,t),$$

for almost all $t \in [0,T]$.

Proof of Theorem 4.2. For convenience, we omit the subscript Ω in ρ_Ω and $\rho_{i,\Omega}$ in this proof. Let $\Omega_t^+ = \{x \in \Omega : \rho(x,t) - \rho_i(x,t) \geq 0\}$. Applying Lemma 3.1, with $(\rho - \rho_i)$ taking place of ρ_Ω in Equation 3.27, and with the test function $\psi \equiv 1$, we have

$$\frac{d}{dt}\int_{\Omega_t^+} (\rho - \rho_i) \leq \int_{\Omega_t^+} (\rho - \rho_i)N(\cdot,t) + \int_{\Omega_t^+} \rho_i[N(\cdot,t) - N(\cdot,t;\tau_{i-1})], \tag{4.9}$$

by Lemma 4.1. All the boundary integrals vanish, except $\int_{\partial\Omega_t^+} D_w(\rho-\rho_i)\cdot\nu$, since $(\rho-\rho_i)|_{\partial\Omega_t^+} = 0$. Moreover, $\int_{\partial\Omega_t^+} D_w(\rho - \rho_i)\cdot\nu \leq 0$, due to the similar argument for $\int_{\partial B_R} D_w\rho\cdot\nu \leq 0$ in Proposition 3.1. Combining Equation 3.45 and Equation 4.5, Equation 4.9 can be estimated as

$$\frac{d}{dt}\int_{\Omega_t^+} (\rho - \rho_i) \leq C\int_{\Omega_t^+} (\rho - \rho_i) + \tilde{C}(t - \tau_{i-1})^\alpha \int_\Omega \rho. \tag{4.10}$$

To estimate $\int_\Omega \rho$, we apply Lemma 3.1 to ρ, with the test function $\psi \equiv 1$, to get

$$\frac{d}{dt}\int_\Omega \rho \leq \int_\Omega \rho N \leq C\int_\Omega \rho,$$

which implies that

$$\int_\Omega \rho \le C \int_\Omega \sigma_{0,\Omega}, \tag{4.11}$$

where C is a generic constant, depending on T, for all $0 \le t \le T$. Thus,

$$\frac{d}{dt}\int_{\Omega_t^+} (\rho - \rho_i) \le C \int_{\Omega_t^+} (\rho - \rho_i) + \tilde{C}(t - \tau_{i-1})^\alpha \int_\Omega \sigma_{0,\Omega}.$$

Multiplying $e^{-\tilde{C}(t-\tau_{i-1})}$ on both sides and integrating from τ_{i-1} to t, we get

$$\int_{\Omega_t^+} (\rho - \rho_i)(x,t) \le e^{\tilde{C}(t-\tau_{i-1})} \int_{\Omega^+_{\tau_{i-1}}} (\rho - \rho_i)(x,\tau_{i-1}) + C\frac{(t-\tau_{i-1})^{1+\alpha}}{1+\alpha} e^{\tilde{C}(t-\tau_{i-1})},$$

where C is a constant, which depends on T and $\int_\Omega \sigma_{0,\Omega}$. Similarly, we also find for $\Omega_t^- = \{x \in \Omega :\ \rho(x,t) - \rho_i(x,t) < 0\}$, that

$$\int_{\Omega_t^-} (\rho_i - \rho)(x,t) \le e^{\tilde{C}(t-\tau_{i-1})} \int_{\Omega^-_{\tau_{i-1}}} (\rho_i - \rho)(x,\tau_{i-1}) + C\frac{(t-\tau_{i-1})^{1+\alpha}}{1+\alpha} e^{\tilde{C}(t-\tau_{i-1})}.$$

Consequently, we have

$$\begin{aligned}\int_\Omega |\rho - \rho_i|(x,t) \le & e^{\tilde{C}(t-\tau_{i-1})} \left[\int_\Omega |\rho - \rho_i|(x,\tau_{i-1}) + C\frac{(t-\tau_{i-1})^{1+\alpha}}{1+\alpha}\right] \\ \le & e^{\tilde{C}(t-\tau_{i-1})} \left[\int_\Omega |\rho - \rho_{i-1}|(x,\tau_{i-1}) + C\frac{(t-\tau_{i-1})^{1+\alpha}}{1+\alpha}\right],\end{aligned} \tag{4.12}$$

since $\rho_i(x,\tau_{i-1}) = \rho_{i-1}(x,\tau_{i-1})$, for $i = 1,2,\cdots,k$. Applying Equation 4.12 recursively, we obtain

$$\begin{aligned}
&\int_\Omega |\rho-\rho_k|(x,\tau_k) \le e^{\tilde{C}(\tau_k-\tau_{k-1})}\left[\int_\Omega |\rho-\rho_{k-1}|(x,\tau_{k-1}) + C\frac{(\tau_k-\tau_{k-1})^{1+\alpha}}{1+\alpha}\right]\\
\le& e^{\tilde{C}T}\int_\Omega |\rho-\rho_0|(x,0) + \frac{C}{1+\alpha}\Big[(\tau_k-\tau_{k-1})^{1+\alpha}e^{\tilde{C}(\tau_k-\tau_{k-1})} + (\tau_{k-1}-\tau_{k-2})^{1+\alpha}e^{\tilde{C}(\tau_k-\tau_{k-2})}\\
&\qquad\qquad + \cdots + (\tau_1-\tau_0)^{1+\alpha}e^{\tilde{C}(\tau_k-\tau_0)}\Big]\\
=&\frac{C}{1+\alpha}\frac{T^{1+\alpha}}{k^{1+\alpha}}\left(e^{\tilde{C}\frac{T}{k}} + e^{\tilde{C}\frac{2T}{k}} + \cdots + e^{\tilde{C}\frac{kT}{k}}\right) \le \frac{C}{k^\alpha},
\end{aligned}$$

where C is a constant, which depends on α, T and $\int_\Omega \sigma_{0,\Omega}$. It is then clear that $\int_\Omega |\rho-\rho_k| \to 0$, as $k \to \infty$. □

CHAPTER 5

IMPLEMENTATION OF OUR ALGORITHM WITH 1-D STATE

In this chapter, we shall discuss the difficulties in the implementation of our algorithm. As discussed in Chapter 2, when we pre-compute the FKE Equation 2.13, we shall choose the orthogonal basis function $\{\phi_n\}_{n=0}^{\infty}$ to be the generalized Hermite functions $H_n^{\alpha,\beta}(x)$.

5.1 Generalized Hermite functions and orthogonal projection

Let $L^2(\mathbb{R})$ be the Lebesgue space, equipped with the norm $||\cdot|| = (\int_{\mathbb{R}} |\cdot|^2 dx)^{\frac{1}{2}}$ and the scalar product $\langle\cdot,\cdot\rangle$. Let $\mathrm{H}_n(x)$ be the physical Hermite polynomials given by $\mathrm{H}_n(x) = (-1)^n e^{x^2}\partial_x^n e^{-x^2}$, $n \in \mathbb{Z}$ and $n \geq 0$. The three-term recurrence

$$\mathrm{H}_0(x) \equiv 1, \quad \mathrm{H}_1(x) = 2x \quad \text{and} \quad \mathrm{H}_{n+1}(x) = 2x\mathrm{H}_n(x) - 2n\mathrm{H}_{n-1}(x) \tag{5.1}$$

will be used in our implementations. One of the well-known and useful facts of Hermite polynomials is that they are mutually orthogonal with respect to the weight $w(x) = e^{-x^2}$. We define our generalized Hermite functions as

$$H_n^{\alpha,\beta}(x) = \frac{1}{\sqrt{2^n n!}}\mathrm{H}_n(\alpha(x-\beta))e^{-\frac{1}{2}\alpha^2(x-\beta)^2}, \tag{5.2}$$

for $n \in \mathbb{Z}$ and $n \geq 0$, where $\alpha > 0$, $\beta \in \mathbb{R}$ are constants, namely the scaling factor and the translating factor, respectively. It is easy to derive the following properties for the generalized Hermite functions:

1. The $\{H_n^{\alpha,\beta}\}_{n=0}^{\infty}$ forms an orthogonal basis of $L^2(\mathbb{R})$, i.e.

$$\int_{\mathbb{R}} H_n^{\alpha,\beta}(x) H_m^{\alpha,\beta}(x) dx = \frac{\sqrt{\pi}}{\alpha} \delta_{nm}, \tag{5.3}$$

   where δ_{nm} is the Kronecker function.

2. $H_n^{\alpha,\beta}(x)$ is the n^{th} eigenfunction of the following Strum-Liouville problem

$$e^{\frac{1}{2}\alpha^2(x-\beta)^2} \partial_x (e^{-\alpha^2(x-\beta)^2} \partial_x (e^{\frac{1}{2}\alpha^2(x-\beta)^2} u(x))) + \lambda_n u(x) = 0, \tag{5.4}$$

   with the corresponding eigenvalue $\lambda_n = 2\alpha^2 n$.

3. By convention, $H_n^{\alpha,\beta} \equiv 0$, for $n < 0$. For $n \in \mathbb{Z}$ and $n \geq 0$, the following three-term recurrence holds:

$$\begin{aligned} 2\alpha(x-\beta) H_n^{\alpha,\beta}(x) =& \sqrt{2n} H_{n-1}^{\alpha,\beta}(x) + \sqrt{2(n+1)} H_{n+1}^{\alpha,\beta}(x); \\ \text{or} \quad 2\alpha^2(x-\beta) H_n^{\alpha,\beta}(x) =& \sqrt{\lambda_n} H_{n-1}^{\alpha,\beta}(x) + \sqrt{\lambda_{n+1}} H_{n+1}^{\alpha,\beta}(x). \end{aligned} \tag{5.5}$$

4. The derivative of $H_n^{\alpha,\beta}(x)$ is a linear combination of $H_{n-1}^{\alpha,\beta}(x)$ and $H_{n+1}^{\alpha,\beta}(x)$:

$$\partial_x H_n^{\alpha,\beta}(x) = \frac{1}{2}\sqrt{\lambda_n} H_{n-1}^{\alpha,\beta}(x) - \frac{1}{2}\sqrt{\lambda_{n+1}} H_{n+1}^{\alpha,\beta}(x) = \sqrt{\frac{n}{2}}\alpha H_{n-1}^{\alpha,\beta}(x) - \sqrt{\frac{n+1}{2}}\alpha H_{n+1}^{\alpha,\beta}(x). \tag{5.6}$$

5. Properties 1) and 4) yield the "orthogonality" of $\{\partial_x H_n^{\alpha,\beta}(x)\}_{n=0}^{\infty}$:

$$\int_{\mathbb{R}} \partial_x H_n^{\alpha,\beta}(x)\partial_x H_m^{\alpha,\beta}(x)dx = \begin{cases} \sqrt{\pi}\alpha(n+\frac{1}{2}) = \frac{\sqrt{\pi}}{4\alpha}(\lambda_n + \lambda_{n+1}), & \text{if } m = n; \\ -\frac{\alpha}{2}\sqrt{\pi(l+1)(l+2)} = -\frac{\sqrt{\pi}}{4\alpha}\sqrt{\lambda_{l+1}\lambda_{l+2}}, & \\ \qquad\qquad l = \min\{n,m\}, \quad \text{if } |n-m| = 2; & \\ 0, \quad \text{otherwise.} \end{cases} \tag{5.7}$$

The generalized Hermite functions form a complete orthogonal basis in $L^2(\mathbb{R})$. That is, any function $u \in L^2(\mathbb{R})$ can be written in the form

$$u(x) = \sum_{n=0}^{\infty} \hat{u}_n H_n^{\alpha,\beta}(x),$$

where $\{\hat{u}_n\}_{n=0}^{\infty}$ are the Fourier-Hermite coefficients, given by

$$\hat{u}_n = \frac{\alpha}{\sqrt{\pi}} \int_{\mathbb{R}} u(x) H_n^{\alpha,\beta}(x)dx. \tag{5.8}$$

Let us denote the subspace spanned by the first $N+1$ generalized Hermite functions as $\mathcal{R}_N$:

$$\mathcal{R}_N = \text{span}\left\{H_0^{\alpha,\beta}(x), \cdots, H_N^{\alpha,\beta}(x)\right\}. \tag{5.9}$$

We follow the convection in the asymptotic analysis that $a \sim b$ means that there exists some constants $C_1, C_2 > 0$ such that $C_1 a \leq b \leq C_2 a$; $a \lesssim b$ means that there exists some constant $C_3 > 0$ such that $a \leq C_3 b$.

It is shown in (57) that for $\alpha > 0$, $\beta = 0$ the difference between an arbitrary function and its orthogonal projection onto $\mathcal{R}_N$ in some suitable function space could be precisely estimated in terms of the scaling factor α and the truncation mode N. Let us first introduce the function space $W^r_{\alpha,\beta}(\mathbb{R})$, for any integer $r \geq 0$,

$$W^r_{\alpha,\beta}(\mathbb{R}) := \left\{u \in L^2(\mathbb{R}) : \, ||u||_{r,\alpha,\beta} < \infty, ||u||^2_{r,\alpha,\beta} := \sum_{k=0}^{\infty} \lambda^r_{k+1} \hat{u}_k^2\right\}, \tag{5.10}$$

where λ_k is in Equation 6.3 and $\hat{u}_k$ is the Fourier-Hermite coefficient in Equation 5.8. We shall denote the space by $W^r(\mathbb{R})$ for short, hoping that no confusion will arise. Also, the norms will be denoted briefly as $||\cdot||_r$. The larger r is, the smaller the space $W^r(\mathbb{R})$ is, and the smoother the functions in $W^r(\mathbb{R})$ are. The index r can be viewed as the indicator of the regularity of the functions.

Let us define the L^2-orthogonal projection $P_N^{\alpha,\beta} : \, L^2(\mathbb{R}) \to \mathcal{R}_N$, of a given $v \in L^2(\mathbb{R})$, by

$$\langle v - P_N^{\alpha,\beta} v, \phi \rangle = 0, \quad \forall \phi \in \mathcal{R}_N. \tag{5.11}$$

The superscript α, β will be dropped in $P_N^{\alpha,\beta}$ in the sequel, since no confusion should arise. More precisely,

$$P_N v(x) := \sum_{n=0}^{N} \hat{v}_n H_n^{\alpha,\beta}(x),$$

where $\hat{v}_n$ are the Fourier-Hermite coefficients defined in Equation 5.8. The truncated error $||u - P_N u||_r$, for any integer $r \geq 0$, has been essentially estimated in Theorem 2.3 in (27), for $\alpha = 1$, $\beta = 0$, and in Theorem 2.1 in (57), for arbitrary $\alpha > 0$ and $\beta = 0$. For arbitrary $\alpha > 0$ and $\beta \neq 0$, the estimate still holds.

Theorem 5.1. *For any $u \in W^r(\mathbb{R})$ and any integer $0 \leq \mu \leq r$, we have*

$$|u - P_N u|_\mu \lesssim \alpha^{\mu - r - \frac{1}{2}} N^{\frac{\mu - r}{2}} ||u||_r, \tag{5.12}$$

where $|u|_\mu := ||\partial_x^\mu u||$ are the seminorms, if $N \gg 1$.

Proof. The proof is extremely similar to those in (27) and (57). We use induction, and first establish it for $\mu = 0$. For any integer $r \geq 0$,

$$||u - P_N u||^2 = \frac{\sqrt{\pi}}{\alpha} \sum_{n=N+1}^{\infty} \hat{u}_n^2 = \frac{\sqrt{\pi}}{\alpha} \sum_{n=N+1}^{\infty} \lambda_{n+1}^{-r} \lambda_{n+1}^{r} \hat{u}_n^2 \lesssim \alpha^{-2r-1} N^{-r} ||u||_r^2. \tag{5.13}$$

Suppose that for $1 \leq \mu \leq r$, Equation 5.12 holds for $\mu - 1$. We need to show that Equation 5.12 is also valid for μ. It is clear that

$$|u - P_N u|_\mu \leq |\partial_x u - P_N \partial_x u|_{\mu-1} + |P_N \partial_x u - \partial_x P_N u|_{\mu-1}. \tag{5.14}$$

On the one hand, due to the assumption for $\mu - 1$, we apply Equation 5.12 to $\partial_x u$ and replace μ and r with $\mu - 1$ and $r - 1$, respectively:

$$|\partial_x u - P_N \partial_x u|_{\mu-1} \leq \alpha^{\mu-r-\frac{1}{2}} N^{\frac{\mu-r}{2}} ||\partial_x u||_{r-1} \lesssim \alpha^{\mu-r-\frac{1}{2}} N^{\frac{\mu-r}{2}} ||u||_r, \tag{5.15}$$

where the last inequality holds because of the observation, that

$$||\partial_x u||_{r-1}^2 = \sum_{n=0}^{\infty} \lambda_{n+1}^{r-1} \widehat{(\partial_x u)}_n^2$$

and

$$\begin{aligned}
\widehat{(\partial_x u)}_n =& \frac{\alpha}{\sqrt{\pi}} \int_{\mathbb{R}} \partial_x u H_n^{\alpha,\beta}(x) dx = -\frac{\alpha}{\sqrt{\pi}} \int_{\mathbb{R}} u \partial_x H_n^{\alpha,\beta}(x) dx \\
=& \frac{\alpha\sqrt{\lambda_{n+1}}}{2\sqrt{\pi}} \int_{\mathbb{R}} u H_{n+1}^{\alpha,\beta}(x) dx - \frac{\alpha\sqrt{\lambda_n}}{2\sqrt{\pi}} \int_{\mathbb{R}} u H_{n-1}^{\alpha,\beta}(x) dx \\
=& \frac{\sqrt{\lambda_{n+1}}}{2} \hat{u}_{n+1} - \frac{\sqrt{\lambda_n}}{2} \hat{u}_{n-1}.
\end{aligned}$$

Here we used integration by parts and Equation 5.6. On the other hand, by virtue of Equation 5.6

$$\begin{aligned}
P_N\partial_x u - \partial_x P_N u =& P_N \sum_{n=0}^{\infty} \hat{u}_n \partial_x H_n^{\alpha,\beta}(x) - \sum_{n=0}^{N} \hat{u}_n \partial_x H_n^{\alpha,\beta}(x) \\
=& -\frac{1}{2}\sum_{n=0}^{N-1} \sqrt{\lambda_{n+1}}\hat{u}_n H_{n+1}^{\alpha,\beta}(x) + \frac{1}{2}\sum_{n=0}^{N+1}\sqrt{\lambda_n}\hat{u}_n H_{n-1}^{\alpha,\beta}(x) \\
& - \left[-\frac{1}{2}\sum_{n=0}^{N} \sqrt{\lambda_{n+1}}\hat{u}_n H_{n+1}^{\alpha,\beta}(x) + \frac{1}{2}\sum_{n=0}^{N}\sqrt{\lambda_n}\hat{u}_n H_{n-1}^{\alpha,\beta}\right] \\
=& \frac{1}{2}\sqrt{\lambda_{N+1}}\left[\hat{u}_N H_{N+1}^{\alpha,\beta}(x) + \hat{u}_{N+1} H_N^{\alpha,\beta}(x)\right].
\end{aligned}$$

This yields that

$$|P_N\partial_x u - \partial_x P_N u|_{\mu-1}^2 \lesssim \lambda_{N+1}\left(\hat{u}_N^2 |H_{N+1}^{\alpha,\beta}(x)|_{\mu-1}^2 + \hat{u}_{N+1}^2 |H_N^{\alpha,\beta}(x)|_{\mu-1}^2\right), \tag{5.16}$$

due to the property of seminorms. Moreover, we estimate $\hat{u}_k^2$ and $|H_k^{\alpha,\beta}(x)|_{\mu-1}^2$, for $k = N, N+1$:

$$\hat{u}_N^2 \leq \sum_{n=N}^{\infty} \hat{u}_n^2 \leq \frac{\alpha}{\sqrt{\pi}}||u - P_{N-1}u||^2 \lesssim \alpha^{-2r}N^{-r}||u||_r^2, \tag{5.17}$$

by Equation 5.13. Similarly, $\hat{u}_{N+1}^2 \lesssim \alpha^{-2r}N^{-r}||u||_r^2$. And

$$|H_N^{\alpha,\beta}|_{\mu-1}^2 = ||\partial_x^{\mu-1}H_N^{\alpha,\beta}(x)||^2 \lesssim \alpha^{-1}||H_N^{\alpha,\beta}(x)||_{\mu-1}^2 = \alpha^{-1}\lambda_N^{\mu-1} \leq \alpha^{-1}\lambda_{N+1}^{\mu-1}, \tag{5.18}$$

by Lemma 5.2, since $\widehat{(H_N^{\alpha,\beta})}_k = \delta_{kN}$, for $k \in \mathbb{Z}^+$. Similarly, $|H_{N+1}^{\alpha,\beta}|_{\mu-1}^2 \lesssim \alpha^{-1}\lambda_{N+1}^{\mu-1}$. Substituting Equation 5.17 and Equation 5.18 into Equation 5.16, we get

$$|P_N\partial_x u - \partial_x P_N u|_{\mu-1}^2 \lesssim \alpha^{-2r-1}N^{-r}\lambda_{N+1}^{\mu}||u||_r^2 \lesssim \alpha^{2\mu-2r-1}N^{\mu-r}||u||_r^2, \tag{5.19}$$

by the fact that $\lambda_N = 2N\alpha^2$. The conclusion follows immediately from Equation 5.14, Equation 5.15 and Equation 5.19. □

5.2 Hermite spectral method to 1D forward Kolmogorov equation (FKE)

The general 1D FKE is in the form

$$\begin{cases} u_t(x,t) = p(x,t)u_{xx}(x,t) + q(x,t)u_x(x,t) + r(x,t)u(x,t), \quad \text{for } (x,t) \in \mathbb{R}\times\mathbb{R}_+ \\ u(x,0) = \sigma_0(x). \end{cases} \tag{5.20}$$

The well-posedness of 1D FKE has been investigated in (5). We state its key result here.

Lemma 5.1 (Besala, (5))**.** *Let $p(x,t)$, $q(x,t)$, $r(x,t)$ (real valued) together with p_x, p_{xx}, q_x be locally Hölder continuous in $\mathcal{D} = (t_0,t_1)\times\mathbb{R}$. Assume that*

1. *$p(x,t) \geq \lambda > 0$, $\forall\,(x,t) \in \mathcal{D}$, for some constant λ;*
2. *$r(x,t) \leq 0$, $\forall\,(x,t) \in \mathcal{D}$;*
3. *$(r - q_x + p_{xx})(x,t) \leq 0$, $\forall\,(x,t) \in \mathcal{D}$.*

Then the Cauchy problem Equation 5.20 with the initial condition $u(x,t_0)=u_0(x)$ has a fundamental solution $\Gamma(x,t;z,s)$ which satisfies

$$0 \le \Gamma(x,t;z,s) \le c(t-s)^{-\frac{1}{2}}$$

for some constant c and

$$\int_{-\infty}^{\infty} \Gamma(x,t;z,s)dz \le 1; \quad \int_{-\infty}^{\infty} \Gamma(x,t;z,s)dx \le 1.$$

Moreover, if $u_0(x)$ is continuous and bounded, then

$$u(x,t) = \int_{-\infty}^{\infty} \Gamma(x,t;z,t_0)u_0(z)dz$$

is a bounded solution of Equation 5.20.

Through the transformation

$$w(x,t) = e^{\frac{1}{2}\int_{-\infty}^{x} \tilde{q}(s,t)ds} u\left(\int_{-\infty}^{x} p^{\frac{1}{2}}(s,t)ds, t\right), \tag{5.21}$$

where

$$\tilde{q}(x,t) = p^{-\frac{1}{2}}(x,t)\left[q(x,t) - \frac{1}{2}p^{-\frac{1}{2}}p_x(x,t) + \frac{1}{2}\int_{-\infty}^{x} p^{-\frac{1}{2}}p_t(s,t)ds\right], \tag{5.22}$$

Equation 5.20 can be simplified to the following FKE, with the diffusion coefficient equal to 1 and without a convection term:

$$\begin{cases} w_t(x,t) = w_{xx}(x,t) + V(x,t)w(x,t), \quad \text{for } \mathbb{R} \times \mathbb{R}_+ \\ w(x,0) = w_0(x), \end{cases} \tag{5.23}$$

where

$$V(x,t) = \left[-\frac{1}{4}\tilde{q}^2(x,t) - \frac{1}{2}\tilde{q}_x(x,t) + \frac{1}{2}\int_{-\infty}^{x} \tilde{q}_t(s,t)ds + r(x,t) \right]. \tag{5.24}$$

Remark 5.1. *From the computational point of view, the form in Equation 5.23 is superior to the original form in Equation 5.20 in general, when implementing with the HSM.*

(i) If both the potential $V(x,t)$ and the initial data $w(x,0)$ are even functions in x, so is the solution to Equation 5.23. With the fact that the odd modes of the Fourier-Hermite coefficients of the even functions are identically zeros, it requires half amount of computations to resolve the even functions.

(ii) Even when $V(x,t)$ and $w(x,0)$ are not even, it is still wise to get rid of the convection term, since this term will drive the states to left and right, and probably out of the current "window". Shifting of the windows frequently by the moving-window technique will definitely affect the computational efficiency.

5.2.1 Formulation and convergence analysis

Let us consider the FKE Equation 5.23 with some source term $F(x,t)$. Let us use u instead of w in Equation 5.23. The weak formulation of HSM is to find $u_N(x,t) \in \mathcal{R}_N$ such that

$$\begin{cases} \langle \partial_t u_N(x,t), \varphi \rangle = - \langle \partial_x u_N(x,t), \varphi_x \rangle + \langle V(x,t) u_N(x,t), \varphi \rangle + \langle F(x,t), \varphi \rangle, \\ \qquad u_N(x,0) = P_N u_0(x), \end{cases} \tag{5.25}$$

for all $\varphi \in \mathcal{R}_N$. The convergence rate is stated below:

Theorem 5.2. *Assume*

$$-(1+|x|^2)^\gamma \lesssim V(x,t) \leq C,$$

for all $(x,t) \in \mathbb{R} \times (0,T)$*, for some* $\gamma > 0$ *and some constant* C*. If* $u_0 \in W^r(\mathbb{R})$ *and* u *is the solution to Equation 5.23 with source term* $F(x,t)$*, then for* $u \in L^\infty(0,T;W^r(\mathbb{R})) \cap L^2(0,T;W^r(\mathbb{R}))$ *with* $r > 2\gamma$ *and*

$$N \gg \max\left\{ \alpha^{\frac{4\gamma-2r+2}{2\gamma-1}} \max\left\{(\alpha\beta)^{4\gamma}, 1\right\}^{\frac{1}{1-2\gamma}}, \alpha^{2-\frac{r}{\gamma}} \max\left\{(\alpha\beta)^{4\gamma}, 1\right\}^{-\frac{1}{2\gamma}} \right\},$$

we have

$$||u - u_N||^2(t) \lesssim c^* \alpha^{-4\gamma-1} \max\left\{(\alpha\beta)^{4\gamma}, 1\right\} N^{2\gamma-r}, \tag{5.26}$$

where c^* *depends only on* T*,* $||u||_{L^\infty(0,T;W^r(\mathbb{R}))}$ *and* $||u||_{L^2(0,T;W^r(\mathbb{R}))}$.

Before we prove Theorem 5.2, we need some estimate on $||x^{r_1}\partial_x^{r_2}u(x)||^2$, for any integers $r_1, r_2 \geq 0$:

Lemma 5.2. *For any function $u \in W^{r_1+r_2}(\mathbb{R})$, with some integers $r_1, r_2 \geq 0$, we have*

$$||x^{r_1}\partial_x^{r_2}u||^2 \lesssim \alpha^{-2r_1-1}\max\{(\alpha\beta)^{2r_1}, 1\}||u||^2_{r_1+r_2}. \tag{5.27}$$

Proof. For any integers $r_1, r_2 \geq 0$,

$$||x^{r_1}\partial_x^{r_2}u||^2 = \left|\left|\sum_{n=0}^{\infty}\hat{u}_n x^{r_1}\partial_x^{r_2}H_n^{\alpha,\beta}(x)\right|\right|^2 \sim \left|\left|\frac{1}{\alpha^{2r_1}}\sum_{n=0}^{\infty}\hat{u}_n\sum_{k=-r_2-r_1}^{r_2+r_1}a_{n,k}H_{n+k}^{\alpha,\beta}(x)\right|\right|^2,$$

by Equation 5.5 and Equation 5.6, where for each n fixed, $a_{n,k}$ is a product of $2(r_1+r_2)$ factors of $\alpha^2\beta$ or $\sqrt{\lambda_{n+j}}$, with $-r_2-r_1 \leq j \leq r_2+r_1$. Let $n^* \geq 0$ such that $\alpha^2\beta \sim \sqrt{\lambda_{n^*+1}}$. And notice that $\lambda_{n+j} \sim \lambda_{n+1}$ for $n+j \geq 0$ and $H_{n+j}^{\alpha,\beta}(x) \equiv 0$ for $n+j < 0$. Hence, we have

$$\begin{aligned}||x^{r_1}\partial_x^{r_2}u(x)||^2 &\lesssim \alpha^{-1}\beta^{2r_1}\sum_{n=0}^{n^*}\lambda_{n+1}^{r_2+r_1}\hat{u}_n^2 + \alpha^{-2r_1-1}\sum_{n=n^*+1}^{\infty}\lambda_{n+1}^{r_2+r_1}\hat{u}_n^2\\ &\leq \alpha^{-2r_1-1}\max\{(\alpha\beta)^{2r_1}, 1\}||u||^2_{r_1+r_2},\end{aligned}$$

for any integers $r_1, r_2 \geq 0$, by Equation 5.3. □

Proof of Theorem 5.2. Denote $U_N = P_N u$ for simplicity. By Equation 5.23 with source term $F(x,t)$ and the definition of U_N, we obtain that

$$0 = \langle \partial_t (u - U_N), \varphi \rangle = -\langle u_x, \varphi_x \rangle + \langle V(x,t)u, \varphi \rangle + \langle F(x,t), \varphi \rangle - \langle \partial_t U_N, \varphi \rangle$$

$$\Rightarrow \quad \langle \partial_t U_N, \varphi \rangle = -\langle u_x, \varphi_x \rangle + \langle V(x,t)u, \varphi \rangle + \langle F(x,t), \varphi \rangle, \tag{5.28}$$

for all $\varphi \in \mathcal{R}_N$. Combining the above with Equation 5.25, yields that

$$\langle \partial_t (u_N - U_N), \varphi \rangle = -\langle \partial_x (u_N - u), \varphi_x \rangle + \langle V(x,t)(u_N - u), \varphi \rangle,$$

for all $\varphi \in \mathcal{R}_N$. Set $\varrho_N = u_N - U_N$. Choosing the function $\varphi = 2\varrho_N$, we have

$$\partial_t ||\varrho_N||^2 = -2||\partial_x \varrho_N||^2 - 2\langle \partial_x (U_N - u), \partial_x \varrho_N \rangle + 2\langle V(x,t)\varrho_N, \varrho_N \rangle + 2\langle V(x,t)(U_N - u), \varrho_N \rangle. \tag{5.29}$$

It follows from Young's inequality that

$$|\langle \partial_x (U_N - u), \partial_x \varrho_N \rangle| \leq \frac{1}{4}||\partial_x (U_N - u)||^2 + ||\partial_x \varrho_N||^2. \tag{5.30}$$

The assumption $V(x,t) \leq C$ for $(x,t) \in \mathbb{R} \times (0,T)$ then yields

$$\langle V(x,t)\varrho_N, \varrho_N \rangle \leq C||\varrho_N||^2, \tag{5.31}$$

for $(x,t) \in \mathbb{R} \times (0,T)$. Moreover, we have

$$|\langle V(x,t)(U_N - u), \varrho_N \rangle| \leq \frac{1}{2}||V(U_N - u)||^2 + \frac{1}{2}||\varrho_N||^2, \tag{5.32}$$

by the Cauchy-Schwartz inequality. Substituting Equation 5.30 - Equation 5.32 into Equation 5.29, we obtain

$$\partial_t||\varrho_N||^2 - (C+1)\,||\varrho_N||^2 \leq ||V(U_N - u)||^2 + \frac{1}{2}||\partial_x(U_N - u)||^2. \tag{5.33}$$

Notice that $V \gtrsim -(1+|x|^2)^\gamma$, for some $\gamma > 0$. By the estimate in Lemma 5.2, we have

$$\begin{aligned} ||V(U_N - u)||^2 &\lesssim ||(1+|x|^2)^\gamma (U_N - u)||^2 \lesssim ||(x^{2\gamma}+1)(U_N - u)||^2 \\ &\lesssim \alpha^{-4\gamma-1} \max\{(\alpha\beta)^{4\gamma}, 1\} \sum_{n=N+1}^{\infty} \lambda_{n+1}^{2\gamma} \hat{u}_n^2 + ||U_N - u||^2 \\ &\lesssim \alpha^{-4\gamma-1} \max\{(\alpha\beta)^{4\gamma}, 1\} N^{2\gamma-r} ||u||_r^2 + \alpha^{-2r-1} N^{-r} ||u||_r^2. \end{aligned} \tag{5.34}$$

The estimate of the second term on the right-hand side of Equation 5.34 follows from Theorem 5.1. Again by Theorem 5.1, we obtain

$$||\partial_x(U_N - u)||^2 = |U_N - u|_1^2 \lesssim \alpha^{-2r+1} N^{1-r} ||u||_r^2. \tag{5.35}$$

Substituting Equation 5.34 and Equation 5.35 into Equation 5.33, we obtain

$$\partial_t ||\varrho_N||^2 - (C+1)||\varrho_N||^2 \lesssim \alpha^{-4\gamma-1} \max\{(\alpha\beta)^{4\gamma}, 1\} N^{2\gamma-r} ||u||_r^2,$$

provided that

$$N \gg \max\left\{\alpha^{\frac{4\gamma-2r+2}{2\gamma-1}} \max\{(\alpha\beta)^{4\gamma}, 1\}^{\frac{1}{1-2\gamma}}, \alpha^{2-\frac{r}{\gamma}} \max\{(\alpha\beta)^{4\gamma}, 1\}^{-\frac{1}{2\gamma}}\right\}.$$

Therefore, we have

$$||\varrho_N||^2(t) \lesssim \alpha^{-4\gamma-1} \max\{(\alpha\beta)^{4\gamma}, 1\} N^{2\gamma-r} \int_0^t e^{-(C+1)(t-s)} ||u||_r^2(s) ds.$$

By the triangle inequality and Theorem 5.1,

$$\begin{aligned} ||u - u_N||^2(t) &\leq ||\varrho_N||^2 + ||u - U_N||^2 \\ &\lesssim \alpha^{-4\gamma-1} N^{2\gamma-r} \left[||u||_r^2 + \max\{(\alpha\beta)^{4\gamma}, 1\} \int_0^t e^{-(C+1)(t-s)} ||u||_r^2(s) ds \right] \\ &\lesssim c^* \alpha^{-4\gamma-1} \max\{(\alpha\beta)^{4\gamma}, 1\} N^{2\gamma-r}, \end{aligned}$$

where c^* is a constant depending on $||u||_{L^\infty(0,T;W^r(\mathbb{R}))}$, $||u||_{L^2(0,T;W^r(\mathbb{R}))}$ and T. □

5.2.2 Guidelines of the scaling factor

From Theorem 5.1, it is known for sure that any function in $W^r(\mathbb{R})$ could be approximated well by the generalized Hermite functions, provided that the truncation N is large enough.

However, in practice, "sufficiently" large N challenges the computer capacity. To improve the resolution of Hermite functions with reasonably large N, we need the scaling factor α, as pointed out in (7). Many efforts have been made along this direction, refer to (6), (7), (55), etc. However, the optimal choice of α (with respect to the truncation error) is still an open problem. In this subsection, we give a practical guideline to choose an appropriate scaling factor for the Gaussian type and super-Gaussian type functions.

It is well known that, for smooth functions $f(x) = \sum_{n=0}^{\infty} \hat{f}_n H_n^{\alpha,\beta}(x)$, the exponential decay of the Fourier-Hermite coefficients $\left|\hat{f}_n\right|$ with respect to n implies that the infinite sum is dominated by the first N terms, that is,

$$\left| f(x) - \sum_{n=0}^{N} \hat{f}_n H_n^{\alpha,\beta}(x) \right| \approx \mathcal{O}\left(\hat{f}_{N+1}\right),$$

for $N \gg 1$. Thus, the suitable scaling factor is proposed to get the Fourier-Hermite coefficients decaying as fast as possible. Once the coefficient approaching the machine error (say 10^{-16}), many other factors such as the roundoff error will come into play. Hence, it is wise to truncate the series here. Therefore, we need some guidelines for choosing not only the suitable scaling factor α but also the corresponding truncation mode N.

Suppose the function $f(x)$ peaks in the neighborhood of the origin and behaves asymptotically as $e^{-p|x|^k}$ with some $p > 0$ and $k \geq 2$, as $|x| \to +\infty$. Our guidelines are motivated by the following observations:

1. The function f decays exponentially fast, as $|x| \to \infty$, so that $\hat{f}_n \approx \int_{-L}^{L} f(x) H_n^{\alpha,\beta}(x) dx$, provided L is large enough, due to Equation 5.8.

2. For the exact Gaussian function e^{-px^2}, $p > 0$, the optimal α is naturally to be $\sqrt{2p}$ with the truncated mode $N = 1$. In fact, with this choice, $e^{-px^2} = H_0^{\alpha,0}(x)$, e^{-px^2} is orthogonal to all the rest of $H_n^{\alpha,0}$, $n > 0$. That is, $(\widehat{e^{-px^2}})_0 \neq 0$ and $(\widehat{e^{-px^2}})_n \equiv 0$, $n \geq 1$. This suggests that the closer the asymptotical behavior of f is to $e^{-\frac{1}{2}\alpha^2 x^2}$, the faster the Fourier-Hermite coefficients decays, and the smaller the truncation mode N is.

3. It is natural to adopt the Gaussian-Hermite quadrature method to compute the Fourier-Hermite coefficients by Equation 5.8. The truncation mode N has to be chosen such that the roots of Hermite polynomial $\mathcal{H}_{N+1}$ cover the domain $[-\alpha L, \alpha L]$ where the integral Equation 5.8 is contributed most from both f and $H_n^{\alpha,0}$, $n = 0, \cdots, N$.

We describe our guidelines for the Gaussian type and the super-Gaussian type functions separately as follows.

Case I. Gaussian type, i.e. $f(x) \sim e^{-px^2}$, $p > 0$, as $|x| \to +\infty$.

1. $e^{-px^2} \sim e^{-\frac{1}{2}\alpha^2 x^2}$ as $|x| \to +\infty$, which yields $\alpha \approx \sqrt{2p}$;

2. The integrand in Equation 5.8 is approximately e^{-2px^2}. Using the machine error 10^{-16} to decide the domain of interest L, i.e. $e^{-2pL^2} \approx 10^{-16}$, it yields that $L \approx \sqrt{8p^{-1} \ln 10}$;

3. Determine the truncation mode N such that the roots of Hermite polynomial H_{N+1} covers approximately $(-\alpha L, \alpha L)$, where $\alpha L \approx 4\sqrt{\ln 10}$.

Case II. Super-Gaussian type, i.e. $f(x) \sim e^{-px^k}$, as $|x| \to +\infty$ for some $k > 2$, $p > 0$.

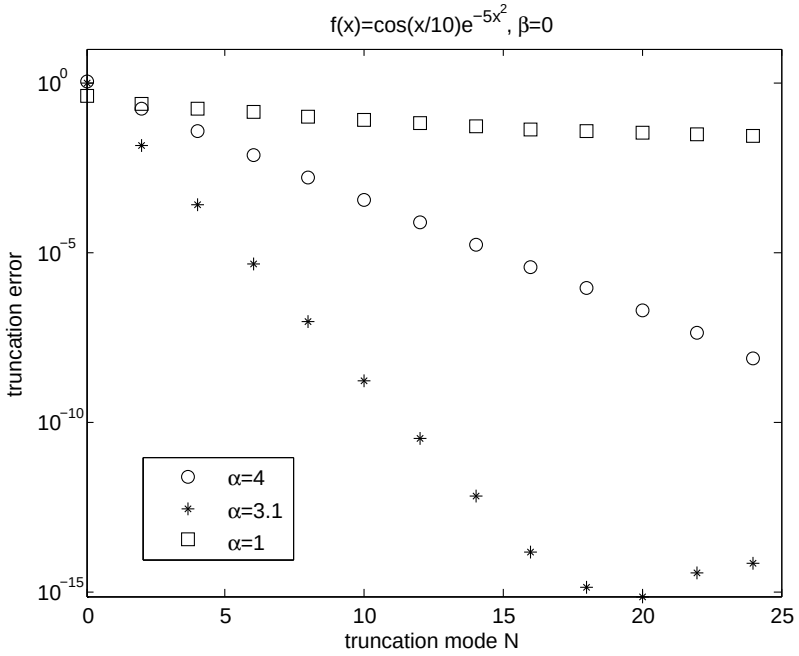

Figure 1. The truncation error v.s. the truncation mode for $f(x) = \cos\left(\frac{x}{10}\right)e^{-5x^2}$ is plotted, with $\beta = 0$ and $\alpha = 4, 3.1$ or 1.

1. Notice that $e^{-\frac{1}{2}\alpha^2 x^2} \gg e^{-px^k}$, when $x \gg 1$. Thus, we require that $e^{-\frac{1}{2}\alpha^2 x^2} \approx 10^{-16}$, which implies that $\alpha L \approx \sqrt{32 \ln 10}$;

2. We match $e^{-px^k} \approx e^{-\frac{1}{2}\alpha^2 x^2}$ near $x = \pm L$ yields that $\alpha \approx \sqrt{2p} L^{\frac{k}{2}-1}$. Hence, $L \approx (16p^{-1} \ln 10)^{\frac{1}{k}}$, $\alpha \approx 2^{\frac{5}{2}-\frac{4}{k}} p^{\frac{1}{k}} (\ln 10)^{\frac{1}{2}-\frac{1}{k}}$;

3. Determine the truncation mode N such that the roots of the Hermite polynomial H_{N+1} cover approximately $(-\alpha L, \alpha L)$.

To examine the feasibility of our guidelines, we explore the Gaussian type $f(x) = e^{-5x^2} \cos\left(\frac{x}{10}\right)$. According to the strategy in Case I, we choose the scaling factor $\alpha \approx \sqrt{10} \approx 3.1$, $L \approx \sqrt{\frac{8 \ln 10}{5}} \approx$

1.9194 and $N \approx 24$. As shown in Figure 1, the truncation error with $\alpha = 3.1$ decays the most rapidly with respect to the truncation mode N, and approaches the machine error near the 20^{th} mode. Meanwhile, the decay of the truncation error with $\alpha = 4$ and $\alpha = 1$ are much slower. Moreover, the truncation mode $N = 24$ is appropriate in the sense that the next few coefficients start to grow, due to the roundoff error.

Remark 5.2. *1) These guidelines are very practical. However, it is not the optimal scaling factor α. For example, if $f(x) = e^{-\frac{1}{2}x^2}$, then the optimal scaling factor $\alpha = 1$ and $N = 0$, while $N = 24$ is chosen according to our guidelines.*

2) Although the scaling factor helps to resolve the function concentrated in the neighborhood of the origin, it helps little if the function is peaked away from the origin. The numerical evidence could be found in Table I. This is the exact reason why we need to introduce the translating factor to the generalized Hermite functions when applying to the NLF problems.

5.2.3 Numerical verification of the convergence rate

To verify the convergence rate of HSM shown in Theorem 5.2, we explore a 1D FKE with some source $F(x,t)$. The exact solution could be found explicitly and is served as our benchmark. We consider the 1D FKE

$$\begin{cases} u_t = u_{xx} - x^2 u + (\sin t + \cos t + 3x)e^{-\frac{1}{2}x^2} \\ u(x,0) = xe^{-\frac{1}{2}x^2}, \end{cases} \tag{5.36}$$

for $(x,t) \in \mathbb{R} \times [0,T]$. It is easy to verify that $u(x,t) = (x + \sin t)e^{-\frac{1}{2}x^2}$ is the exact solution.

Notice that the initial data, the potential and the source in Equation 5.36 are all concentrated around the origin. So, we set the translating factor $\beta = 0$. For notational convenience, we drop β in this example. As to the suitable scaling factor α, from our strategy in section 5.2.2, we know that it is better to let $\alpha = 1$. However, if we do so, the first two modes will give us an extremely good approximation. Hence, the error v.s. the truncation mode won't be seen clearly. Due to this consideration, we pick $\alpha = 1.4$ (a little bit away from 1, but not too far away so that it won't affect the resolution too much). The weak formulation (Equation 5.25) yields

$$\langle \partial_t u_N, \varphi \rangle = -\langle \partial_x u_N, \partial_x \varphi \rangle - \langle x u_N, x\varphi \rangle + \langle F(x,t), \varphi \rangle, \tag{5.37}$$

for all $\varphi \in \mathcal{R}_N$. Take the test functions $\varphi = H_n^\alpha(x)$, $n = 0, 1, \cdots, N$, in Equation 5.37. Since $u_N \in \mathcal{R}_N$, it can be written in the form

$$u_N(x,t) = \sum_{n=0}^{N} a_n(t) H_n^\alpha(x).$$

The matrix form of Equation 5.37 follows from Equation 5.5 and Equation 5.7:

$$\frac{d}{dt}\vec{a}(t) = A\vec{a}(t) + \vec{f}(t), \tag{5.38}$$

where $\vec{a}(t) = (a_0(t), a_1(t), \cdots, a_N(t))^T$, $\vec{f}(t) = \left(\hat{f}_0(t), \hat{f}_1(t), \cdots, \hat{f}_N(t)\right)^T$ are column vectors with $N+1$ entries, $\hat{f}_i(t)$, $i = 0, 1, \cdots, N$, are the Fourier-Hermite coefficients of $F(x,t)$ and A is a penta-diagonal $(N+1) \times (N+1)$ constant matrix, where $A = -A_1 - A_2$, with

$$A_1(i,j) = \begin{cases} -\dfrac{\alpha^2}{2}\sqrt{(k+1)(k+2)}, \ k = \min\{i,j\}, \quad |i-j| = 2, \\ \alpha^2\left(i + \dfrac{1}{2}\right), \quad i = j, \\ 0, \quad \text{otherwise}, \end{cases}$$

and

$$A_2(i,j) = \begin{cases} \dfrac{\sqrt{(k+1)(k+2)}}{2\alpha^2}, \ k = \min\{i,j\}, \quad |i-j| = 2, \\ \dfrac{(2i+1)}{2\alpha^2}, \quad i = j, \\ 0, \quad \text{otherwise}. \end{cases}$$

The L^2 errors v.s. the truncation mode N at time $T = 0.1$ is plotted in Figure 2. The ODE Equation 5.38 is numerically solved by central difference scheme in time with the time step $dt = 10^{-5}$. It indeed illustrates the spectral accuracy of HSM.

5.3 Application to nonlinear filtering problems

Recalling the brief description of our algorithm in Chapter 2, the off-line computation is to numerically solve the FKE Equation 2.13 repeatedly on each interval $[\tau_i, \tau_{i+1}]$. Equation 2.13 is in the form of Equation 5.20 with

$$p(x,t) = \frac{1}{2}Qg^2; \quad q(x,t) = Q(g^2)_x - f_x; \quad r(x,t) = -\frac{1}{2}h^2/S + Q(g_x^2 + gg_{xx}) - f_x,$$

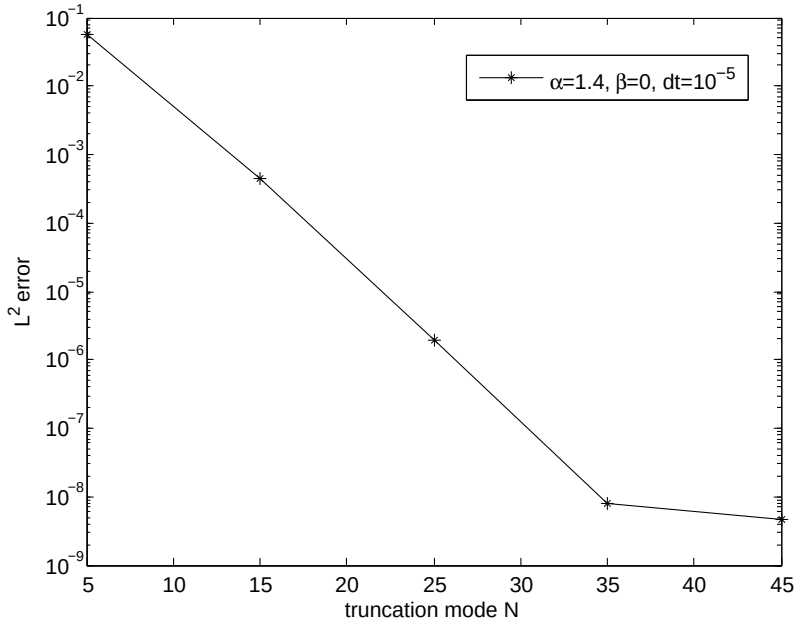

Figure 2. The L^2-errors of the HSM to FKE Equation 5.36 v.s. the truncation mode $N = 5, 15, 25, 35$ and 45 is plotted, with $\alpha = 1.4$, $\beta = 0$ and the time step $dt = 10^{-5}$.

where Q, S, f, g and h are in Equation 2.1.

5.3.1 Existence and uniqueness of the solution to 1D FKE

We interpret the well-posedness theorem, i.e. Lemma 5.1, for general 1D FKE in the framework of the NLF problems.

Proposition 5.1 (Existence)**.** *Let f, g, h in Equation 2.1 be Hölder continuous functions in $\mathcal{D} := \mathbb{R} \times (t_0, t_1)$. Also, assume that g_x, g_{xx} and f_x exist and are also Hölder continuous in $\mathcal{D}$. Assume further that*

1. *$Qg^2 \geq \lambda > 0$, for some $\lambda > 0$;*

2. $S > 0$*;*

3. $-\frac{1}{2}h^2/S - f_x + Q(g_x^2 + gg_{xx}) \leq C$*, for some constant* C*,*

for $(x,t) \in \mathcal{D}$*. Then there exists a bounded solution* $u(x,t)$ *to Equation 5.20, if the initial condition* $u_0(x)$ *is continuous and bounded.*

Proof. Conditions 1)-3) in Lemma 5.1 are directly translated into conditions 1)-3) in this proposition with $C \leq 0$. For $C > 0$, let $v(x,t) = e^{-C(t-t_0)}u(x,t)$, then v satisfies

$$v_t(x,t) = p(x,t)v_{xx}(x,t) + q(x,t)v_x(x,t) + (r(x,t) - C)v(x,t), \tag{5.39}$$

for $(x,t) \in \mathcal{D}$, with the initial condition $v(x,t_0) = u_0(x)$. The coefficients in Equation 5.39 satisfy the conditions in Lemma 5.1. Thus, we apply Lemma 5.1 directly to Equation 5.39. The existence of the solution to Equation 5.20 follows immediately. □

Remark 5.3. *In practice, the initial data of the conditional density function has either compact support or exponentially decay as* $|x| \to +\infty$*. So, the assumption on the initial data in Proposition 5.1 always holds.*

For simplicity, we establish the uniqueness for Equation 5.23, instead of Equation 5.20. They can be easily transformed into each other, due to the bijective transformation Equation 5.21.

Proposition 5.2 (Uniqueness)**.** *There exists a unique solution to Equation 5.23 in the class that* $\{u : \lim_{|x|\to\infty} uu_x = 0\}$ *if* $V(x,t)$ *is bounded from above in* $\mathcal{D}$*.*

Proof. Case I: Assume $V(x,t) \leq 0$ in $\mathcal{D}$. Suppose there exist two distinct solutions to Equation 5.23, say u_1 and u_2. Denote $\eta := u_1 - u_2$, and then η satisfies

$$\eta_t = \eta_{xx} + V(x,t)\eta, \tag{5.40}$$

in $\mathcal{D}$ with the initial condition $\eta(x,t_0) = 0$. Using the standard energy estimate, i.e. multiplying Equation 5.40 with η and integrating with respect to x in $\mathbb{R}$:

$$\frac{1}{2}||\eta||_t^2 = -||\eta_x||^2 + \int_{\mathbb{R}} V(x,t)\eta^2 dx \leq -||\eta_x||^2 \leq 0,$$

by integration by parts, and the facts that $\lim_{|x|\to\infty} \eta\eta_x = 0$ and $V(x,t) \leq 0$ in $\mathcal{D}$. This yields that

$$||\eta||^2(t) \leq ||\eta||^2(t_0),$$

for $t \in (t_0, t_1)$. With the fact that $\eta(x,t_0) = 0$, we conclude that $\eta \equiv 0$ in $\mathcal{D}$, i.e. $u_1 \equiv u_2$.

Case II: Assume $V(x,t) \leq C$, for some $C > 0$. We use the strategy in the proof of Proposition 5.1. Let $v(x,t) = e^{-C(t-t_0)}u(x,t)$, then v satisfies Equation 5.23 with the potential $V(x,t) - C \leq 0$ in $\mathcal{D}$. By case I, we conclude the uniqueness of v, and thus that of u. □

Remark 5.4. *Similar conditions as in Proposition 5.1 were used to guarantee the well-posedness of the "pathwise-robust" DMZ equation (see Chapter 3) and to establish the convergence of our algorithm (see Chapter 4). They essentially require that h has to grow faster than f. They*

are not restrictive in the sense that most of the polynomial sensors are included. For example, $f(x) = f_0 x^j$, $g(x) = g_0(1 + x^2)^k$ *and* $h(x) = h_0 x^l$, *with* $S, Q > 0$, f_0, g_0 *and* h_0 *are constants,* $j, k, l \in \mathbb{N}$, *provided* $l > \max\left\{\frac{j-1}{2}, 2k - 1\right\}$.

5.3.2 Translating factor β and moving-window technique

As we mentioned before, the untranslated Hermite functions with suitable scaling factor could resolve functions concentrated in the neighborhood of the origin accurately and effectively. However, the states of the NLF problems could be driven to left and right during the on-line experiments. It is not hard to imagine that the "peaking" area of the density function escapes from the current "window". As numerical evidence, Figure 6 is the plot of the normalized density function of the cubic sensor.

The idea of the translating factor is that, under the circumstance that the function is peaking far away from the "window" covered by the current Hermite functions, we translate the current Hermite functions to the "support" of the function, by letting the translating factor β be near the "peaking" area of the function.

In Table I, we list the truncation error of the Gaussian function $f(x) = e^{-\frac{1}{2}(x-p_0)^2}$ with various $p_0 = -1, 0, \cdots, 4$ and different translating factors $\beta = 0$ or 3. The truncation errors with different translating factor β is denoted as error_β, which is defined as $||f - \sum_{n=0}^{N} \hat{f}_n H_n^{\alpha,\beta}||$. According to the guidelines in section 5.2.2, the scaling factor is $\alpha = 1$ and the truncation mode is $N = 24$. As shown in Table I, the further the function is peaking away from the origin, the larger the error is with untranslated Hermite functions. But with appropriate translating

p_0	error_0	error_3
-1	3.3×10^{-13}	1.1×10^{-3}
0	8.2×10^{-15}	7.7×10^{-6}
1	1.6×10^{-13}	1.8×10^{-9}
2	1.8×10^{-9}	3.3×10^{-13}
3	7.7×10^{-6}	8.2×10^{-15}
4	1.1×10^{-3}	1.6×10^{-13}

TABLE I

TRUNCTION ERROR V.S. THE "PEAKING" P_0 OF THE GAUSSIAN FUNCTION $F(X) = E^{-\frac{1}{2}(X-P_0)^2}$.

factor, the function could be resolved very well with the *same* scaling factor, for example, $\text{error}_3 \approx 10^{-16}$ for $f(x) = e^{-\frac{1}{2}(x-3)^2}$.

Indeed, this fact motivates the idea of a *moving-window technique*. The suitable width of the window could be pre-determined if the trunction error of the density function v.s. various "peaking" p_0 is investigated beforehand. To be more precise, suppose we know the asymptotic behavior of the density function of the NLF problem from the asymptotical analysis, say $\sim e^{-px^k}$, with some $p > 0$, $k \geq 2$. According to the guidelines in section 5.2.2, the suitable scaling factor α and the truncation mode N with $\beta = 0$ could be chosen. With these parameters, a table similar to Table I could be obtained, i.e. the truncation error (error_0) of the function $e^{-p(x-p_0)^k}$ v.s. various p_0. If the error tolerance is given, then the appropriate width of the window is obtained according to the table. Let us take Table I as an example. If the asymptotic behavior of the density function is $e^{-\frac{1}{2}x^2}$, then the scaling factor $\alpha = 1$ and the truncation mode

$N = 24$. Suppose we set the error tolerance to be 10^{-5}, then the suitable width of the window would be $3 + 3 = 6$, from the first two columns of Table I. The window, that covers the origin, would be $[-3, 3]$.

Our algorithm with the moving-window technique is illustrated in the flowchart Figure 3. It reads as follows. Without loss of generality, assume that the expectation of the initial distribution of the state is near 0. During the experimental time, say $[0, T]$, the state remains inside some bounded interval $[-L, L]$, for some $L > 0$. We first cover the neighborhood of 0 by the untranslated Hermite functions $\{H_n^{\alpha,0}\}_{n=0}^{N}$, where α, N can be chosen according to the guidelines in section 5.2.2. With the given error tolerance, the suitable width of the window could be pre-defined, denoted as L_w. If $[-L, L] \subset [-L_w, L_w]$, then no moving-window technique is needed. Hence, the on-line experiment runs always within the left half loop in Figure 3. Otherwise, $\{\beta_j\}_{j=0}^{J}$, for some $J > 0$, need to be prepared beforehand, such that $[-L, L] \subset \cup_{j=0}^{J} (-L_w + \beta_j, \beta_j + L_w)$. The off-line data corresponding to different intervals $(-L_w + \beta_j, \beta_j + L_w)$ have to be pre-computed and stored ahead of time. During the on-line experiment, if the expectation of the state $\mathbb{E}[x_t]$ moves across the boundary of the current "window" (the condition in the rhombic box in Figure 3 is satisfied), the current "window" is shifted to the nearby window, into which $\mathbb{E}[x_t]$ falls. That is, the right half loop in Figure 3 is performed once.

Let us analyze the computational cost of our algorithm. Notice that only the storage capacity of the off-line data and the number of the flops for on-line performance need to be taken into consideration in our algorithm. Without loss of generality, let us assume, as before,

that $\mathbb{E}[x](0)$ is near 0 and our state is inside $[-L, L] \subset \cup_{j=0}^{J} (-L_w + \beta_j, \beta_j + L_w)$. For simplicity and clarity, let us first assume further that

1. The operator $\left(L - \frac{1}{2} h^T S^{-1} h\right)$ is not explicitly time-dependent;

2. The time steps are the same, i.e. $\tau_{i+1} - \tau_i = \triangle t$.

The storage of the off-line data, on each interval $(-L_w + \beta_j, \beta_j + L_w)$, requires storing $(N+1)^2$ floating point numbers. Hence, the $(J+1)$ intervals requires to store $(J+1)(N+1)^2$ floating point numbers. As to the number of the flops in the on-line computations, if no moving-window technique is adopted during the experiment, for each time step, it requires $\mathcal{O}((N+1)^2)$ flops. The number of the flops to complete the experiment during $[0, T] = \cup_{i=0}^{k-1} [\tau_i, \tau_{i+1}]$ is $\mathcal{O}(k(N+1)^2)$. Suppose the number of window shifts during $[0, T]$ is P, then the total number of flops is $\mathcal{O}\left((k+P)(N+1)^2\right)$.

Remark 5.5. *Even if either assumption 1) or 2) is not satisfied, the real-time manner of our algorithm won't be affected. This is because the number of the flops in the on-line experiment remains the same. But the off-line data will take more storage as the trade-off. To be more specific, on each interval* $(-L_w + \beta_j, \beta_j + L_w)$*, it requires to store* $k \times (N+1)^2$ *floating point numbers, where* k *is the total number of time steps. Therefore, the total storage is* $k(J+1)(N+1)^2$ *floating point numbers.*

5.4 Numerical simulations

In this subsection, we shall validate our algorithm by solving three NLF problems: two "time-invariant" cases and one "time-varying" case. Our algorithm is compared with either the

extended Kalman filter (EKF) or the particle filters (PF). The particle filters are implemented based on the algorithm described in (1), and systematic resampling is adopted if the effective sample size drops below 50% of the total number of particles. As we shall see, to achieve similar accuracy our algorithm surpasses both the EKF and the PF in the real-time manner.

5.4.1 "time-invariant" case: 1D almost linear filter

The signal observation model we are considering here is

$$\begin{cases} dx_t = dv_t \\ \\ dy_t = x_t(1 + 0.25\cos x_t)dt + dw_t, \end{cases}$$

where $x_t,\ y_t \in \mathbb{R},\ v_t,\ w_t$ are scalar Brownian motion processes with $E[dv_t^T dv_t] = 1$ and $E[dw_t^T dw_t] = 1$. Suppose the signal at the beginning is somewhere near the origin.

The corresponding FKE Equation 2.13 in this case is

$$u_t = \frac{1}{2}u_{xx} - \frac{1}{2}x^2(1+\cos x)^2 u \tag{5.41}$$

Assume further that the initial distribution of x_0 is $u_0(x) = e^{\frac{-x^2}{2}}$. This assumption is not crucial at all. The non-Gaussian ones, for example $u_0(x) = e^{\frac{-x^4}{2}}$, will give the similar results as the Gaussian one.

It is easy to see that the asymptotic behavior of the solution to Equation 5.41 is $e^{-\frac{x^2}{2}}$. With the guidelines, we choose $\alpha = 1$, $\beta = 0$ and $N = 25$ for the starting interval. We shall run the experiment for the total time $T = 50$. Thus, we expect the density function probably will move

out of the starting interval. Table I suggests that the appropriate width of the window should be 3, if the error tolerance is set to be 10^{-5}. We shall overlap the adjacent windows a little bit to prevent frequent shifting of windows. Let us take the width of the overlaped region to be 0.5. Therefore, as the preparation for the moving-window technique, we shall prepare the off-line data for $[-19.5, -13.5]$, $[-14, -8]$, $[-8.5, -2.5]$, $[-3, 3]$, $[2.5, 8.5]$, $[8, 14]$ and $[13.5, 19.5]$. The correpsonding β's are $-16.5, -11, -5.5, 0, 5.5, 11$ and 16.5. The barrier in the rhombic box in the flowchart Figure 3 should be 3 (the width of the "window").

Our algorithm is compared with the PF with 10 or 50 particles in Figure 4 for the total experimental time $T = 50$. The time step is $\triangle t = 0.01$. All three filters show acceptable experimental results. It is clear (between time 10 to 30) that the PF with 50 particles gives closer estimation to our algorithm than that with 10 particles. But as to the efficiency, our algorithm is superior to the PF, since the CPU times of PF with 10 and 50 particles are 5.00s and 35.75s respectively, while that of our algorithm is only 2.62s. As to the storage, the size of the binary file to keep the off-line data is only 35.5kB. During this particular on-line experiment, the window has been shifted 13 times, which can't be seen from the figure at all. It also seems that the moving-window technique doesn't affect the efficiency of our algorithm.

5.4.2 "time-invariant" case: cubic sensor in the channel

We consider cubic sensor in the channel $x_t \in [-3, 3]$:

$$\begin{cases} dx_t = dv_t \\ \\ dy_t = x_t^3 dt + dw_t, \end{cases} \tag{5.42}$$

where x_t, $y_t \in \mathbb{R}$, v_t, w_t are scalar Brownian motion processes with $E[dv_t^T dv_t] = 1$, $E[dw_t^T dw_t] = 1$. Assume the initial state is somewhere near 0.

The FKE Equation 2.13 is

$$u_t = \frac{1}{2}u_{xx} - \frac{1}{2}x^6 u. \tag{5.43}$$

Furthermore, we assume the initial distribution is $u_0(x) = e^{-x^4/4}$. We set our translating factor $\beta = 0$ and the moving-window technique won't be used. According to the guidelines in section 5.2.2, we choose the scaling factor $\alpha \approx 2^{\frac{3}{2}} \left(\frac{\ln 10}{4}\right)^{\frac{1}{4}} \approx 2.4637$, and the truncated mode $N \approx 45$.

In Figure 5, we compare our algorithm with the PF with 50 particles for $T = 50$. The observation data come in every 0.01. Figure 5 reads that both filters work very well. The result of our algorithm nearly overlaps with that of the particle filter, for all times. However, the CPU time of our algorithm is 4.90s, while that of PF is 37.17s. With our algorithm, the on-line computational time for every estimation of the state is around 0.001s, which is 10 times less than the update time 0.01s. This indicates that our algorithm is indeed a real-time solver. The normalized density functions, which is defined as $\frac{u(x,t)}{\max_{x\in\mathbb{R}} u(x,t)}$, have been plotted every other 1s in Figure 6.

5.4.3 "time-varying" case: the 1D almost linear sensor

The 1D almost linear "time-varying" sensor we are considering is

$$\begin{cases} dx_t = [1 + 0.1\cos(20\pi t)]dv_t \\ dy_t = x_t[1 + 0.25\cos(x_t)]dt + dw_t, \end{cases} \tag{5.44}$$

where x_t, $y_t \in \mathbb{R}$, v_t, w_t are scalar Brownian motion processes with $E[dv_t^T dv_t] = E[dw_t^T dw_t] = 1$. The FKE Equation 2.13 in this example is

$$u_t = \frac{1}{2}[1 + 0.1\cos(20\pi t)]^2 u_{xx} - \frac{1}{2}x^2[1 + 0.25\cos(x)]^2 u,$$

with the initial data $u_0(x) = e^{-x^2/2}$ and the updated initial data

$$u_i(x, \tau_i) = e^{x^2[1+0.25\cos(x)]\cdot dy_t} u_{i-1}(x, \tau_i),$$

$i = 1, 2, \cdots, k$. In Figure 7, our algorithm tracks the state's expectation at least as well as the EKF. The total simulation time is $T = 60$, and the update time step is $dt = \tau_{i+1} - \tau_i = 0.01$. It costs our algorithm only around $3.17s$ to complete the simulation, i.e. the on-line computational time is less than $5 \times 10^{-4}s$, which is around 20 times shorter than the updated time.

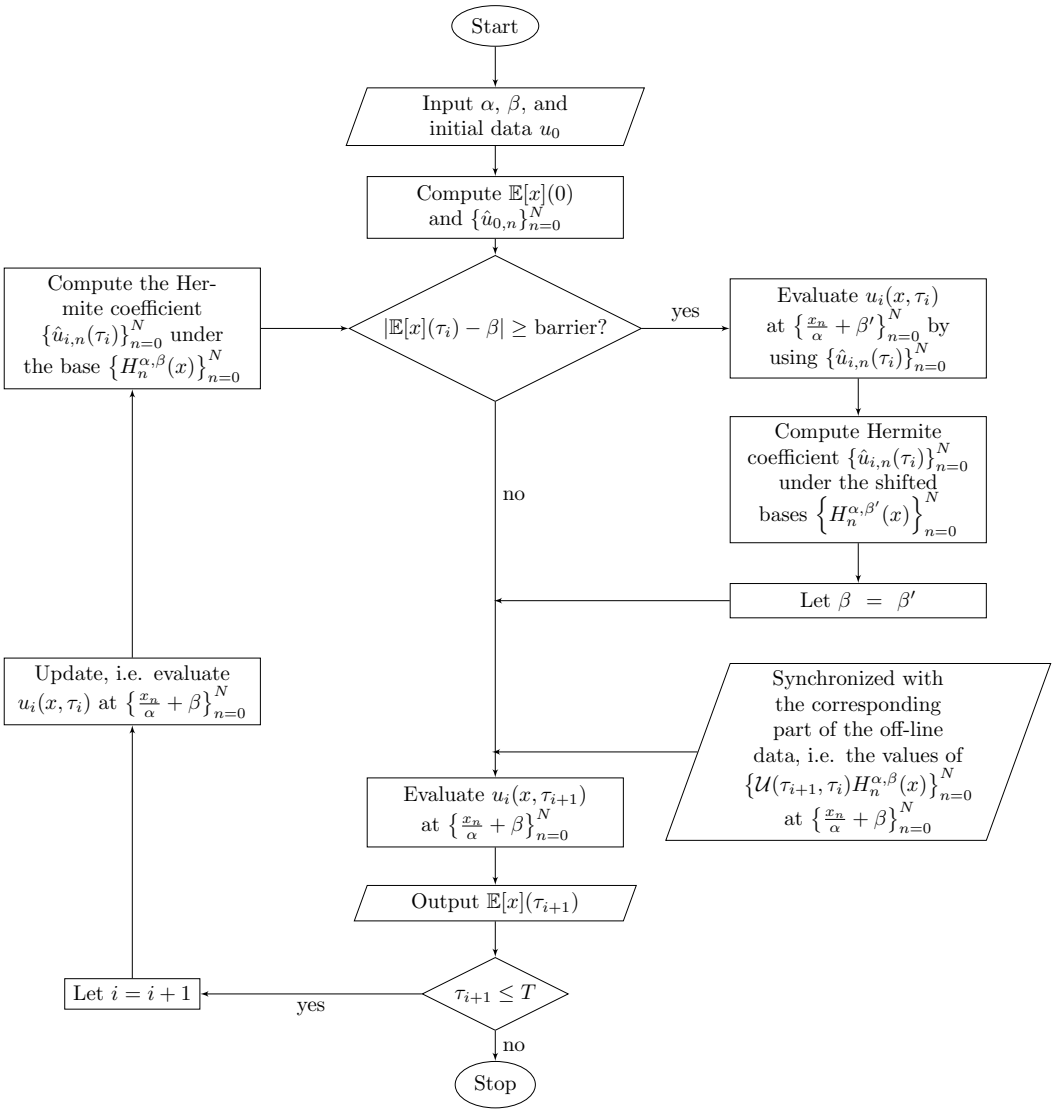

Figure 3. The flowchart of our algorithm, where $\beta' \in \{\beta_j\}_{j=0}^{J}$.

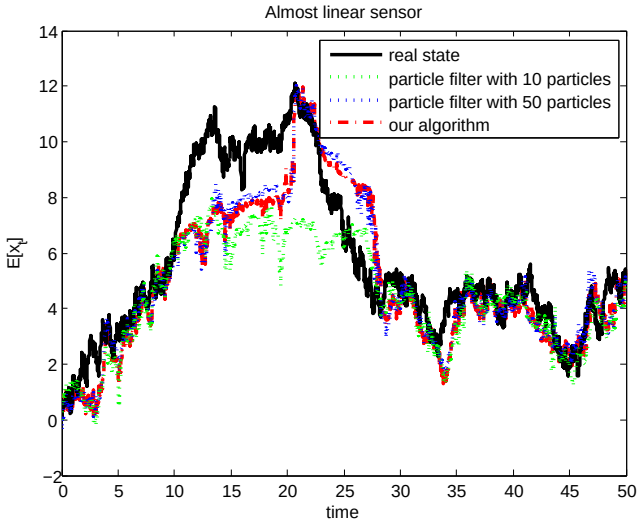

Figure 4. Almost linear filter is investigated with our algorithm and the particle filter with 10 and 50 particles. The total experimental time is $T = 50s$. And the update time is $\triangle t = 0.01$.

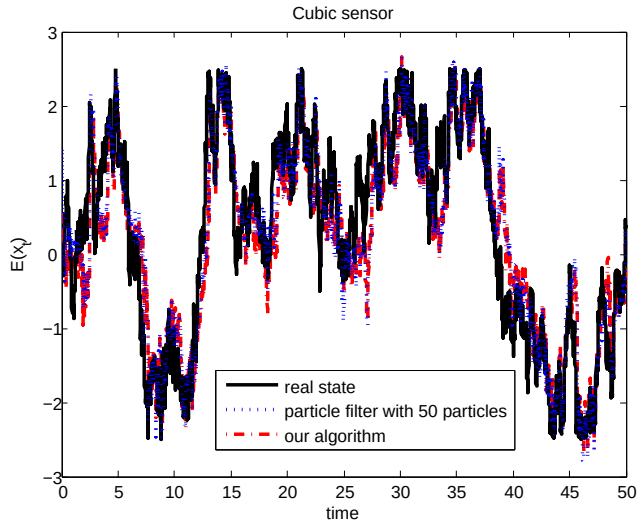

Figure 5. Cubic sensor in the channel is experimented for $T = 50$, with the time step $\triangle t = 0.01$s, by both particle filter and our algorithm.

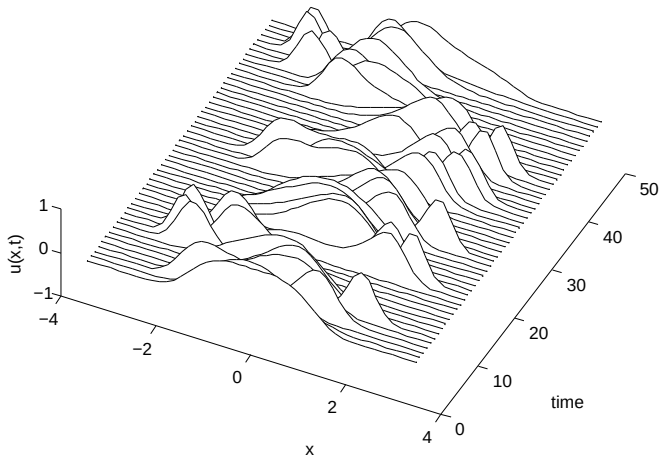

Figure 6. The normalized density functions are plotted every other 1s for the cubic sensor in the channel.

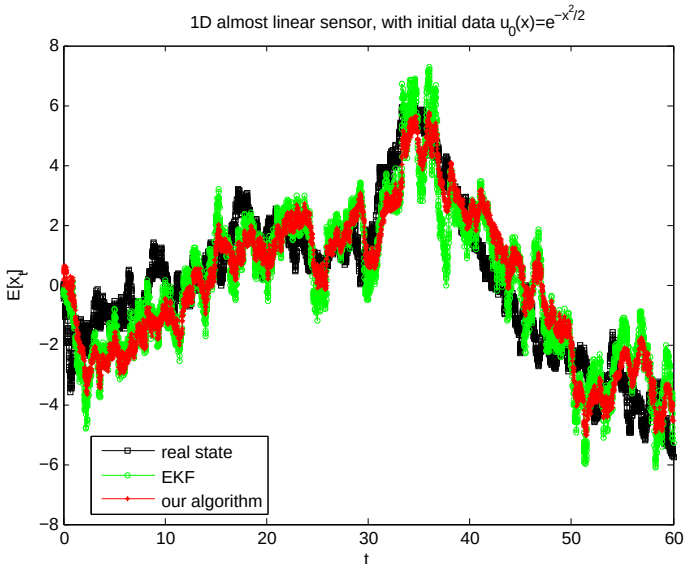

Figure 7. 1D "time-varying" almost linear sensor, with the initial condition $u_0(x) = e^{-\frac{x^2}{2}}$. Black: real state; Green: extended Kalman filter; Red: our algorithm.

CHAPTER 6

OUR ALGORITHM IN HIGHER DIMENSIONS

In the design of our algorithm, it is central to solve the FKE accurately and update rapidly the initial data at the beginning of each interval. The main difficulty on applying our algorithm to high-dimensional NLF problems is the so-called "curse of dimensionality". As mentioned in Chapter 1, we shall resort to the HSM, combined with the sparse grids algorithms.

6.1 Hyperbolic cross (HC) approximation with generalized Hermite functions

6.1.1 Notations

Let us first clarify the notations to be used in this chapter.

- ◇ Let $\mathbb{R}$(resp., $\mathbb{N}$) denote all the real numbers (resp., natural numbers), and let $\mathbb{N}_0 = \mathbb{N} \cup \{0\}$.
- ◇ For any $d \in \mathbb{N}$, we use boldface lowercase letters to denote d-dimensional multi-indices and vectors, e.g., $\boldsymbol{k} = (k_1, k_2, \ldots, k_d) \in \mathbb{N}_0^d$ and $\boldsymbol{\alpha} = (\alpha_1, \alpha_2, \ldots, \alpha_d) \in \mathbb{R}^d$.
- ◇ Let $\mathbf{1} = (1, 1, \ldots, 1) \in \mathbb{N}^d$, and let $\boldsymbol{e}_i = (0, \ldots, 1 \ldots, 0)$ be the i^{th} unit vector in $\mathbb{R}^d$. For any scalar $s \in \mathbb{R}$, we define the componentwise operations:

$$\begin{aligned} \boldsymbol{\alpha} \pm \boldsymbol{k} =& (\alpha_1 \pm k_1, \ldots, \alpha_d \pm k_d), \quad \boldsymbol{\alpha} \pm s := \boldsymbol{\alpha} \pm s\mathbf{1} = (\alpha_1 \pm s, \ldots, \alpha_d \pm s), \\ \frac{1}{\boldsymbol{\alpha}} =& \left(\frac{1}{\alpha_1}, \ldots, \frac{1}{\alpha_d}\right), \quad \boldsymbol{\alpha}^{\boldsymbol{k}} = \alpha_1^{k_1} \cdots \alpha_d^{k_d}, \end{aligned}$$

and

$$\boldsymbol{\alpha} \geq \boldsymbol{k} \Leftrightarrow \alpha_j \geq k_j, \quad \forall\, 1 \leq j \leq d; \quad \boldsymbol{\alpha} \geq s \Leftrightarrow \alpha_j \geq s, \quad \forall\, 1 \leq j \leq d.$$

◇ The frequently used norms are denoted as

$$|\boldsymbol{k}|_1 = \sum_{j=1}^{d} k_j; \quad |\boldsymbol{k}|_\infty = \max_{1 \leq j \leq d} k_j; \quad |\boldsymbol{k}|_{\text{mix}} = \prod_{j=1}^{d} \bar{k}_j,$$

where $\bar{k}_j = \max\{1, k_j\}$.

◇ Given a multivariate function $u(\boldsymbol{x})$, we denote, the $\boldsymbol{k}^{\text{th}}$ mixed partial derivative by

$$\boldsymbol{\partial}_{\boldsymbol{x}}^{\boldsymbol{k}} u = \frac{\partial^{|\boldsymbol{k}|_1} u}{\partial x_1^{k_1} \cdots \partial x_d^{k_d}} = \partial_{x_1}^{k_1} \cdots \partial_{x_d}^{k_d} u.$$

In particular, we denote $\boldsymbol{\partial}_{\boldsymbol{x}}^{s} u = \boldsymbol{\partial}_{\boldsymbol{x}}^{s\mathbf{1}} u = \boldsymbol{\partial}_{\boldsymbol{x}}^{(s,s,\ldots,s)} u$.

◇ Let $L^2(\mathbb{R}^d)$ be the Lebesgue space in $\mathbb{R}^d$, equipped with the norm $||\cdot|| = \left(\int_{\mathbb{R}^d} |\cdot|^2 d\boldsymbol{x}\right)^{\frac{1}{2}}$ and the scalar product $\langle \cdot, \cdot \rangle$.

◇ We follow the convention in the asymptotic analysis, $a \sim b$ means that there exists some constants $C_1, C_2 > 0$ such that $C_1 a \leq b \leq C_2 a$; $a \lesssim b$ means that there exists some constant $C_3 > 0$ such that $a \leq C_3 b$.

◇ We denote C as some generic positive constant, which may vary from line to line.

6.1.2 Generalized Hermite functions and its properties

For the convenience of analysis in this chapter, we define the generalized Hermite functions slightly different from those in Chapter 5. Let us define the univariate generalized Hermite functions as

$$\mathcal{H}_n^{\alpha,\beta}(x) = \left(\frac{\alpha}{2^n n!\sqrt{\pi}}\right)^{\frac{1}{2}} \mathrm{H}_n(\alpha(x-\beta))e^{-\frac{1}{2}\alpha^2(x-\beta)^2}, \tag{6.1}$$

for $n \geq 0$, where $\alpha > 0$ is the scaling factor,$\beta \in \mathbb{R}$ is the translating factor, and $\{\mathrm{H}_n(x)\}_{n\in\mathbb{N}_0}$ is the physical Hermite polynomials as introduction in Chapter 5. It is readily to derive the following properties for $\{\mathcal{H}_n^{\alpha,\beta}(x)\}_{n\in\mathbb{N}_0}$:

- ◇ The $\{\mathcal{H}_n^{\alpha,\beta}\}_{n\in\mathbb{N}_0}$ forms an orthonormal basis of $L^2(\mathbb{R})$, i.e.

$$\int_{\mathbb{R}} \mathcal{H}_n^{\alpha,\beta}(x)\mathcal{H}_m^{\alpha,\beta}(x)dx = \delta_{nm}, \tag{6.2}$$

  where δ_{nm} is the Kronecker function.

- ◇ $\mathcal{H}_n^{\alpha,\beta}(x)$ is the n^{th} eigenfunction of the following Strum-Liouville problem

$$e^{\frac{1}{2}\alpha^2(x-\beta)^2}\partial_x(e^{-\alpha^2(x-\beta)^2}\partial_x(e^{\frac{1}{2}\alpha^2(x-\beta)^2}u(x))) + \lambda_n u(x) = 0, \tag{6.3}$$

  with the corresponding eigenvalue $\lambda_n = 2\alpha^2 n$.

◇ By convention, $\mathcal{H}_n^{\alpha,\beta} \equiv 0$, for $n < 0$. For $n \geq 0$, the three-term recurrence is inherited from the Hermite polynomials:

$$2\alpha^2(x-\beta)\mathcal{H}_n^{\alpha,\beta}(x) = \sqrt{\lambda_n}\mathcal{H}_{n-1}^{\alpha,\beta}(x) + \sqrt{\lambda_{n+1}}\mathcal{H}_{n+1}^{\alpha,\beta}(x). \tag{6.4}$$

◇ The derivative of $\mathcal{H}_n^{\alpha,\beta}(x)$ is explicitly expressed, namely

$$\partial_x\mathcal{H}_n^{\alpha,\beta}(x) = \frac{1}{2}\sqrt{\lambda_n}\mathcal{H}_{n-1}^{\alpha,\beta}(x) - \frac{1}{2}\sqrt{\lambda_{n+1}}\mathcal{H}_{n+1}^{\alpha,\beta}(x). \tag{6.5}$$

◇ Let $\mathcal{D}_x = \partial_x + \alpha^2(x-\beta)$. Then

$$\mathcal{D}_x^k\mathcal{H}_n^{\alpha,\beta}(x) = \sqrt{\mu_{n,k}}\mathcal{H}_{n-k}^{\alpha,\beta}(x), \quad \forall\, n \geq k \geq 1, \tag{6.6}$$

where

$$\mu_{n,k} = \prod_{j=0}^{k-1}\lambda_{n-j} = \frac{2^k\alpha^{2k}n!}{(n-k)!}, \quad \text{for } n \geq k \geq 1. \tag{6.7}$$

◇ The orthogonality of $\{\mathcal{D}_x^k\mathcal{H}_n^{\alpha,\beta}(x)\}_{n\in\mathbb{N}_0}$ holds, i.e.,

$$\int_{\mathbb{R}}\mathcal{D}_x^k\mathcal{H}_n^{\alpha,\beta}(x)\mathcal{D}_x^k\mathcal{H}_m^{\alpha,\beta}(x)dx = \mu_{n,k}\delta_{nm}. \tag{6.8}$$

For notational convenience, we extend $\mu_{n,k}$ for all $n, k \geq 0$.

$$\mu_{n,k} = \begin{cases} 1, & \text{if } k = 0, n \geq 0, \\ 0, & \text{if } k > n \geq 0. \end{cases} \tag{6.9}$$

Now we define the d-dimensional generalized Hermite functions by

$$\boldsymbol{\mathcal{H}}_{\boldsymbol{n}}^{\boldsymbol{\alpha},\boldsymbol{\beta}}(\boldsymbol{x}) = \prod_{j=1}^{d} \mathcal{H}_{n_j}^{\alpha_j,\beta_j}(x_j),$$

for $\boldsymbol{\alpha} > 0$, $\boldsymbol{\beta} \in \mathbb{R}^d$ and $\boldsymbol{x} \in \mathbb{R}^d$. It verifies readily that the properties Equation 6.6 - Equation 6.8 can be extended correspondingly to multivariate generalized Hermite functions. Let $\boldsymbol{\mathcal{D}}_{\boldsymbol{x}}^{\boldsymbol{k}} = \mathcal{D}_{x_1}^{k_1} \cdots \mathcal{D}_{x_d}^{k_d}$, then

$$\boldsymbol{\mathcal{D}}_{\boldsymbol{x}}^{\boldsymbol{k}} \boldsymbol{\mathcal{H}}_{\boldsymbol{n}}^{\boldsymbol{\alpha},\boldsymbol{\beta}} = \sqrt{\boldsymbol{\mu}_{\boldsymbol{n},\boldsymbol{k}}} \boldsymbol{\mathcal{H}}_{\boldsymbol{n}-\boldsymbol{k}}^{\boldsymbol{\alpha},\boldsymbol{\beta}}; \tag{6.10}$$

and

$$\int_{\mathbb{R}^d} \boldsymbol{\mathcal{D}}_{\boldsymbol{x}}^{\boldsymbol{k}} \boldsymbol{\mathcal{H}}_{\boldsymbol{n}}^{\boldsymbol{\alpha},\boldsymbol{\beta}}(\boldsymbol{x}) \boldsymbol{\mathcal{D}}_{\boldsymbol{x}}^{\boldsymbol{k}} \boldsymbol{\mathcal{H}}_{\boldsymbol{m}}^{\boldsymbol{\alpha},\boldsymbol{\beta}}(\boldsymbol{x}) d\boldsymbol{x} = \boldsymbol{\mu}_{\boldsymbol{n},\boldsymbol{k}} \boldsymbol{\delta}_{\boldsymbol{n}\boldsymbol{m}}, \tag{6.11}$$

for $\boldsymbol{\alpha} > 0$, $\boldsymbol{\beta} \in \mathbb{R}^d$, where

$$\boldsymbol{\mu}_{\boldsymbol{n},\boldsymbol{k}} = \prod_{j=1}^{d} \mu_{n_j,k_j} \quad \text{and} \quad \boldsymbol{\delta}_{\boldsymbol{n}\boldsymbol{m}} = \prod_{j=1}^{d} \delta_{n_j m_j}. \tag{6.12}$$

Here, $\mu_{\cdot,\cdot}$ is defined in Equation 6.7 and Equation 6.9, and $\boldsymbol{\delta_{nm}}$ is the tensorial Kronecker function.

The generalized Hermite functions $\{\boldsymbol{\mathcal{H}_n^{\alpha,\beta}}(\boldsymbol{x})\}_{\boldsymbol{n}\in\mathbb{N}_0^d}$ form an orthonormal basis of $L^2(\mathbb{R}^d)$. That is, for any function $u \in L^2(\mathbb{R}^d)$ can be written in the form

$$u(\boldsymbol{x}) = \sum_{\boldsymbol{n}\geq 0} \hat{u}_{\boldsymbol{n}}^{\boldsymbol{\alpha},\boldsymbol{\beta}} \mathcal{H}_{\boldsymbol{n}}^{\boldsymbol{\alpha},\boldsymbol{\beta}}(\boldsymbol{x}), \quad \text{with} \quad \hat{u}_{\boldsymbol{n}}^{\boldsymbol{\alpha},\boldsymbol{\beta}} = \int_{\mathbb{R}^d} u(\boldsymbol{x}) \mathcal{H}_{\boldsymbol{n}}^{\boldsymbol{\alpha},\boldsymbol{\beta}}(\boldsymbol{x}) d\boldsymbol{x}. \tag{6.13}$$

Hence, we have $\boldsymbol{\mathcal{D}_x^k} u(\boldsymbol{x}) = \sum_{\boldsymbol{n}\geq\boldsymbol{k}} \hat{u}_{\boldsymbol{n}}^{\boldsymbol{\alpha},\boldsymbol{\beta}} \boldsymbol{\mathcal{D}_x^k} \mathcal{H}_{\boldsymbol{n}}^{\boldsymbol{\alpha},\boldsymbol{\beta}}(\boldsymbol{x})$. Furthermore,

$$||\boldsymbol{\mathcal{D}_x^k} u||^2 = \sum_{\boldsymbol{n}\geq\boldsymbol{k}} \boldsymbol{\mu_{n,k}} |\hat{u}_{\boldsymbol{n}}^{\boldsymbol{\alpha},\boldsymbol{\beta}}|^2 = \sum_{\boldsymbol{n}\in\mathbb{N}_0^d} \boldsymbol{\mu_{n,k}} |\hat{u}_{\boldsymbol{n}}^{\boldsymbol{\alpha},\boldsymbol{\beta}}|^2, \tag{6.14}$$

by Equation 6.9.

6.1.3 Multivariate orthogonal projection and approximations

In this subsection, we aim to arrive at some typical error esitmates of the form

$$\inf_{U_N\in X_N} ||u - U_N||_l \lesssim N^{-c(l,r)} ||u||_r,$$

where $c(l,r)$ is some positive constant depending on l and r, $||\cdot||_l$ is the norm of some function space, l indicates the regularity of the function in some sense, and X_N is an approximation space. In the sequel, X_N is defined as

$$X_N^{\boldsymbol{\alpha},\boldsymbol{\beta}} = \text{span}\{\boldsymbol{\mathcal{H}_n^{\alpha,\beta}} : \boldsymbol{n} \in \boldsymbol{\Omega}_N\}, \tag{6.15}$$

where $\mathbf{\Omega}_N \subset \mathbb{N}_0^d$ is some index set. With different choices of $\mathbf{\Omega}_N$, we arrive at different approximations, including the full grid, regular hyperbolic cross (RHC) and optimized hyperbolic cross (OHC), etc.

Let us denote the orthogonal projection operator $P_N^{\boldsymbol{\alpha},\boldsymbol{\beta}} : L^2(\mathbb{R}^d) \to X_N^{\boldsymbol{\alpha},\boldsymbol{\beta}}$, i.e., for any $u \in L^2(\mathbb{R}^d)$,

$$\langle (u - P_N^{\boldsymbol{\alpha},\boldsymbol{\beta}} u), v \rangle = 0, \quad \forall\, v \in X_N^{\boldsymbol{\alpha},\boldsymbol{\beta}},$$

or, equivalently,

$$P_N^{\boldsymbol{\alpha},\boldsymbol{\beta}} u(\boldsymbol{x}) = \sum_{\boldsymbol{n} \in \mathbf{\Omega}_N} \hat{u}_N^{\boldsymbol{\alpha},\boldsymbol{\beta}} \boldsymbol{\mathcal{H}}_{\boldsymbol{n}}^{\boldsymbol{\alpha},\boldsymbol{\beta}}(\boldsymbol{x}). \tag{6.16}$$

We shall estimate how close the projected function $P_N^{\boldsymbol{\alpha},\boldsymbol{\beta}} u$ is to u, with respect to various index sets $\mathbf{\Omega}_N$ and norms.

6.1.3.1 Appoximations on the full grid

The index set $\mathbf{\Omega}_N$ corresponding to the d-dimensional full tensor grid is

$$\mathbf{\Omega}_N = \{\boldsymbol{n} \in \mathbb{N}_0^d : \ |\boldsymbol{n}|_\infty \leq N\}.$$

And $X_N^{\boldsymbol{\alpha},\boldsymbol{\beta}}$ is defined in Equation 6.15. Let us define the Sobolev-type space as

$$\mathcal{W}_{\boldsymbol{\alpha},\boldsymbol{\beta}}^m(\mathbb{R}^d) = \{u : \ \boldsymbol{\mathcal{D}}_{\boldsymbol{x}}^{\boldsymbol{k}} u \in L^2(\mathbb{R}^d), 0 \leq |\boldsymbol{k}|_1 \leq m\}, \quad \forall\, m \in \mathbb{N}_0, \tag{6.17}$$

equipped with the norm

$$||u||_{\mathcal{W}^m_{\boldsymbol{\alpha},\boldsymbol{\beta}}(\mathbb{R}^d)} = \left(\sum_{0\leq|\boldsymbol{k}|_1\leq m} \left|\left| \boldsymbol{\mathcal{D}}^{\boldsymbol{k}}_{\boldsymbol{x}} u \right|\right|^2 \right)^{\frac{1}{2}}, \tag{6.18}$$

and seminorm

$$|u|_{\mathcal{W}^m_{\boldsymbol{\alpha},\boldsymbol{\beta}}(\mathbb{R}^d)} = \left(\sum_{j=1}^{d} \left|\left| \mathcal{D}^m_{x_j} u \right|\right|^2 \right)^{\frac{1}{2}}. \tag{6.19}$$

It is clear that $\mathcal{W}^0_{\boldsymbol{\alpha},\boldsymbol{\beta}}(\mathbb{R}^d) = L^2(\mathbb{R}^d)$, and

$$|u|^2_{\mathcal{W}^m_{\boldsymbol{\alpha},\boldsymbol{\beta}}(\mathbb{R}^d)} = \sum_{j=1}^{d} \sum_{\boldsymbol{n}\in\mathbb{N}_0^d} \mu_{n_j,m} \left| \hat{u}^{\boldsymbol{\alpha},\boldsymbol{\beta}}_{\boldsymbol{n}} \right|^2, \tag{6.20}$$

by Equation 6.14.

Theorem 6.1. *Given $u \in \mathcal{W}^m_{\boldsymbol{\alpha},\boldsymbol{\beta}}(\mathbb{R}^d)$, we have for any $0 \leq l < m$,*

$$\left| P_N^{\boldsymbol{\alpha},\boldsymbol{\beta}} u - u \right|_{\mathcal{W}^l_{\boldsymbol{\alpha},\boldsymbol{\beta}}(\mathbb{R}^d)} \lesssim |\boldsymbol{\alpha}|_\infty^{l-m} N^{\frac{l-m}{2}} |u|_{\mathcal{W}^m_{\boldsymbol{\alpha},\boldsymbol{\beta}}(\mathbb{R}^d)}, \tag{6.21}$$

for $N \gg 1$. Furthermore,

$$\left|\left| P_N^{\boldsymbol{\alpha},\boldsymbol{\beta}} u - u \right|\right|_{\mathcal{W}^l_{\boldsymbol{\alpha},\boldsymbol{\beta}}(\mathbb{R}^d)} \lesssim C_{\boldsymbol{\alpha},l,m} N^{\frac{l-m}{2}} |u|_{\mathcal{W}^m_{\boldsymbol{\alpha},\boldsymbol{\beta}}(\mathbb{R}^d)},$$

where $C_{\boldsymbol{\alpha},l,m}$ is some constant depending on $\boldsymbol{\alpha}$, l and m.

Proof. The argument is similar to that in (51). Let $\mathbf{\Omega}_N^c = \{\boldsymbol{n} \in \mathbb{N}_0^d : \ |\boldsymbol{n}|_\infty > N\}$. By Equation 6.16, Equation 6.19 and Equation 6.20,

$$\left| P_N^{\boldsymbol{\alpha},\boldsymbol{\beta}} u - u \right|^2_{\mathcal{W}^l_{\boldsymbol{\alpha},\boldsymbol{\beta}}(\mathbb{R}^d)} = \sum_{j=1}^{d} \sum_{\boldsymbol{n}\in\mathbf{\Omega}_N^c} \mu_{n_j,l} \left| \hat{u}_{\boldsymbol{n}}^{\boldsymbol{\alpha},\boldsymbol{\beta}} \right|^2. \tag{6.22}$$

For any $1 \le j \le d$,

$$\sum_{\boldsymbol{n}\in\mathbf{\Omega}_N^c} \mu_{n_j,l} \left| \hat{u}_{\boldsymbol{n}}^{\boldsymbol{\alpha},\boldsymbol{\beta}} \right|^2 = \sum_{\boldsymbol{n}\in\mathbf{\Lambda}_N^{1,j}} \mu_{n_j,l} \left| \hat{u}_{\boldsymbol{n}}^{\boldsymbol{\alpha},\boldsymbol{\beta}} \right|^2 + \sum_{\boldsymbol{n}\in\mathbf{\Lambda}_N^{2,j}} \mu_{n_j,l} \left| \hat{u}_{\boldsymbol{n}}^{\boldsymbol{\alpha},\boldsymbol{\beta}} \right|^2 := \mathrm{XII} + \mathrm{XIII}, \tag{6.23}$$

where $\mathbf{\Lambda}_N^{1,j} = \{\boldsymbol{n} \in \mathbf{\Omega}_N^c : \ n_j > N\}$ and $\mathbf{\Lambda}_N^{2,j} = \{\boldsymbol{n} \in \mathbf{\Omega}_N^c : \ n_j \le N\}$. For XII:

$$\mathrm{XII} \le \max_{\boldsymbol{n}\in\mathbf{\Lambda}_N^{1,j}} \left\{ \frac{\mu_{n_j,l}}{\mu_{n_j,m}} \right\} \sum_{\boldsymbol{n}\in\mathbf{\Lambda}_N^{1,j}} \mu_{n_j,m} \left| \hat{u}_{\boldsymbol{n}}^{\boldsymbol{\alpha},\boldsymbol{\beta}} \right|^2 \lesssim |\boldsymbol{\alpha}|_\infty^{2(l-m)} N^{l-m} |u|^2_{\mathcal{W}^m_{\boldsymbol{\alpha},\boldsymbol{\beta}}(\mathbb{R}^d)}. \tag{6.24}$$

In fact,

$$\begin{aligned} \max_{\boldsymbol{n}\in\mathbf{\Lambda}_N^{1,j}} \left\{ \frac{\mu_{n_j,l}}{\mu_{n_j,m}} \right\} &= \max_{\boldsymbol{n}\in\mathbf{\Lambda}_N^{1,j}} \left\{ \frac{2^{l-m} \alpha_j^{2(l-m)}}{(n_j-l)(n_j-l-1)\cdots(n_j-m+1)} \right\} \\ &\le 2^{l-m} |\boldsymbol{\alpha}|_\infty^{2(l-m)} (N-m+1)^{l-m}. \end{aligned}$$

For XIII, if $\boldsymbol{n} \in \mathbf{\Lambda}_N^{2,j}$, there exists some $k \ne j$, such that $n_k > N$.

$$\mathrm{XIII} \le \max_{\boldsymbol{n}\in\mathbf{\Lambda}_N^{2,j}} \left\{ \frac{\mu_{n_j,l}}{\mu_{n_k,m}} \right\} \sum_{\boldsymbol{n}\in\mathbf{\Lambda}_N^{2,j}} \mu_{n_k,m} \left| \hat{u}_{\boldsymbol{n}}^{\boldsymbol{\alpha},\boldsymbol{\beta}} \right|^2 \lesssim |\boldsymbol{\alpha}|_\infty^{2l} \left| \frac{1}{\boldsymbol{\alpha}} \right|_\infty^{2m} N^{l-m-2} |u|^2_{\mathcal{W}^m_{\boldsymbol{\alpha},\boldsymbol{\beta}}(\mathbb{R}^d)}, \tag{6.25}$$

since

$$\begin{aligned}\max_{\boldsymbol{n}\in\boldsymbol{\Lambda}_N^{2,j}}\left\{\frac{\mu_{n_j,l}}{\mu_{n_k,m}}\right\} &= \max_{\boldsymbol{n}\in\boldsymbol{\Lambda}_N^{2,j}}\left\{2^{l-m}\frac{\alpha_j^{2l}}{\alpha_k^{2m}}\frac{\frac{n_j!}{(n_j-l)!}}{\frac{n_k!}{(n_k-m)!}}\right\} \leq 2^{l-m}|\boldsymbol{\alpha}|_\infty^{2l}\left|\frac{1}{\boldsymbol{\alpha}}\right|_\infty^{2m}\frac{\frac{N!}{(N-l)!}}{\frac{(N+1)!}{(N+1-m)!}}\\ &=2^{l-m}|\boldsymbol{\alpha}|_\infty^{2l}\left|\frac{1}{\boldsymbol{\alpha}}\right|_\infty^{2m}\frac{1}{N+1}\frac{1}{(N-l)(N-l-1)\cdots(N-m)}\\ &\leq 2^{l-m}|\boldsymbol{\alpha}|_\infty^{2l}\left|\frac{1}{\boldsymbol{\alpha}}\right|_\infty^{2m}(N-m)^{l-m-2}.\end{aligned}$$

Combining Equation 6.22 - Equation 6.25, we obtain the result. Furthermore, the mixed derivatives of order equal to or less than m can be bounded by the seminorm $|u|_{\mathcal{W}^m_{\boldsymbol{\alpha},\boldsymbol{\beta}}(\mathbb{R}^d)}$. □

Remark 6.1. *It is clear that the convergence rate deteriorates rapidly with respect to the cardinality of the full grid. That is,*

$$\left|\left|P_N^{\boldsymbol{\alpha},\boldsymbol{\beta}}u-u\right|\right|_{\mathcal{W}^l_{\boldsymbol{\alpha},\boldsymbol{\beta}}(\mathbb{R}^d)} \lesssim C_{\boldsymbol{\alpha},l,m}M^{\frac{l-m}{2d}}|u|_{\mathcal{W}^m_{\boldsymbol{\alpha},\boldsymbol{\beta}}(\mathbb{R}^d)},$$

where $M=\mathrm{card}(\boldsymbol{\Omega}_N)=N^d$.

6.1.3.2 Regular hyperbolic cross (RHC) approximation

As we mentioned in Chapter 1, the HC approximation is an efficient tool to overcome the "curse of dimensionality" in some degree. The index set of the RHC approximation is $\boldsymbol{\Omega}_N=\{\boldsymbol{n}\in\mathbb{N}_0^d:\ |\boldsymbol{n}|_{\mathrm{mix}}\leq N\}$. It is known that the cardinality of $\boldsymbol{\Omega}_N$ is $\mathcal{O}(N(\ln N)^{d-1})$, see (24). Correspondingly, the finite dimensional subspace $X_N^{\boldsymbol{\alpha},\boldsymbol{\beta}}$ is

$$X_N^{\boldsymbol{\alpha},\boldsymbol{\beta}}=\mathrm{span}\{\boldsymbol{\mathcal{H}}_{\boldsymbol{n}}^{\boldsymbol{\alpha},\boldsymbol{\beta}}:\ |\boldsymbol{n}|_{\mathrm{mix}}\leq N\}. \tag{6.26}$$

Let the orthogonal projection operator $P_N^{\boldsymbol{\alpha},\boldsymbol{\beta}} : L^2(\mathbb{R}^d) \to X_N^{\boldsymbol{\alpha},\boldsymbol{\beta}}$ be defined as before. Denote the $\boldsymbol{k}-$complement of $\boldsymbol{\Omega}_N$ by

$$\boldsymbol{\Omega}_{N,\boldsymbol{k}}^c := \{\boldsymbol{n} \in \mathbb{N}_0^d : \ |\boldsymbol{n}|_{\text{mix}} > N \text{ and } \boldsymbol{n} \geq \boldsymbol{k}\}, \quad \forall\, \boldsymbol{k} \in \mathbb{N}_0^d. \tag{6.27}$$

We define the Koborov-type space as

$$\mathcal{K}_{\boldsymbol{\alpha},\boldsymbol{\beta}}^r(\mathbb{R}^d) = \{u : \ \mathcal{D}_{\boldsymbol{x}}^{\boldsymbol{k}} u \in L^2(\mathbb{R}^d),\ 0 \leq |\boldsymbol{k}|_\infty \leq r\}, \quad \forall\, m \in \mathbb{N}_0^d, \tag{6.28}$$

equipped with the norm

$$||u||_{\mathcal{K}_{\boldsymbol{\alpha},\boldsymbol{\beta}}^r(\mathbb{R}^d)} = \left(\sum_{0 \leq |\boldsymbol{k}|_\infty \leq r} \left|\left| \mathcal{D}_{\boldsymbol{x}}^{\boldsymbol{k}} u \right|\right|^2 \right)^{\frac{1}{2}}, \tag{6.29}$$

and seminorm

$$|u|_{\mathcal{K}_{\boldsymbol{\alpha},\boldsymbol{\beta}}^r(\mathbb{R}^d)} = \left(\sum_{|\boldsymbol{k}|_\infty = r} \left|\left| \mathcal{D}_{\boldsymbol{x}}^{\boldsymbol{k}} u \right|\right|^2 \right)^{\frac{1}{2}}. \tag{6.30}$$

Remark 6.2. *It is easy to see from the definitions that* $\mathcal{K}_{\boldsymbol{\alpha},\boldsymbol{\beta}}^0(\mathbb{R}^d) = L^2(\mathbb{R}^d)$ *and* $\mathcal{W}_{\boldsymbol{\alpha},\boldsymbol{\beta}}^{dl}(\mathbb{R}^d) \subset \mathcal{K}_{\boldsymbol{\alpha},\boldsymbol{\beta}}^l(\mathbb{R}^d) \subset \mathcal{W}_{\boldsymbol{\alpha},\boldsymbol{\beta}}^l(\mathbb{R}^d)$.

Theorem 6.2. *Given* $u \in \mathcal{K}_{\boldsymbol{\alpha},\boldsymbol{\beta}}^m(\mathbb{R}^d)$, *for* $0 \leq \boldsymbol{l} < m$, *we have*

$$\left|\left| \mathcal{D}_{\boldsymbol{x}}^{\boldsymbol{l}} \left(P_N^{\boldsymbol{\alpha},\boldsymbol{\beta}} u - u \right) \right|\right| \leq C_{\boldsymbol{\alpha},\boldsymbol{l},m,d} N^{\frac{|\boldsymbol{l}|_\infty - m}{2}} |u|_{\mathcal{K}_{\boldsymbol{\alpha},\boldsymbol{\beta}}^m(\mathbb{R}^d)},$$

where $C_{\boldsymbol{\alpha},\boldsymbol{l},m,d}$ *is some constant depending on* $\boldsymbol{\alpha}$, $\boldsymbol{l}$, m *and* d, *for* $N \gg 1$. *In particular, if* $\boldsymbol{\alpha} = \mathbf{1}$, *then*

$$C_{\mathbf{1},\boldsymbol{l},m,d} = 2^{|\boldsymbol{l}|_\infty - m} m^{(2d-1)m - |\boldsymbol{l}|_1 - (d-1)|\boldsymbol{l}|_\infty}.$$

Proof. From Equation 6.16 and Equation 6.14, we have

$$\begin{aligned}\left|\left|\mathcal{D}_{\boldsymbol{x}}^{\boldsymbol{l}}(P_N^{\boldsymbol{\alpha},\boldsymbol{\beta}}u - u)\right|\right|^2 &= \sum_{\boldsymbol{n}\in\boldsymbol{\Omega}_N^c} \boldsymbol{\mu}_{\boldsymbol{n},\boldsymbol{l}} \left|\hat{u}_{\boldsymbol{n}}^{\boldsymbol{\alpha},\boldsymbol{\beta}}\right|^2 = \sum_{\boldsymbol{n}\in\boldsymbol{\Omega}_{N,m}^c} \boldsymbol{\mu}_{\boldsymbol{n},\boldsymbol{l}} \left|\hat{u}_{\boldsymbol{n}}^{\boldsymbol{\alpha},\boldsymbol{\beta}}\right|^2 + \sum_{\boldsymbol{n}\in\boldsymbol{\Omega}_{N,\boldsymbol{l}}^c\setminus\boldsymbol{\Omega}_{N,m}^c} \boldsymbol{\mu}_{\boldsymbol{n},\boldsymbol{l}} \left|\hat{u}_{\boldsymbol{n}}^{\boldsymbol{\alpha},\boldsymbol{\beta}}\right|^2 \\ &:= \mathrm{XIV} + \mathrm{XV}.\end{aligned}$$

For XIV:

$$\mathrm{XIV} \le \max_{\boldsymbol{n}\in\boldsymbol{\Omega}_{N,m}^c} \left\{\frac{\boldsymbol{\mu}_{\boldsymbol{n},\boldsymbol{l}}}{\boldsymbol{\mu}_{\boldsymbol{n},m}}\right\} \sum_{\boldsymbol{n}\in\boldsymbol{\Omega}_{N,m}^c} \boldsymbol{\mu}_{\boldsymbol{n},m} \left|\hat{u}_{\boldsymbol{n}}^{\boldsymbol{\alpha},\boldsymbol{\beta}}\right|^2.$$

Using the facts that

$$\begin{aligned}\frac{\boldsymbol{\mu}_{\boldsymbol{n},\boldsymbol{l}}}{\boldsymbol{\mu}_{\boldsymbol{n},m}} =& 2^{|\boldsymbol{l}|_1 - dm} \prod_{j=1}^d \alpha_j^{2(l_j-m)} \prod_{j=1}^d \frac{1}{(n_j - l_j)\cdots(n_j - m + 1)} \\ =& 2^{|\boldsymbol{l}|_1 - dm} \prod_{j=1}^d \alpha_j^{2(l_j-m)} \prod_{j=1}^d n_j^{l_j - m} \prod_{j=1}^d \left(1 - \frac{l_j}{n_j}\right)^{-1} \cdots \left(1 - \frac{m-1}{n_j}\right)^{-1} \\ \le& 2^{|\boldsymbol{l}|_1 - dm} \prod_{j=1}^d \alpha_j^{2(l_j-m)} N^{|\boldsymbol{l}|_\infty - m} \prod_{j=1}^d \left(1 - \frac{l_j}{n_j}\right)^{-1} \cdots \left(1 - \frac{m-1}{n_j}\right)^{-1}, \end{aligned} \tag{6.31}$$

by Equation 6.27, and

$$\max_{\boldsymbol{n}\in\boldsymbol{\Omega}^c_{N,m}}\left\{\prod_{j=1}^{d}\left(1-\frac{l_j}{n_j}\right)^{-1}\cdots\left(1-\frac{m-1}{n_j}\right)^{-1}\right\}\leq\max_{\boldsymbol{n}\in\boldsymbol{\Omega}^c_{N,m}}\left\{\prod_{j=1}^{d}\left(1-\frac{m-1}{n_j}\right)^{l_j-m}\right\}$$
$$\leq\prod_{j=1}^{d}m^{m-l_j}=m^{dm-|\boldsymbol{l}|_1}, \tag{6.32}$$

we find that

$$\text{XIV}\leq\left(\frac{m}{2}\right)^{dm-|\boldsymbol{l}|_1}\prod_{j=1}^{d}\alpha_j^{2(l_j-m)}N^{|\boldsymbol{l}|_\infty-m}\left|\left|\mathcal{D}_{\boldsymbol{x}}^{m\cdot\mathbf{1}}u\right|\right|^2. \tag{6.33}$$

For XV: The index set $\boldsymbol{\Omega}^c_{N,\boldsymbol{l}}\setminus\boldsymbol{\Omega}^c_{N,m}$ is

$$\boldsymbol{\Omega}^c_{N,\boldsymbol{l}}\setminus\boldsymbol{\Omega}^c_{N,m}=\{\boldsymbol{n}\in\mathbb{N}_0^d:\ |\boldsymbol{n}|_{\text{mix}}>N\text{ and }\boldsymbol{n}\geq\boldsymbol{l},\ \exists\, j,\ \text{such that } n_j<m\}.$$

Let us divide the index $1\leq j\leq d$ into two parts

$$\mathcal{N}:=\{j:\ l_j\leq n_j<m,\,1\leq j\leq d\},\quad\mathcal{N}^c:=\{j:\ n_j\geq m,\,1\leq j\leq d\}. \tag{6.34}$$

It is easy to see that neither $\mathcal{N}$ nor $\mathcal{N}^c$ is the empty set. We denote

$$\tilde{\boldsymbol{\mu}}_{\boldsymbol{n},\boldsymbol{l},m}=\left(\prod_{j\in\mathcal{N}}\mu_{n_j,l_j}\right)\left(\prod_{i\in\mathcal{N}^c}\mu_{n_i,m}\right):=\boldsymbol{\mu}_{\boldsymbol{n},\boldsymbol{k}}, \tag{6.35}$$

where $\boldsymbol{k}$ is a d-dimensional index consisting of l_j for $j \in \mathcal{N}$ and m for $j \in \mathcal{N}^c$. Now, we estimate XV as

$$\mathrm{XV} \leq \max_{\boldsymbol{n} \in \boldsymbol{\Omega}^c_{N,\boldsymbol{l}} \setminus \boldsymbol{\Omega}^c_{N,m}} \left\{ \frac{\boldsymbol{\mu}_{\boldsymbol{n},\boldsymbol{l}}}{\boldsymbol{\mu}_{\boldsymbol{n},\boldsymbol{k}}} \right\} \sum_{\boldsymbol{n} \in \boldsymbol{\Omega}^c_{N,\boldsymbol{l}} \setminus \boldsymbol{\Omega}^c_{N,m}} \boldsymbol{\mu}_{\boldsymbol{n},\boldsymbol{k}} \left| \hat{u}^{\boldsymbol{\alpha},\boldsymbol{\beta}}_{\boldsymbol{n}} \right|^2 \leq \max_{\boldsymbol{n} \in \boldsymbol{\Omega}^c_{N,\boldsymbol{l}} \setminus \boldsymbol{\Omega}^c_{N,m}} \left\{ \frac{\boldsymbol{\mu}_{\boldsymbol{n},\boldsymbol{l}}}{\boldsymbol{\mu}_{\boldsymbol{n},\boldsymbol{k}}} \right\} |u|^2_{\mathcal{K}^m_{\boldsymbol{\alpha},\boldsymbol{\beta}}(\mathbb{R}^d)}, \tag{6.36}$$

since $|\boldsymbol{k}|_\infty = m$, where the first inequality follows from Equation 6.35. It remains to estimate the maximum in Equation 6.36:

$$\begin{aligned} \frac{\boldsymbol{\mu}_{\boldsymbol{n},\boldsymbol{l}}}{\boldsymbol{\mu}_{\boldsymbol{n},\boldsymbol{k}}} =& 2^{|\boldsymbol{l}|_1 - |\boldsymbol{k}|_1} \prod_{j \in \mathcal{N}^c} \alpha_j^{2(l_j - m)} \frac{1}{(n_j - l_j) \cdots (n_j - m + 1)} \\ =& 2^{|\boldsymbol{l}|_1 - |\boldsymbol{k}|_1} \prod_{j \in \mathcal{N}^c} \alpha_j^{2(l_j - m)} \prod_{j \in \mathcal{N}^c} n_j^{l_j - m} \prod_{j \in \mathcal{N}^c} \left(1 - \frac{l_j}{n_j}\right)^{-1} \cdots \left(1 - \frac{m-1}{n_j}\right)^{-1}. \end{aligned} \tag{6.37}$$

Observe that $j \in \mathcal{N}^c$ implies that $n_j \geq m > \boldsymbol{l} \geq 0$. That is, $n_j \geq 1$. Hence, $\bar{n}_j = n_j$, for all $j = 1, \cdots, d$. In view of $|\boldsymbol{n}|_{\mathrm{mix}} > N$, we deduce that

$$\prod_{j \in \mathcal{N}^c} \bar{n}_j > \frac{N}{\prod_{j \in \mathcal{N}} \bar{n}_j} > \frac{N}{\prod_{j \in \mathcal{N}} m}.$$

With the same estimate as in Equation 6.32 and the fact that

$$2^{|\boldsymbol{l}|_1 - |\boldsymbol{k}|_1} = 2^{\sum_{j \in \mathcal{N}^c} (l_j - m)} \leq 2^{|\boldsymbol{l}|_\infty - m}, \tag{6.38}$$

we find that

$$\max_{\boldsymbol{n}\in\boldsymbol{\Omega}^c_{N,\boldsymbol{l}}\setminus\boldsymbol{\Omega}^c_{N,m}}\left\{\frac{\boldsymbol{\mu}_{\boldsymbol{n},\boldsymbol{l}}}{\boldsymbol{\mu}_{\boldsymbol{n},\boldsymbol{k}}}\right\}\leq C_{\boldsymbol{\alpha},\boldsymbol{l},m}2^{|\boldsymbol{l}|_\infty-m}m^{(2d-1)m-|\boldsymbol{l}|_1-(d-1)|\boldsymbol{l}|_\infty}N^{|\boldsymbol{l}|_\infty-m}, \tag{6.39}$$

where $C_{\boldsymbol{\alpha},\boldsymbol{l},m}$ denotes some constant depending on $\boldsymbol{\alpha}$, $\boldsymbol{l}$ and m. The desired result follows immediately from Equation 6.33, Equation 6.36 and Equation 6.39. □

Corollary 6.1.

$$\left|\left|P_N^{\boldsymbol{\alpha},\boldsymbol{\beta}}u-u\right|\right|_{\mathcal{K}^l_{\boldsymbol{\alpha},\boldsymbol{\beta}}(\mathbb{R}^d)}\leq C_{\boldsymbol{\alpha},l,m,d}N^{\frac{l-m}{2}}|u|_{\mathcal{K}^m_{\boldsymbol{\alpha},\boldsymbol{\beta}}(\mathbb{R}^d)},\quad\forall\,0\leq l<m,$$

where $C_{\boldsymbol{\alpha},l,m,d}$ is some constant depending on $\boldsymbol{\alpha}$, l, m and d.

Remark 6.3. *Recall that $M=\mathrm{card}(\boldsymbol{\Omega}_N)=\mathcal{O}(N(\ln N)^{d-1})\leq CN^{1+\epsilon(d-1)}$, for arbitrarily small $\epsilon>0$. Then*

$$\left|\left|P_N^{\boldsymbol{\alpha},\boldsymbol{\beta}}u-u\right|\right|_{\mathcal{K}^l_{\boldsymbol{\alpha},\boldsymbol{\beta}}(\mathbb{R}^d)}\leq C_{\boldsymbol{\alpha},l,m,d}M^{\frac{l-m}{2(1+\epsilon(d-1))}}|u|_{\mathcal{K}^m_{\boldsymbol{\alpha},\boldsymbol{\beta}}(\mathbb{R}^d)},\quad\forall\,0\leq l<m,$$

where $C_{\boldsymbol{\alpha},l,m,d}$ is some constant depending on $\boldsymbol{\alpha}$, l, m and d. It is clear to see that the convergence rate deteriorates slightly with increasing d.

6.1.3.3 Optimized hyperbolic cross (OHC) approximation

In order to completely break the curse of dimensionality, we consider the index set introduced in (24)

$$\boldsymbol{\Omega}_{N,\gamma} := \{\boldsymbol{n} \in \mathbb{N}_0^d : \ |\boldsymbol{n}|_{\text{mix}}|\boldsymbol{n}|_\infty^{-\gamma} \leq N^{1-\gamma}\}, \quad -\infty \leq \gamma < 1. \tag{6.40}$$

The cardinality of $\boldsymbol{\Omega}_{N,\gamma}$ is $\mathcal{O}(N)$, for $\gamma \in (0,1)$, where the dependence of dimension is in the big-$\mathcal{O}$, see (24). The family of spaces are defined as

$$X_{N,\gamma}^{\boldsymbol{\alpha},\boldsymbol{\beta}} := \text{span}\{\boldsymbol{\mathcal{H}}_{\boldsymbol{n}}^{\boldsymbol{\alpha},\boldsymbol{\beta}} : \ \boldsymbol{n} \in \boldsymbol{\Omega}_{N,\gamma}\}. \tag{6.41}$$

Remark 6.4. *In particular, we have* $X_{N,0}^{\boldsymbol{\alpha},\boldsymbol{\beta}} = X_N^{\boldsymbol{\alpha},\boldsymbol{\beta}}$ *in RHC, see Equation 6.26, and* $X_{N,-\infty}^{\boldsymbol{\alpha},\boldsymbol{\beta}} = \text{span}\{\mathcal{H}_{\boldsymbol{n}}^{\boldsymbol{\alpha},\boldsymbol{\beta}} : \ |\boldsymbol{n}|_\infty \leq N\}$*, i.e., the full grid.*

We denote the projection operator as $P_{N,\gamma}^{\boldsymbol{\alpha},\boldsymbol{\beta}} : \ L^2(\mathbb{R}^d) \to X_{N,\gamma}^{\boldsymbol{\alpha},\boldsymbol{\beta}}$. In this case, the $\boldsymbol{k}-$complement of index set of $\boldsymbol{\Omega}_{N,\gamma}$ is

$$\boldsymbol{\Omega}_{N,\gamma,\boldsymbol{k}}^c = \{\boldsymbol{n} \in \mathbb{N}_0^d : \ \boldsymbol{n} \in \boldsymbol{\Omega}_{N,\gamma}^c \text{ and } \boldsymbol{n} \geq \boldsymbol{k}\}, \quad \forall\, \boldsymbol{k} \in \mathbb{N}_0^d. \tag{6.42}$$

Although (51) obtains a similar result for Jacobi polynomials as Theorem 6.3 below, we believe that there is a gap in their error analysis of OHC, namely in Theorem 2.3 in (51). We circumvent it here with a more delicate analysis.

Theorem 6.3. *For any $u \in \mathcal{K}^m_{\boldsymbol{\alpha},\boldsymbol{\beta}}(\mathbb{R}^d)$, $d \geq 2$, and $0 \leq |\boldsymbol{l}|_1 < m$,*

$$\left|\left|\boldsymbol{\mathcal{D}}^{\boldsymbol{l}}_{\boldsymbol{x}}\left(P^{\boldsymbol{\alpha},\boldsymbol{\beta}}_{N,\gamma}u - u\right)\right|\right| \leq C_{\boldsymbol{\alpha},\boldsymbol{l},m,d,\gamma}|u|_{\mathcal{K}^m_{\boldsymbol{\alpha},\boldsymbol{\beta}}(\mathbb{R}^d)} \begin{cases} N^{\frac{|\boldsymbol{l}|_1-m}{2}}, & \text{if } 0 < \gamma \leq \dfrac{|\boldsymbol{l}|_1}{m} \\ N^{\frac{(1-\gamma)[|\boldsymbol{l}|_1-(d-1)m]}{d-1-\gamma}}, & \text{if } \dfrac{|\boldsymbol{l}|_1}{m} \leq \gamma < 1, \end{cases} \tag{6.43}$$

where $C_{\boldsymbol{\alpha},\boldsymbol{l},m,d,\gamma}$ is some constant depending on $\boldsymbol{\alpha}$, $\boldsymbol{l}$, m, d and γ. In particular, if $\boldsymbol{\alpha} = \mathbf{1}$, then

$$C_{\mathbf{1},\boldsymbol{l},m,d,\gamma} = m^{dm-|\boldsymbol{l}|_1} \begin{cases} 2^{|\boldsymbol{l}|_\infty - m} m^{\frac{(d-1)(\gamma m - |\boldsymbol{l}|_1)}{1-\gamma}}, & \text{if } 0 < \gamma \leq \dfrac{|\boldsymbol{l}|_1}{m} \\ 2^{|\boldsymbol{l}|_1 - dm}, & \text{if } \dfrac{|\boldsymbol{l}|_1}{m} \leq \gamma < 1. \end{cases}$$

Proof. As argued in the proof of Theorem 6.2, we arrive at

$$\begin{aligned} \left|\left|\boldsymbol{\mathcal{D}}^{\boldsymbol{l}}_{\boldsymbol{x}}\left(P^{\boldsymbol{\alpha},\boldsymbol{\beta}}_{N,\gamma}u - u\right)\right|\right|^2 \leq & \max_{\boldsymbol{n}\in\boldsymbol{\Omega}^c_{N,\gamma,m}} \left\{\frac{\boldsymbol{\mu}_{\boldsymbol{n},\boldsymbol{l}}}{\boldsymbol{\mu}_{\boldsymbol{n},m}}\right\} \sum_{\boldsymbol{n}\in\boldsymbol{\Omega}^c_{N,\gamma,m}} \boldsymbol{\mu}_{\boldsymbol{n},m}\left|\hat{u}^{\boldsymbol{\alpha},\boldsymbol{\beta}}_{\boldsymbol{n}}\right|^2 \\ & + \max_{\boldsymbol{n}\in\boldsymbol{\Omega}^c_{N,\gamma,\boldsymbol{l}}\setminus\boldsymbol{\Omega}^c_{N,\gamma,m}} \left\{\frac{\boldsymbol{\mu}_{\boldsymbol{n},\boldsymbol{l}}}{\tilde{\boldsymbol{\mu}}_{\boldsymbol{n},\boldsymbol{l},m}}\right\} \sum_{\boldsymbol{n}\in\boldsymbol{\Omega}^c_{N,\gamma,\boldsymbol{l}}\setminus\boldsymbol{\Omega}^c_{N,\gamma,m}} \tilde{\boldsymbol{\mu}}_{\boldsymbol{n},\boldsymbol{l},m}\left|\hat{u}^{\boldsymbol{\alpha},\boldsymbol{\beta}}_{\boldsymbol{n}}\right|^2 \\ := & \text{XVI} + \text{XVII}, \end{aligned} \tag{6.44}$$

where $\tilde{\boldsymbol{\mu}}_{\boldsymbol{n},\boldsymbol{l},m}$ is defined as in Equation 6.35. To estimate XVI, like in Equation 6.31, we have

$$\begin{aligned} \frac{\boldsymbol{\mu}_{\boldsymbol{n},\boldsymbol{l}}}{\boldsymbol{\mu}_{\boldsymbol{n},m}} = & 2^{|\boldsymbol{l}|_1-dm}\prod_{j=1}^d \alpha_j^{2(l_j-m)} \prod_{j=1}^d \left(1-\frac{l_j}{n_j}\right)^{-1} \cdots \left(1-\frac{m-1}{n_j}\right)^{-1} \prod_{j=1}^d n_j^{l_j-m} \\ := & D_1 \prod_{j=1}^d n_j^{l_j-m}. \end{aligned} \tag{6.45}$$

The estimate of $\max_{\boldsymbol{n}\in\boldsymbol{\Omega}^c_{N,\gamma,m}} D_1$ follows by a similar argument as in Equation 6.32, i.e.,

$$\max_{\boldsymbol{n}\in\boldsymbol{\Omega}^c_{N,\gamma,m}} D_1 \leq \left(\frac{m}{2}\right)^{dm-|\boldsymbol{l}|_1} \prod_{j=1}^{d} \alpha_j^{2(l_j-m)}. \tag{6.46}$$

Notice that for any $\boldsymbol{n} \in \boldsymbol{\Omega}^c_{N,\gamma}$,

$$|\boldsymbol{n}|_{\text{mix}}|\boldsymbol{n}|_\infty^{-\gamma} > N^{1-\gamma} \Rightarrow \left(\frac{|\boldsymbol{n}|_\infty^\gamma}{|\boldsymbol{n}|_{\text{mix}}}\right)^{\frac{1}{1-\gamma}} < \frac{1}{N} \tag{6.47}$$

and furthermore, if $\boldsymbol{n} \in \boldsymbol{\Omega}^c_{N,\gamma,m}$,

$$\frac{|\boldsymbol{n}|_\infty}{|\boldsymbol{n}|_{\text{mix}}} \leq \frac{1}{m^{d-1}}. \tag{6.48}$$

Moreover,

$$|\boldsymbol{n}|_\infty^{d-\gamma} \geq |\boldsymbol{n}|_{\text{mix}}|\boldsymbol{n}|_\infty^{-\gamma} > N^{1-\gamma} \Rightarrow |\boldsymbol{n}|_\infty > N^{\frac{1-\gamma}{d-\gamma}}. \tag{6.49}$$

Let us estimate the product in the right-hand side of Equation 6.45:

$$\prod_{j=1}^{d} n_j^{l_j-m} = \left(\prod_{j=1}^{d} n_j^{l_j}\right)\left(\prod_{j=1}^{d} n_j\right)^{-m} \leq \left(\prod_{j=1}^{d} |\boldsymbol{n}|_\infty^{l_j}\right)|\boldsymbol{n}|_{\text{mix}}^{-m} = |\boldsymbol{n}|_\infty^{|\boldsymbol{l}|_1}|\boldsymbol{n}|_{\text{mix}}^{-m}. \tag{6.50}$$

If $0 < \gamma \leq \frac{|\boldsymbol{l}|_1}{m}$, then

$$\begin{aligned}\max_{\boldsymbol{n}\in\boldsymbol{\Omega}^c_{N,\gamma,m}} \prod_{j=1}^{d} n_j^{l_j-m} &\leq \max_{\boldsymbol{n}\in\boldsymbol{\Omega}^c_{N,\gamma,m}} \left\{ \left(\frac{|\boldsymbol{n}|_\infty^\gamma}{|\boldsymbol{n}|_{\text{mix}}}\right)^{\frac{m-|\boldsymbol{l}|_1}{1-\gamma}} \left(\frac{|\boldsymbol{n}|_\infty}{|\boldsymbol{n}|_{\text{mix}}}\right)^{\frac{|\boldsymbol{l}|_1-\gamma m}{1-\gamma}} \right\} \\ &< m^{\frac{(d-1)(\gamma m-|\boldsymbol{l}|_1)}{1-\gamma}} N^{|\boldsymbol{l}|_1-m}, \end{aligned} \tag{6.51}$$

by Equation 6.50, Equation 6.47 and Equation 6.48. Otherwise, if $\frac{|\boldsymbol{l}|_1}{m} \leq \gamma < 1$, then

$$\max_{\boldsymbol{n}\in\boldsymbol{\Omega}^c_{N,\gamma,m}} \prod_{j=1}^{d} n_j^{l_j-m} \leq \max_{\boldsymbol{n}\in\boldsymbol{\Omega}^c_{N,\gamma,m}} \left\{ \left(\frac{|\boldsymbol{n}|_\infty^\gamma}{|\boldsymbol{n}|_{\text{mix}}}\right)^{m} |\boldsymbol{n}|_\infty^{|\boldsymbol{l}|_1-\gamma m} \right\} \leq N^{\frac{1-\gamma}{d-\gamma}(|\boldsymbol{l}|_1-\gamma m)-(1-\gamma)m}, \tag{6.52}$$

by Equation 6.50, Equation 6.47 and Equation 6.49. Combining Equation 6.46, Equation 6.51 and Equation 6.52, the first term on the right-hand side of Equation 6.44 has the upper bound

$$\text{XVI} \leq \left(\frac{m}{2}\right)^{dm-|\boldsymbol{l}|_1} \prod_{j=1}^{d} \alpha_j^{2(l_j-m)} \left|\left|\boldsymbol{\mathcal{D}}_{\boldsymbol{x}}^{m\cdot\mathbf{1}} u\right|\right|^2 \begin{cases} m^{\frac{(d-1)(\gamma m-|\boldsymbol{l}|_1)}{1-\gamma}} N^{|\boldsymbol{l}|_1-m}, & \text{if } 0 < \gamma \leq \dfrac{|\boldsymbol{l}|_1}{m} \\ N^{\frac{1-\gamma}{d-\gamma}(|\boldsymbol{l}|_1-\gamma m)-(1-\gamma)m}, & \text{if } \dfrac{|\boldsymbol{l}|_1}{m} \leq \gamma < 1. \end{cases} \tag{6.53}$$

Next, we consider XVII. Define $\mathcal{N}$ and $\mathcal{N}^c$ as in Equation 6.34. As in Equation 6.36, we obtain that

$$\text{XVII} \leq \max_{\boldsymbol{n}\in\boldsymbol{\Omega}^c_{N,\gamma,\boldsymbol{l}}\setminus\boldsymbol{\Omega}^c_{N,\gamma,m}} \left\{ \frac{\boldsymbol{\mu}_{\boldsymbol{n},\boldsymbol{l}}}{\boldsymbol{\mu}_{\boldsymbol{n},\boldsymbol{k}}} \right\} |u|^2_{\mathcal{K}^m_{\boldsymbol{\alpha},\boldsymbol{\beta}}(\mathbb{R}^d)}. \tag{6.54}$$

We then estimate the maximum similarly as in Equation 6.38:

$$\begin{aligned}\frac{\boldsymbol{\mu}_{\boldsymbol{n},\boldsymbol{l}}}{\boldsymbol{\mu}_{\boldsymbol{n},\boldsymbol{k}}} =& 2^{|\boldsymbol{l}|_1-|\boldsymbol{k}|_1}\prod_{j\in\mathcal{N}^c}\alpha_j^{2(l_j-m)}\prod_{j\in\mathcal{N}^c}n_j^{l_j-m}\prod_{j\in\mathcal{N}^c}\left(1-\frac{l_j}{n_j}\right)^{-1}\cdots\left(1-\frac{m-1}{n_j}\right)^{-1}\\ :=& D_2\prod_{j\in\mathcal{N}^c}n_j^{l_j-m}.\end{aligned} \tag{6.55}$$

Similar arguments as in Equation 6.32 yields

$$\max_{\boldsymbol{n}\in\boldsymbol{\Omega}^c_{N,\gamma,\boldsymbol{l}}\setminus\boldsymbol{\Omega}^c_{N,\gamma,m}} D_2 \le 2^{|\boldsymbol{l}|_1-|\boldsymbol{k}|_1}\prod_{j\in\mathcal{N}^c}\alpha_j^{2(l_j-m)}m^{dm-|\tilde{\boldsymbol{l}}|_1}, \tag{6.56}$$

where

$$\tilde{\boldsymbol{l}} = (l_1,\cdots,l_d) = \begin{cases} l_j, & \text{if } j\in\mathcal{N}^c \\ 0, & \text{otherwise.}\end{cases} \tag{6.57}$$

Then we verify that

$$\prod_{j\in\mathcal{N}^c}n_j^{l_j-m} \le \left(\prod_{j\in\mathcal{N}^c}|\tilde{\boldsymbol{n}}|_\infty^{l_j}\right)\left(\prod_{j\in\mathcal{N}^c}n_j\right)^{-m} = |\tilde{\boldsymbol{n}}|_\infty^{|\tilde{\boldsymbol{l}}|_1}|\tilde{\boldsymbol{n}}|_{\text{mix}}^{-m} \le |\tilde{\boldsymbol{n}}|_\infty^{|\boldsymbol{l}|_1}|\tilde{\boldsymbol{n}}|_{\text{mix}}^{-m}, \tag{6.58}$$

where $\tilde{\boldsymbol{n}}$ is defined similarly as $\tilde{\boldsymbol{l}}$ in Equation 6.57. With a similar argument as in Equation 6.47, we deduce that for any $\boldsymbol{n}\in\boldsymbol{\Omega}^c_{N,\gamma}$,

$$N^{1-\gamma} < |\boldsymbol{n}|_{\text{mix}}|\boldsymbol{n}|_\infty^{-\gamma} \le m^{d-1}|\tilde{\boldsymbol{n}}|_{\text{mix}}|\tilde{\boldsymbol{n}}|_\infty^{-\gamma} \Rightarrow \left(\frac{|\tilde{\boldsymbol{n}}|_\infty^\gamma}{|\tilde{\boldsymbol{n}}|_{\text{mix}}}\right)^{\frac{1}{1-\gamma}} < m^{\frac{d-1}{1-\gamma}}N^{-1}. \tag{6.59}$$

Similarly as in Equation 6.48, we have for any $\boldsymbol{n} \in \boldsymbol{\Omega}^c_{N,\gamma,m}$,

$$\frac{|\tilde{\boldsymbol{n}}|_\infty}{|\tilde{\boldsymbol{n}}|_{\text{mix}}} \leq \frac{1}{m^{d-2}}, \tag{6.60}$$

and

$$N^{1-\gamma} < m^{d-1}|\tilde{\boldsymbol{n}}|_{\text{mix}}|\tilde{\boldsymbol{n}}|_\infty^{-\gamma} \leq m^{d-1}|\tilde{\boldsymbol{n}}|_\infty^{d-1-\gamma} \Rightarrow |\tilde{\boldsymbol{n}}|_\infty > \left(\frac{N^{1-\gamma}}{m^{d-1}}\right)^{\frac{1}{d-1-\gamma}}, \tag{6.61}$$

by Equation 6.60. If $0 < \gamma \leq \frac{|\boldsymbol{l}|_1}{m}$, then

$$\begin{aligned}\max_{\boldsymbol{n}\in\boldsymbol{\Omega}^c_{N,\gamma,\boldsymbol{l}}\setminus\boldsymbol{\Omega}^c_{N,\gamma,m}} \prod_{j\in\mathcal{N}^c} n_j^{l_j-m} &< \max_{\boldsymbol{n}\in\boldsymbol{\Omega}^c_{N,\gamma,\boldsymbol{l}}\setminus\boldsymbol{\Omega}^c_{N,\gamma,m}} \left\{ \left(\frac{|\tilde{\boldsymbol{n}}|_\infty^\gamma}{|\tilde{\boldsymbol{n}}|_{\text{mix}}}\right)^{\frac{m-|\boldsymbol{l}|_1}{1-\gamma}} \left(\frac{|\tilde{\boldsymbol{n}}|_\infty}{|\tilde{\boldsymbol{n}}|_{\text{mix}}}\right)^{\frac{|\boldsymbol{l}|_1-\gamma m}{1-\gamma}} \right\} \\ &\leq m^{\frac{1}{1-\gamma}\{[(\gamma+1)d-(2\gamma+1)]m-(2d-3)|\boldsymbol{l}|_1\}} N^{|\boldsymbol{l}|_1-m},\end{aligned} \tag{6.62}$$

by Equation 6.58 - Equation 6.60. Otherwise, if $\frac{|\boldsymbol{l}|_1}{m} \leq \gamma < 1$, then

$$\begin{aligned}\max_{\boldsymbol{n}\in\boldsymbol{\Omega}^c_{N,\gamma,\boldsymbol{l}}\setminus\boldsymbol{\Omega}^c_{N,\gamma,m}} \prod_{j\in\mathcal{N}^c} n_j^{l_j-m} &< \max_{\boldsymbol{n}\in\boldsymbol{\Omega}^c_{N,\gamma,\boldsymbol{l}}\setminus\boldsymbol{\Omega}^c_{N,\gamma,m}} \left\{ \left(\frac{|\tilde{\boldsymbol{n}}|_\infty^\gamma}{|\tilde{\boldsymbol{n}}|_{\text{mix}}}\right)^m |\tilde{\boldsymbol{n}}|_\infty^{|\boldsymbol{l}|_1-\gamma m} \right\} \\ &\leq m^{(d-1)\left[m-\frac{|\boldsymbol{l}|_1-\gamma m}{d-1-\gamma}\right]} N^{\frac{(1-\gamma)[|\boldsymbol{l}|_1-(d-1)m]}{d-1-\gamma}},\end{aligned} \tag{6.63}$$

by Equation 6.58, Equation 6.59 and Equation 6.61. Combining Equation 6.38, Equation 6.54, Equation 6.56, Equation 6.62 and Equation 6.63, we arrive at

$$\begin{aligned}\text{XVII} \leq & 2^{|\boldsymbol{l}|_\infty - m} \prod_{j \in \mathcal{N}^c} \alpha_j^{2(l_j - m)} m^{dm - |\boldsymbol{l}|_1} |u|^2_{\mathcal{K}^m_{\boldsymbol{\alpha},\boldsymbol{\beta}}(\mathbb{R}^d)} \\ & \begin{cases} m^{\frac{1}{1-\gamma}\{[(\gamma+1)d-(2\gamma+1)]m-(2d-3)|\boldsymbol{l}|_1\}} N^{|\boldsymbol{l}|_1 - m}, & \text{if } 0 < \gamma \leq \dfrac{|\boldsymbol{l}|_1}{m} \\ m^{(d-1)\left[m - \frac{|\boldsymbol{l}|_1 - \gamma m}{d-1-\gamma}\right]} N^{\frac{(1-\gamma)[|\boldsymbol{l}|_1 - (d-1)m]}{d-1-\gamma}}, & \text{if } \dfrac{|\boldsymbol{l}|_1}{m} \leq \gamma < 1. \end{cases}\end{aligned} \tag{6.64}$$

Therefore, the desired result follows immediately from Equation 6.53 and Equation 6.64. □

Corollary 6.2. *For any $u \in \mathcal{K}^m_{\boldsymbol{\alpha},\boldsymbol{\beta}}(\mathbb{R}^d)$, $0 \leq l < m$, and $0 < \gamma \leq \frac{l}{m}$,*

$$\left|\left| P^{\boldsymbol{\alpha},\boldsymbol{\beta}}_{N,\gamma} u - u \right|\right|_{\mathcal{W}^l_{\boldsymbol{\alpha},\boldsymbol{\beta}}(\mathbb{R}^d)} \leq C_{\boldsymbol{\alpha},\boldsymbol{l},m,d,\gamma} N^{\frac{l-m}{2}} |u|_{\mathcal{K}^m_{\boldsymbol{\alpha},\boldsymbol{\beta}}(\mathbb{R}^d)}.$$

where $C_{\boldsymbol{\alpha},\boldsymbol{l},m,d,\gamma}$ is some constant depending on $\boldsymbol{\alpha}$, $\boldsymbol{l}$, m, d and γ.

Remark 6.5. *Due to the fact that $M = \operatorname{card}(\boldsymbol{\Omega}_{N,\gamma}) = \mathcal{O}(N) \leq CN$, we obtain*

$$\left|\left| P^{\boldsymbol{\alpha},\boldsymbol{\beta}}_{N,\gamma} u - u \right|\right|_{\mathcal{W}^l_{\boldsymbol{\alpha},\boldsymbol{\beta}}(\mathbb{R}^d)} \leq C_{\boldsymbol{\alpha},\boldsymbol{l},m,d,\gamma} M^{\frac{l-m}{2}} |u|_{\mathcal{K}^m_{\boldsymbol{\alpha},\boldsymbol{\beta}}(\mathbb{R}^d)}.$$

where $C_{\boldsymbol{\alpha},\boldsymbol{l},m,d,\gamma}$ is some constant depending on $\boldsymbol{\alpha}$, $\boldsymbol{l}$, m, d and γ. We see that the convergence rate no longer deteriorates with respect to d. The effect of the dimension goes into the constant in front.

6.1.3.4 Dimensional adaptive approximation

The standard sparse grids are isotropic, treating all of the dimensions equally. Many problems vary rapidly in only some dimensions, remaining less variable in other dimensions. In some situations, the highly changing dimensions can be recognized apriori. Consequently it is advantageous to treat them accordingly. Without loss of generality, we assume the first d_1 dimensions are the rapidly variable ones, and we wish to use the full grid. Meanwhile, the OHC approximation will be used in the remaining $d_2 := d - d_1$ dimensions.

Let us set $\boldsymbol{n} := \boldsymbol{n}_1 \bigoplus \boldsymbol{n}_2$, where $\boldsymbol{n}_1 = (n_1, \cdots, n_{d_1})$ and $\boldsymbol{n}_2 = (n_{d_1+1}, \cdots, n_d)$. The index set is

$$\boldsymbol{\Omega}_{N_1,N_2,\gamma} := \left\{ \boldsymbol{n} \in \mathbb{N}_0^d : \ |\boldsymbol{n}_1|_\infty \leq N_1, \ |\boldsymbol{n}_2|_{\text{mix}} |\boldsymbol{n}_2|_\infty^{-\gamma} \leq N_2^{1-\gamma} \right\}, \quad \forall \ -\infty < \gamma < 1. \tag{6.65}$$

The complement of the index set is

$$\boldsymbol{\Omega}^c_{N_1,N_2,\gamma} := \left\{ \boldsymbol{n} \in \mathbb{N}_0^d : \ |\boldsymbol{n}_1|_\infty > N_1 \quad \text{or} \quad |\boldsymbol{n}_2|_{\text{mix}} |\boldsymbol{n}_2|_\infty^{-\gamma} > N_2^{1-\gamma} \right\},$$

and the $\boldsymbol{k}-$complement of $\boldsymbol{\Omega}_{N_1,N_2,\gamma}$ is defined similarly as in Equation 6.42:

$$\boldsymbol{\Omega}^c_{N_1,N_2,\gamma,\boldsymbol{k}} := \left\{ \boldsymbol{n} \in \boldsymbol{\Omega}^c_{N_1,N_2,\gamma} : \ \boldsymbol{n} \geq \boldsymbol{k} \right\}, \ \forall \, \boldsymbol{k} \in \mathbb{N}_0^d.$$

The subspace $X^{\boldsymbol{\alpha},\boldsymbol{\beta}}_{N_1,N_2}$ is defined accordingly, i.e.,

$$X^{\boldsymbol{\alpha},\boldsymbol{\beta}}_{N_1,N_2} := \mathrm{span}\{\boldsymbol{\mathcal{H}}^{\boldsymbol{\alpha},\boldsymbol{\beta}}_{\boldsymbol{n}}(\boldsymbol{x}) :\ \boldsymbol{n} \in \boldsymbol{\Omega}_{N_1,N_2,\gamma}\}, \tag{6.66}$$

and so is the projection operator $P^{\boldsymbol{\alpha},\boldsymbol{\beta}}_{N_1,N_2,\gamma} :\ L^2(\mathbb{R}^d) \to X^{\boldsymbol{\alpha},\boldsymbol{\beta}}_{N_1,N_2}$.

Theorem 6.4. *For any $u \in \mathcal{K}^m_{\boldsymbol{\alpha},\boldsymbol{\beta}}(\mathbb{R}^d)$, for $0 < l \le m$, we have*

$$\left|P^{\boldsymbol{\alpha},\boldsymbol{\beta}}_{N_1,N_2,\gamma}u - u\right|_{W^l_{\boldsymbol{\alpha},\boldsymbol{\beta}}(\mathbb{R}^d)} \lesssim |\boldsymbol{\alpha}|^{l-m}_\infty \left(N_1^{l-m} + N_2^{\frac{1-\gamma}{d-d_1-\gamma}(l-m)}\right)^{\frac{1}{2}} |u|_{\mathcal{K}^m_{\boldsymbol{\alpha},\boldsymbol{\beta}}(\mathbb{R}^d)}.$$

Proof. Before we proceed to prove the theorem, we divide the index set $\boldsymbol{\Omega}^c_{N_1,N_2,\gamma}$ into two subsets:

$$\begin{aligned}
\Gamma_1 :=& \{\boldsymbol{n} \in \boldsymbol{\Omega}^c_{N_1,N_2,\gamma} :\ |\boldsymbol{n}_1|_\infty > N_1\},\\
\Gamma_2 :=& \{\boldsymbol{n} \in \boldsymbol{\Omega}^c_{N_1,N_2,\gamma} :\ |\boldsymbol{n}_1|_\infty \le N_1 \text{ and } |\boldsymbol{n}_2|_{\mathrm{mix}}|\boldsymbol{n}_2|^{-\gamma}_\infty > N_2^{1-\gamma}\}.
\end{aligned}$$

Our proof mainly follows the proof of Theorem 6.1:

$$\begin{aligned}
\left|P^{\boldsymbol{\alpha},\boldsymbol{\beta}}_{N_1,N_2,\gamma}u - u\right|^2_{\mathcal{W}^l_{\boldsymbol{\alpha},\boldsymbol{\beta}}(\mathbb{R}^d)} =& \sum_{j=1}^d \sum_{\boldsymbol{n}\in\boldsymbol{\Omega}^c_{N_1,N_2,\gamma}} \mu_{n_j,l}\left|\hat{u}^{\boldsymbol{\alpha},\boldsymbol{\beta}}_{\boldsymbol{n}}\right|^2\\
=& \sum_{j=1}^d \sum_{\boldsymbol{n}\in\Gamma_1} \mu_{n_j,l}\left|\hat{u}^{\boldsymbol{\alpha},\boldsymbol{\beta}}_{\boldsymbol{n}}\right|^2 + \sum_{j=1}^d \sum_{\boldsymbol{n}\in\Gamma_2} \mu_{n_j,l}\left|\hat{u}^{\boldsymbol{\alpha},\boldsymbol{\beta}}_{\boldsymbol{n}}\right|^2 := \mathrm{XVIII} + \mathrm{XIX},
\end{aligned} \tag{6.67}$$

by Equation 6.20. For XVIII, for any $1 \leq j \leq d$,

$$\text{XVIII} = \sum_{\boldsymbol{n} \in \Lambda_{N_1}^{1,j}} \mu_{n_j,l} \left| \hat{u}_{\boldsymbol{n}}^{\boldsymbol{\alpha},\boldsymbol{\beta}} \right|^2 + \sum_{\boldsymbol{n} \in \Lambda_{N_1}^{2,j}} \mu_{n_j,l} \left| \hat{u}_{\boldsymbol{n}}^{\boldsymbol{\alpha},\boldsymbol{\beta}} \right|^2 := \text{XVIII}_1 + \text{XVIII}_2,$$

where

$$\Lambda_{N_1}^{1,j} := \{\boldsymbol{n} \in \Gamma_1 : \ n_j > N_1\}, \quad \Lambda_{N_1}^{2,j} := \{\boldsymbol{n} \in \Gamma_1 : \ n_j \leq N_1\}.$$

For XVIII_1:

$$\text{XVIII}_1 \leq \max_{\boldsymbol{n} \in \Lambda_{N_1}^{1,j}} \left\{ \frac{\mu_{n_j,l}}{\mu_{n_j,m}} \right\} \sum_{\boldsymbol{n} \in \Lambda_{N_1}^{1,j}} \mu_{n_j,m} \left| \hat{u}_{\boldsymbol{n}}^{\boldsymbol{\alpha},\boldsymbol{\beta}} \right|^2 \leq 2^{l-m} |\boldsymbol{\alpha}|_\infty^{2(l-m)} (N_1 - m + 1)^{l-m} |u|^2_{\mathcal{K}^m_{\boldsymbol{\alpha},\boldsymbol{\beta}}(\mathbb{R}^d)}, \tag{6.68}$$

by Equation 6.24. For XVIII_2, since $\boldsymbol{n} \in \Gamma_1$, there exists some $j_0 \in \{1, \cdots, d_1\}$ such that $n_{j_0} > N_1$. Then

$$\text{XVIII}_2 \leq \max_{\boldsymbol{n} \in \Lambda_{N_1}^{2,j}} \left\{ \frac{\mu_{n_j,l}}{\mu_{n_{j_0},m}} \right\} \sum_{\boldsymbol{n} \in \Lambda_{N_1}^{2,j}} \mu_{n_{j_0},m} \left| \hat{u}_{\boldsymbol{n}}^{\boldsymbol{\alpha},\boldsymbol{\beta}} \right|^2 \leq 2^{l-m} |\boldsymbol{\alpha}|_\infty^{2l} \left| \frac{1}{\boldsymbol{\alpha}} \right|_\infty^{2m} (N_1 - m)^{l-m-2} |u|_{\mathcal{K}^m_{\boldsymbol{\alpha},\boldsymbol{\beta}}(\mathbb{R}^d)}, \tag{6.69}$$

by Equation 6.25. Hence, combining Equation 6.68 and Equation 6.69, we have

$$\text{XVIII} \lesssim |\boldsymbol{\alpha}|_\infty^{2(l-m)} N_1^{l-m} |u|^2_{\mathcal{K}^m_{\boldsymbol{\alpha},\boldsymbol{\beta}}(\mathbb{R}^d)}. \tag{6.70}$$

For XIX, we deduce, as in Equation 6.49, that

$$|\boldsymbol{n}_2|_{\text{mix}}|\boldsymbol{n}_2|_\infty^{-\gamma} > N_2^{1-\gamma} \Rightarrow |\boldsymbol{n}_2|_\infty > N_2^{\frac{1-\gamma}{d-d_1-\gamma}}. \tag{6.71}$$

With the similar argument for XVIII, we write

$$\text{XIX} = \sum_{\boldsymbol{n}\in\Lambda_{N_2}^{1,j}} \mu_{n_j,l} \left|\hat{u}_{\boldsymbol{n}}^{\boldsymbol{\alpha},\boldsymbol{\beta}}\right|^2 + \sum_{\boldsymbol{n}\in\Lambda_{N_2}^{2,j}} \mu_{n_j,l} \left|\hat{u}_{\boldsymbol{n}}^{\boldsymbol{\alpha},\boldsymbol{\beta}}\right|^2 := \text{XIX}_1 + \text{XIX}_2,$$

where

$$\Lambda_{N_2}^{1,j} := \left\{\boldsymbol{n} \in \Gamma_2 :\ n_j > N_2^{\frac{1-\gamma}{d-d_1-\gamma}}\right\}, \quad \Lambda_{N_2}^{2,j} := \left\{\boldsymbol{n} \in \Gamma_2 :\ n_j \le N_2^{\frac{1-\gamma}{d-d_1-\gamma}}\right\}.$$

Thus,

$$\text{XIX}_1 \le \max_{\boldsymbol{n}\in\Lambda_{N_2}^{1,j}} \left\{\frac{\mu_{n_j,l}}{\mu_{n_j,m}}\right\} \sum_{\boldsymbol{n}\in\Lambda_{N_2}^{1,j}} \mu_{n_j,m} \left|\hat{u}_{\boldsymbol{n}}^{\boldsymbol{\alpha},\boldsymbol{\beta}}\right|^2 \le 2^{l-m}|\boldsymbol{\alpha}|_\infty^{2(l-m)}(N_2^{\frac{1-\gamma}{d-d_1-\gamma}} - m + 1)^{l-m}|u|^2_{\mathcal{K}^m_{\boldsymbol{\alpha},\boldsymbol{\beta}}(\mathbb{R}^d)}, \tag{6.72}$$

by Equation 6.71. There exists some $j_0 \in \{d_1+1, \cdots, d\}$ such that $n_{j_0} > N_2^{\frac{1-\gamma}{d-d_1-\gamma}}$, then

$$\begin{aligned}\text{XIX}_2 \le& \max_{\boldsymbol{n}\in\Lambda_{N_2}^{2,j}} \left\{\frac{\mu_{n_j,l}}{\mu_{n_{j_0},m}}\right\} \sum_{\boldsymbol{n}\in\Lambda_{N_2}^{2,j}} \mu_{n_{j_0},m} \left|\hat{u}_{\boldsymbol{n}}^{\boldsymbol{\alpha},\boldsymbol{\beta}}\right|^2 \\ \le& 2^{l-m}|\boldsymbol{\alpha}|_\infty^{2l} \left|\frac{1}{\boldsymbol{\alpha}}\right|_\infty^{2m} \left(\lfloor N_2^{\frac{1-\gamma}{d-d_1-\gamma}} \rfloor - m\right)^{l-m-2} |u|^2_{\mathcal{K}^m_{\boldsymbol{\alpha},\boldsymbol{\beta}}(\mathbb{R}^d)},\end{aligned} \tag{6.73}$$

where $\lfloor \cdot \rfloor$ denotes the largest integer smaller or equal to $\cdot$. The estimate of XIX follows immediately from Equation 6.72 and Equation 6.73:

$$\text{XIX} \lesssim |\boldsymbol{\alpha}|_{\infty}^{2(l-m)} N_2^{\frac{1-\gamma}{d-d_1-\gamma}(l-m)} |u|^2_{\mathcal{K}^m_{\boldsymbol{\alpha},\boldsymbol{\beta}}(\mathbb{R}^d)}. \tag{6.74}$$

The desired result follows from Equation 6.70 and Equation 6.74. □

6.2 Application to linear parabolic PDE

In this section, we shall study the Galerkin HSM with the HC approximation applied to higher dimensional linear parabolic PDEs. Let us consider a linear parabolic PDE of the general form:

$$\begin{cases} \partial_t u(\boldsymbol{x},t) + \mathcal{L}u(\boldsymbol{x},t) = f(\boldsymbol{x},t), \quad \boldsymbol{x} \in \mathbb{R}^d,\ t \in [0,T] \\ \\ u(\boldsymbol{x},0) = u_0(\boldsymbol{x}), \end{cases} \tag{6.75}$$

where

$$\mathcal{L}u = -\nabla \cdot (\boldsymbol{A}\nabla u) + \boldsymbol{b} \cdot \nabla u + cu, \tag{6.76}$$

with $\boldsymbol{A} = (a_{ij})_{i,j=1}^d :\ \mathbb{R}^d \mapsto \mathbb{R}^{d\times d}$, $\boldsymbol{b} = (b_i)_{i=1}^d :\ \mathbb{R}^d \mapsto \mathbb{R}^d$ and $c :\ \mathbb{R}^d \mapsto \mathbb{R}$. The aim of HSM is to find $u_N \in X$, such that

$$\langle \partial_t u_N, \varphi \rangle - \mathcal{A}(u_N, \varphi) = \langle f, \varphi \rangle, \quad \forall\, \varphi \in X, \tag{6.77}$$

where X is some approximate space, and $\mathcal{A}(u,v)$ is a bilinear form given by

$$\mathcal{A}(u,v) = \int_{\mathbb{R}^d} (\nabla u)^T \boldsymbol{A} \nabla v + v\boldsymbol{b} \cdot \nabla u + cuv \; d\boldsymbol{x}. \tag{6.78}$$

In our context, X could be chosen as $X_N^{\boldsymbol{\alpha},\boldsymbol{\beta}}$ or $X_{N,\gamma}^{\boldsymbol{\alpha},\boldsymbol{\beta}}$ in the previous section.

To guarantee the existence and regularity of the solution to Equation 6.75, we assume that

$(\boldsymbol{C_1})$ The bilinear form is continuous, i.e., there is a constant $C > 0$ such that

$$|\mathcal{A}(u,v)| \leq C||u||_{H_0^1(\mathbb{R}^d)}||v||_{H_0^1(\mathbb{R}^d)}, \quad \forall\, u, v \in H_0^1(\mathbb{R}^d). \tag{6.79}$$

$(\boldsymbol{C_2})$ The bilinear form is coercive, i.e., there exists some $c > 0$ such that

$$\mathcal{A}(u,u) \geq c||u||^2_{H_0^1(\mathbb{R}^d)}, \quad \forall\, u \in H_0^1(\mathbb{R}^d). \tag{6.80}$$

$(\boldsymbol{C_3})$ The coefficients a_{ij}, b_i and c are smooth.

Here, $H_0^1(\mathbb{R}^d)$ denotes the normal Sobolev space with the functions decaying to zero at infinity. More generally, $H_0^m(\mathbb{R}^d)$ is defined as, for any $u \in H^m(\mathbb{R}^d)$, it satisfies $|u| \to 0$, as $|\boldsymbol{x}| \to \infty$ and

$$||u||^2_{H^m(\mathbb{R}^d)} = \sum_{0 \leq |\boldsymbol{k}|_1 \leq m} \left|\left| \boldsymbol{\partial}_{\boldsymbol{x}}^{\boldsymbol{k}} u \right|\right|^2 < \infty. \tag{6.81}$$

Let us first show some relationships between the Sobolev-type space $W^l_{\boldsymbol{\alpha},\boldsymbol{\beta}}(\mathbb{R}^d)$ (see Equation 6.17) and the normal Sobolev space $H^l(\mathbb{R}^d)$.

Lemma 6.1. *For $u \in \mathcal{W}_{\boldsymbol{\alpha},\boldsymbol{\beta}}^{|\boldsymbol{k}|_1+|\boldsymbol{r}|_1}(\mathbb{R}^d)$, for any $\boldsymbol{r}, \boldsymbol{k} \in \mathbb{N}_0^d$, we have*

$$\left|\left|\boldsymbol{x}^{\boldsymbol{r}} \boldsymbol{\partial}_{\boldsymbol{x}}^{\boldsymbol{k}} u\right|\right| \lesssim \left(\prod_{i=1}^{d} \alpha_i^{-r_i}\right) |\boldsymbol{k}+\boldsymbol{r}|_{\text{mix}}^{\frac{1}{2}} \cdot ||u||_{\mathcal{W}_{\boldsymbol{\alpha},\boldsymbol{\beta}}^{|\boldsymbol{k}|_1+|\boldsymbol{r}|_1}(\mathbb{R}^d)}.$$

Proof. For clarity, we show it holds for $d = 1$ in detail. We have

$$\left|\left|x^r \partial_x^k u\right|\right|^2 = \left|\left|\sum_{n=0}^{\infty} \hat{u}_n^{\alpha,\beta} x^r \partial_x^k \mathcal{H}_n^{\alpha,\beta}(x)\right|\right|^2 = \alpha^{-2r} \left|\left|\sum_{n=0}^{\infty} \hat{u}_n^{\alpha,\beta} \sum_{i=-(k+r)}^{k+r} \eta_{n,i} \mathcal{H}_{n+i}^{\alpha,\beta}(x)\right|\right|^2, \tag{6.82}$$

by Equation 6.4 and Equation 6.5, where, for each n, $\eta_{n,i}$ is a product of $k+r$ factors of $\left(\pm\frac{\sqrt{\lambda_{n+i}}}{2}\right)$ or $\frac{\beta}{2}$ with $-(k+r) \le i \le k+r$. Notice that

$$\lambda_{n+i} \sim \lambda_{n+j}, \tag{6.83}$$

provided that $\lambda_{n+i}, \lambda_{n+j} \neq 0$, for all $-(k+r) \le i, j \le k+r$. In fact, it is equivalent to show that $\lambda_n \sim \lambda_{n+l}$, for all $0 \le l \le 2(k+r)$. By convention, $\lambda_n = 0$, if $n \le 0$. Notice that

$$\frac{\lambda_n}{\lambda_{n+l}} = \frac{n}{n+l} \le 1 \quad \text{and} \quad \frac{n}{n+l} \ge \frac{1}{1+l} \ge \frac{1}{1+2(k+r)}, \ \forall n \ge 1.$$

Meanwhile $\lim_{n\to\infty}\frac{n}{n+l}=1$, for all $0\le l\le 2(k+r)$. Therefore, $\frac{\lambda_n}{\lambda_{n+l}}\sim 1$. Hence, $\eta_{n,j}\lesssim\sqrt{\mu_{n,k+r}}$, by Equation 6.7 and Equation 6.83. Thus,

$$\begin{aligned}\left|\left|x^r\partial_x^k u\right|\right|^2 \sim&\alpha^{-2r}\left|\left|\sum_{n=0}^{\infty}\hat{u}_n^{\alpha,\beta}\sqrt{\mu_{n,k+r}}\sum_{i=-(k+r)}^{k+r}\mathcal{H}_{n+i}^{\alpha,\beta}(x)\right|\right|^2\\ =&\alpha^{-2r}\sum_{n=0}^{\infty}\hat{u}_n^{\alpha,\beta}\sqrt{\mu_{n,k+r}}\sum_{i=-(k+r)}^{k+r}\sum_{l=0}^{\infty}\hat{u}_l^{\alpha,\beta}\sqrt{\mu_{l,k+r}}\left\langle\mathcal{H}_{n+i}^{\alpha,\beta}(x),\sum_{j=-(k+r)}^{k+r}\mathcal{H}_{l+j}^{\alpha,\beta}(x)\right\rangle,\end{aligned}\tag{6.84}$$

by Equation 6.82. It is clear that the scalar product in Equation 6.84 is nonzero only if $l=n+i-j$. Also $\mu_{n,k+r}\sim\mu_{n+i-j,k+r}$, for all $-(k+r)\le i,j\le k+r$, which can be verified by Equation 6.7 and Equation 6.83. Therefore,

$$\begin{aligned}\left|\left|x^r\partial_x^k u\right|\right|^2 \sim&\alpha^{-2r}\sum_{n=0}^{\infty}\mu_{n,k+r}\hat{u}_n^{\alpha,\beta}\sum_{\tilde{l}=-2(k+r)}^{2(k+r)}\hat{u}_{n+\tilde{l}}^{\alpha,\beta}\le\alpha^{-2r}\sum_{n=0}^{\infty}\mu_{n,k+r}\sum_{\tilde{l}=-2(k+r)}^{2(k+r)}\left|\hat{u}_n^{\alpha,\beta}\right|\left|\hat{u}_{n+\tilde{l}}^{\alpha,\beta}\right|\\ \le&\alpha^{-2r}\sum_{n=0}^{\infty}\mu_{n,k+r}\frac{1}{2}\sum_{\tilde{l}=-2(k+r)}^{2(k+r)}\left(\left|\hat{u}_n^{\alpha,\beta}\right|^2+\left|\hat{u}_{n+\tilde{l}}^{\alpha,\beta}\right|^2\right)\\ =&\alpha^{-2r}\sum_{n=0}^{\infty}\mu_{n,k+r}\left[2(k+r)\left|\hat{u}_n^{\alpha,\beta}\right|^2+\frac{1}{2}\sum_{\tilde{l}=-2(k+r)}^{2(k+r)}\left|\hat{u}_{n+\tilde{l}}^{\alpha,\beta}\right|^2\right]\\ =&2(k+r)\alpha^{-2r}\sum_{n=0}^{\infty}\mu_{n,k+r}\left|\hat{u}_n^{\alpha,\beta}\right|^2+\frac{1}{2}\alpha^{-2r}\sum_{\tilde{n}=0}^{\infty}\sum_{\tilde{l}=-2(k+r)}^{2(k+r)}\mu_{\tilde{n}-\tilde{l},k+r}\left|\hat{u}_{\tilde{n}}^{\alpha,\beta}\right|^2\\ \sim&\alpha^{-2r}4(k+r)\sum_{n=0}^{\infty}\mu_{n,k+r}\left|\hat{u}_n^{\alpha,\beta}\right|^2\lesssim\alpha^{-2r}(k+r)||u||^2_{\mathcal{W}_{\alpha,\beta}^{k+r}(\mathbb{R})},\end{aligned}$$

where the first inequality is followed by Equation 6.84. Until now, we have shown that Equation 6.85 holds for $d = 1$. For $d \geq 2$, we shall proceed similarly as for $d = 1$. Then

$$\begin{aligned}
\left|\left|\boldsymbol{x}^{\boldsymbol{r}}\partial_{\boldsymbol{x}}^{\boldsymbol{k}}u\right|\right|^2 &= \left(\prod_{\tilde{i}=1}^{d}\alpha_{\tilde{i}}^{-2r_{\tilde{i}}}\right)\left|\left|\sum_{\boldsymbol{n}\in\mathbb{N}_0^d}\hat{u}_{\boldsymbol{n}}^{\boldsymbol{\alpha},\boldsymbol{\beta}}\sum_{-(\boldsymbol{k}+\boldsymbol{r})\leq\boldsymbol{i}\leq\boldsymbol{k}+\boldsymbol{r}}\boldsymbol{\eta}_{\boldsymbol{n},\boldsymbol{i}}\boldsymbol{\mathcal{H}}_{\boldsymbol{n}+\boldsymbol{i}}^{\boldsymbol{\alpha},\boldsymbol{\beta}}(\boldsymbol{x})\right|\right|^2 \\
&\sim \left(\prod_{\tilde{i}=1}^{d}\alpha_{\tilde{i}}^{-2r_{\tilde{i}}}\right)\left|\left|\sum_{\boldsymbol{n}\in\mathbb{N}_0^d}\hat{u}_{\boldsymbol{n}}^{\boldsymbol{\alpha},\boldsymbol{\beta}}\sqrt{\boldsymbol{\mu}_{\boldsymbol{n},\boldsymbol{k}+\boldsymbol{r}}}\sum_{-(\boldsymbol{k}+\boldsymbol{r})\leq\boldsymbol{i}\leq(\boldsymbol{k}+\boldsymbol{r})}\boldsymbol{\mathcal{H}}_{\boldsymbol{n}+\boldsymbol{i}}^{\boldsymbol{\alpha},\boldsymbol{\beta}}(\boldsymbol{x})\right|\right|^2 \\
&\lesssim \left(\prod_{\tilde{i}=1}^{d}\alpha_{\tilde{i}}^{-2r_{\tilde{i}}}\right)\sum_{\boldsymbol{n}\in\mathbb{N}_0^d}\boldsymbol{\mu}_{\boldsymbol{n},\boldsymbol{k}+\boldsymbol{r}}\sum_{-2(\boldsymbol{k}+\boldsymbol{r})\leq\tilde{\boldsymbol{l}}\leq2(\boldsymbol{k}+\boldsymbol{r})}\left(\left|\hat{u}_{\boldsymbol{n}}^{\boldsymbol{\alpha},\boldsymbol{\beta}}\right|^2+\left|\hat{u}_{\boldsymbol{n}+\tilde{\boldsymbol{l}}}^{\boldsymbol{\alpha},\boldsymbol{\beta}}\right|^2\right) \\
&\sim \left(\prod_{\tilde{i}=1}^{d}\alpha_{\tilde{i}}^{-2r_{\tilde{i}}}\right)|\boldsymbol{k}+\boldsymbol{r}|_{\text{mix}}\sum_{\boldsymbol{n}\in\mathbb{N}_0^d}\boldsymbol{\mu}_{\boldsymbol{n},\boldsymbol{k}+\boldsymbol{r}}\left|\hat{u}_{\boldsymbol{n}}^{\boldsymbol{\alpha},\boldsymbol{\beta}}\right|^2 \\
&\lesssim \left(\prod_{\tilde{i}=1}^{d}\alpha_{\tilde{i}}^{-2r_{\tilde{i}}}\right)|\boldsymbol{k}+\boldsymbol{r}|_{\text{mix}}\cdot||u||^2_{\mathcal{W}^{|\boldsymbol{k}|_1+|\boldsymbol{r}|_1}_{\boldsymbol{\alpha},\boldsymbol{\beta}}(\mathbb{R}^d)}.
\end{aligned}$$

Therefore, we obtain the desired result. □

Corollary 6.3. *For $u \in \mathcal{W}^m_{\boldsymbol{\alpha},\boldsymbol{\beta}}(\mathbb{R}^d)$, we have $||u||_{H^m(\mathbb{R}^d)} \lesssim ||u||_{\mathcal{W}^m_{\boldsymbol{\alpha},\boldsymbol{\beta}}(\mathbb{R}^d)}$, for all $m \geq 0$.*

Proof. From the definitions of $W^m_{\boldsymbol{\alpha},\boldsymbol{\beta}}(\mathbb{R}^d)$ and $H^m(\mathbb{R}^d)$ in Equation 6.18 and Equation 6.81, respectively, we need to show that

$$\left|\left|\boldsymbol{\partial}_{\boldsymbol{x}}^{\boldsymbol{k}}u\right|\right|^2 \lesssim \left|\left|\boldsymbol{\mathcal{D}}_{\boldsymbol{x}}^{\boldsymbol{k}}u\right|\right|^2 = \sum_{\boldsymbol{n}\in\mathbb{N}_0^d}\boldsymbol{\mu}_{\boldsymbol{n},\boldsymbol{k}}\left|\hat{u}_{\boldsymbol{n}}^{\boldsymbol{\alpha},\boldsymbol{\beta}}\right|^2, \tag{6.85}$$

for all $0 \leq |\boldsymbol{k}|_1 \leq m$, by Equation 6.14. The desired results follows immediately from Lemma 6.1 by letting $\boldsymbol{r} = \boldsymbol{0}$, i.e.,

$$\left|\left|\boldsymbol{\partial}_{\boldsymbol{x}}^{\boldsymbol{k}} u\right|\right|^2 \lesssim |\boldsymbol{k}|_{\text{mix}} \cdot \left|\left|\boldsymbol{\mathcal{D}}_{\boldsymbol{x}}^{\boldsymbol{k}} u\right|\right|^2.$$

□

The convergence rate of the HSM with the HC approximation under the assumptions **(C_1)**-**(C_3)** is given below.

Theorem 6.5. *Assume that conditions* **(C_1)**-**(C_3)** *are satisfied, and the solution* $u \in L^\infty(0,T;\mathcal{K}^m_{\boldsymbol{\alpha},\boldsymbol{\beta}}(\mathbb{R}^d)) \cap L^2(0,T;\mathcal{K}^m_{\boldsymbol{\alpha},\boldsymbol{\beta}}(\mathbb{R}^d))$*, for* $m > 1$*. Let* u_N *is the approximate solution obtained by HSM, i.e., the solution to Equation 6.77. Then*

$$||u - u_N||(t) \lesssim c^* N^{\frac{1-m}{2}},$$

where c^* *depends on* $\boldsymbol{\alpha}$ *and the norms of* $L^2(0,T;\mathcal{K}^m_{\boldsymbol{\alpha},\boldsymbol{\beta}}(\mathbb{R}^d))$ *and* $L^\infty(0,T;\mathcal{K}^m_{\boldsymbol{\alpha},\boldsymbol{\beta}}(\mathbb{R}^d))$.

Proof. For notational convenience, we denote $U_N = P_N^{\boldsymbol{\alpha},\boldsymbol{\beta}} u$. It is readily verified that

$$\langle \partial_t (u - U_N), \varphi \rangle = 0 \quad \Rightarrow \quad \langle \partial_t U_N, \varphi \rangle = \langle -\mathcal{L}u + f, \varphi \rangle, \quad \forall\, \varphi \in X_N^{\boldsymbol{\alpha},\boldsymbol{\beta}}. \tag{6.86}$$

Combining with the formulation of HSM, i.e., Equation 6.77, we have

$$\begin{aligned}\langle \partial_t (U_N - u_N), \varphi \rangle =& \langle -\mathcal{L}u + f, \varphi \rangle + \mathcal{A}(u_N, \varphi) + \langle f, \varphi \rangle = \mathcal{A}(u_N - u, \varphi) \\ =& - \mathcal{A}(u - U_N, \varphi) - \mathcal{A}(U_N - u_N, \varphi), \quad \forall\, \varphi \in X_N^{\boldsymbol{\alpha},\boldsymbol{\beta}}.\end{aligned}$$

Take $\varphi = 2(U_N - u_N) \in X_N^{\boldsymbol{\alpha},\boldsymbol{\beta}}$, then

$$\begin{aligned}\partial_t ||U_N - u_N||^2 =& - 2\mathcal{A}(u - U_N, U_N - u_N) - 2\mathcal{A}(U_N - u_N, U_N - u_N) \\ \leq& 2C||u - U_N||_{H_0^1(\mathbb{R}^d)} ||U_N - u_N||_{H_0^1(\mathbb{R}^d)} - 2c||U_N - u_N||^2_{H_0^1(\mathbb{R}^d)} \\ \lesssim& ||u - U_N||^2_{H_0^1(\mathbb{R}^d)},\end{aligned}$$

by Equation 6.79, Equation 6.80 and Young's inequality. With Corollary 6.3 and Corollary 6.2 (if the OHC approximation is considered), we have

$$\begin{aligned}& \partial_t ||U_N - u_N||^2 \lesssim ||u - U_N||^2_{\mathcal{W}^1_{\boldsymbol{\alpha},\boldsymbol{\beta}}(\mathbb{R}^d)} \lesssim N^{1-m} |u|^2_{\mathcal{K}^m_{\boldsymbol{\alpha},\boldsymbol{\beta}}(\mathbb{R}^d)} \\ \Rightarrow \quad & ||U_N - u_N||^2(t) \lesssim N^{\frac{1-m}{2}} \left[\int_0^t |u|^2_{\mathcal{K}^m_{\boldsymbol{\alpha},\boldsymbol{\beta}}(\mathbb{R}^d)}(s) ds \right]^{\frac{1}{2}}.\end{aligned}$$

The same estimate holds for the RHC approximation with Corollary 6.2 replaced by Corollary 6.1. Then, we have

$$\begin{aligned}||u-u_N||(t) \leq & ||u-U_N||(t)+||U_N-u_N||(t) \\ \lesssim & N^{-\frac{m}{2}}|u|_{\mathcal{K}^m_{\boldsymbol{\alpha},\boldsymbol{\beta}}(\mathbb{R}^d)}(t)+N^{\frac{1-m}{2}}\left[\int_0^t |u|^2_{\mathcal{K}^m_{\boldsymbol{\alpha},\boldsymbol{\beta}}(\mathbb{R}^d)}(s)ds\right]^{\frac{1}{2}} \lesssim c^* N^{\frac{1-m}{2}},\end{aligned}$$

where c^* depends on $\boldsymbol{\alpha}$, the norms of $L^2(0,T;\mathcal{K}^m_{\boldsymbol{\alpha},\boldsymbol{\beta}}(\mathbb{R}^d))$ and $L^\infty(0,T;\mathcal{K}^m_{\boldsymbol{\alpha},\boldsymbol{\beta}}(\mathbb{R}^d))$. □

However, the assumptions $(\boldsymbol{C_1})$ and $(\boldsymbol{C_2})$ are not easy to verify. In the sequel, we make assumptions on the operator $\mathcal{L}$ and the convergence rate of the HSM is investigated under the conditions below. Assume that:

$(\boldsymbol{C_4})$ The operator $\mathcal{L}$ (c.f. Equation 6.76) is strongly elliptic and uniformly bounded, i.e.,

$$\sum_{i,j=1}^d a_{ij}(\boldsymbol{x})\xi_i\xi_j \geq \theta|\xi|^2, \quad \forall \xi \in \mathbb{R}^d, \quad \text{and} \quad ||\boldsymbol{A}||_\infty = \max_{i,j=1,\cdots,d}||a_{ij}||_\infty < \infty,$$

for $\boldsymbol{x} \in \mathbb{R}^d$, where $\theta > 0$.

$(\boldsymbol{C_5})$ There exists some constant $C > 0$, such that

$$c(\boldsymbol{x}) - \frac{1}{2}\nabla \cdot \boldsymbol{b}(\boldsymbol{x}) \geq -C,$$

for all $\boldsymbol{x} \in \mathbb{R}^d$.

(C_6) There exist some integer indices $\boldsymbol{\gamma}, \boldsymbol{\delta} \in \mathbb{N}_0^d$, such that

$$c(\boldsymbol{x}) \lesssim 1 + \boldsymbol{x}^{2\boldsymbol{\gamma}} \quad \text{and} \quad b_i(\boldsymbol{x}) \lesssim 1 + \boldsymbol{x}^{2\boldsymbol{\delta}}, \quad \forall\, i = 1, 2, \cdots, d,$$

for all $\boldsymbol{x} \in \mathbb{R}^d$.

Theorem 6.6. *Assume that conditions* **(C_3)**-**(C_6)** *are satisfied. The solution to Equation 6.75 is $u \in L^2(0,T;\mathcal{K}^m_{\boldsymbol{\alpha},\boldsymbol{\beta}}(\mathbb{R}^d))$, for some integer $m > \max\{|\boldsymbol{\gamma}|_1, |\boldsymbol{\delta}|_1 + 1\}$. Now let u_N be the approximate solution obtained by HSM, i.e., Equation 6.77, then*

$$||u - u_N||(t) \lesssim c^\sharp N^{\frac{\max\{|\boldsymbol{\gamma}|_1, |\boldsymbol{\delta}|_1+1\}-m}{2}},$$

where $c^\sharp$ depends on $\boldsymbol{\alpha}$, T and the norm of $L^2(0,T;\mathcal{K}^m_{\boldsymbol{\alpha},\boldsymbol{\beta}}(\mathbb{R}^d))$.

Proof. Similarly as we argued in the proof of Theorem 6.5, denote $U_N = P_N^{\boldsymbol{\alpha},\boldsymbol{\beta}} u$ for convenience, and let $\varphi = 2(U_N - u_N) \in X_N^{\boldsymbol{\alpha},\boldsymbol{\beta}}$, then

$$\partial_t ||U_N - u_N||^2 = -2\mathcal{A}(u - U_N, U_N - u_N) - 2\mathcal{A}(U_N - u_N, U_N - u_N) := \mathrm{XX} + \mathrm{XXI}, \quad (6.87)$$

where $\mathcal{A}$ is defined in Equation 6.78. For XXI,

$$\begin{aligned}
-\frac{1}{2}\text{XXI} =& \int_{\mathbb{R}^d} (\nabla(U_N - u_N))^T \boldsymbol{A} (\nabla(U_N - u_N)) + \int_{\mathbb{R}^d} (U_N - u_N)\boldsymbol{b} \cdot \nabla(U_N - u_N) \\
&+ \int_{\mathbb{R}^d} c(U_N - u_N)^2 \\
=& \int_{\mathbb{R}^d} (\nabla(U_N - u_N))^T \boldsymbol{A} (\nabla(U_N - u_N)) + \int_{\mathbb{R}^d} \left(c - \frac{1}{2}\nabla \cdot \boldsymbol{b} \right) (U_N - u_N)^2 \\
\geq& \theta ||\nabla(U_N - u_N)||^2 - C||U_N - u_N||^2, \qquad (6.88)
\end{aligned}$$

by $(\boldsymbol{C_4})$ and $(\boldsymbol{C_5})$. Meanwhile, for XX,

$$\begin{aligned}
|\text{XX}| =& 2 \left[\int_{\mathbb{R}^d} (\nabla(u - U_N))^T \boldsymbol{A} (\nabla(U_N - u_N)) + \int_{\mathbb{R}^d} (U_N - u_N)\boldsymbol{b} \cdot \nabla(u - U_N) \right. \\
&\left. + \int_{\mathbb{R}^d} c(u - U_N)(U_N - u_N) \right] \\
\leq& 2[||\boldsymbol{A}||_\infty ||\nabla(u - U_N)|| \cdot ||\nabla(U_N - u_N)|| + ||\boldsymbol{b} \cdot \nabla(u - U_N)|| \cdot ||U_N - u_N|| \\
&+ ||c(u - U_N)|| \cdot ||U_N - u_N||] \\
\lesssim& C_{||\boldsymbol{A}||_\infty, \theta} ||\nabla(u - U_N)||^2 + 2\theta ||\nabla(U_N - u_N)||^2 + ||\boldsymbol{b} \cdot \nabla(u - U_N)||^2 + ||c(u - U_N)||^2 \\
&+ ||U_N - u_N||^2. \qquad (6.89)
\end{aligned}$$

On the right-hand side of Equation 6.89, the third and forth terms are to be estimated. Firstly,

$$\begin{aligned}
||c(u - U_N)||^2 \lesssim& ||(1 + \boldsymbol{x}^{2\boldsymbol{\gamma}})(u - U_N)||^2 \lesssim ||u - U_N||^2 + ||\boldsymbol{x}^{2\boldsymbol{\gamma}}(u - U_N)||^2 \\
\lesssim& ||u - U_N||^2 + \left(\prod_{i=1}^{d} \alpha_i^{-4\gamma_i} \right) |\boldsymbol{\gamma}|_{\text{mix}} \cdot ||u - U_N||^2_{\mathcal{W}^{|\boldsymbol{\gamma}|_1}_{\boldsymbol{\alpha},\boldsymbol{\beta}}(\mathbb{R}^d)}, \qquad (6.90)
\end{aligned}$$

by $(\boldsymbol{C_6})$ and Lemma 6.1. Similarly, from $(\boldsymbol{C_6})$ again, we deduce that

$$\begin{aligned}
||\boldsymbol{b}\cdot\nabla(u-U_N)||^2 \leq & \sum_{i=1}^{d}||b_i(\boldsymbol{x})\partial_{x_i}(u-U_N)||^2 \lesssim \sum_{i=1}^{d}\left|\left|(1+\boldsymbol{x}^{2\boldsymbol{\delta}})\partial_{x_i}(u-U_N)\right|\right|^2 \\
\leq & \sum_{i=1}^{d}||\partial_{x_i}(u-U_N)||^2 + \sum_{i=1}^{d}\left|\left|\boldsymbol{x}^{2\boldsymbol{\delta}}\partial_{x_i}(u-U_N)\right|\right|^2 \\
\lesssim & ||u-U_N||^2_{\mathcal{W}^1_{\boldsymbol{\alpha},\boldsymbol{\beta}}(\mathbb{R}^d)} + \sum_{i=1}^{d}\left(\prod_{i=1}^{d}\alpha_i^{-4\delta_i}\right)|\boldsymbol{\delta}+\boldsymbol{e}_i|_{\text{mix}}\cdot||u-U_N||^2_{\mathcal{W}^{|\boldsymbol{\delta}|_1+1}_{\boldsymbol{\alpha},\boldsymbol{\beta}}(\mathbb{R}^d)} \\
\lesssim & ||u-U_N||^2_{\mathcal{W}^1_{\boldsymbol{\alpha},\boldsymbol{\beta}}(\mathbb{R}^d)} + d\left(\prod_{i=1}^{d}\alpha_i^{-4\delta_i}\right)|\boldsymbol{\delta}+1|_{\text{mix}}\cdot||u-U_N||^2_{\mathcal{W}^{|\boldsymbol{\delta}|_1+1}_{\boldsymbol{\alpha},\boldsymbol{\beta}}(\mathbb{R}^d)}.
\end{aligned} \tag{6.91}$$

Combining Equation 6.87 - Equation 6.89, we have

$$\begin{aligned}
\partial_t||u_N-U_N||^2 \lesssim & ||\nabla(u-U_N)||^2 + ||\boldsymbol{b}\cdot\nabla(u-U_N)||^2 + ||c(u-U_N)||^2 + C||u_N-U_N||^2 \\
\lesssim & ||\nabla(u-U_N)||^2 + C||u_N-U_N||^2 + ||u-U_N||^2_{\mathcal{W}^1_{\boldsymbol{\alpha},\boldsymbol{\beta}}(\mathbb{R}^d)} \\
& + ||u-U_N||^2_{\mathcal{W}^{|\boldsymbol{\delta}|_1+1}_{\boldsymbol{\alpha},\boldsymbol{\beta}}(\mathbb{R}^d)} + ||u-U_N||^2_{\mathcal{W}^{|\boldsymbol{\gamma}|_1}_{\boldsymbol{\alpha},\boldsymbol{\beta}}(\mathbb{R}^d)} \\
\lesssim & C||u_N-U_N||^2 + N^{\max\{|\boldsymbol{\gamma}|_1,|\boldsymbol{\delta}|_1+1\}-m}|u|^2_{\mathcal{K}^m_{\boldsymbol{\alpha},\boldsymbol{\beta}}(\mathbb{R}^d)},
\end{aligned}$$

by Equation 6.90, Equation 6.91 and Corollary 6.1 or Corollary 6.2. Hence,

$$\begin{aligned}
||u_N-U_N||^2(t) \leq & e^{Ct}||u_N-U_N||^2(0) + N^{\max\{|\boldsymbol{\gamma}|_1,|\boldsymbol{\delta}|_1+1\}-m}e^{Ct}\int_0^t e^{-Cs}|u|^2_{\mathcal{K}^m_{\boldsymbol{\alpha},\boldsymbol{\beta}}(\mathbb{R}^d)}(s)ds \\
\leq & N^{\max\{|\boldsymbol{\gamma}|_1,|\boldsymbol{\delta}|_1+1\}-m}\int_0^t e^{C(t-s)}|u|^2_{\mathcal{K}^m_{\boldsymbol{\alpha},\boldsymbol{\beta}}(\mathbb{R}^d)}(s)ds.
\end{aligned}$$

dim	2	3	4	5
# of indices in RHC	176	712	2485	7922
# of indices in OHC ($\gamma = 0.5$)	136	440	1264	3392

TABLE II

THE NUMBER OF INDICES FOR $N = 31$ WITH DIMENSION RANGING FROM 2 TO 5.

Therefore,

$$\begin{aligned}||u - u_N||^2(t) \leq & ||u - U_N||^2(t) + ||u_N - U_N||^2(t) \\ \lesssim & N^{1-m}|u|^2_{\mathcal{K}^m_{\boldsymbol{\alpha},\boldsymbol{\beta}}(\mathbb{R}^d)}(t) + N^{\max\{|\boldsymbol{\gamma}|_1,|\boldsymbol{\delta}|_1+1\}-m}\int_0^t e^{C(t-s)}|u|^2_{\mathcal{K}^m_{\boldsymbol{\alpha},\boldsymbol{\beta}}(\mathbb{R}^d)}(s)ds \\ \lesssim & N^{\max\{|\boldsymbol{\gamma}|_1,|\boldsymbol{\delta}|_1+1\}-m}\int_0^T |u|^2_{\mathcal{K}^m_{\boldsymbol{\alpha},\boldsymbol{\beta}}(\mathbb{R}^d)}(s)ds.\end{aligned}$$

The desired result is obtained. □

6.3 Numerical results

6.3.1 HC approximations with Hermite functions

In Figure 8, we display the indices of RHC and OHC (with $\gamma = 0.5$) in dimension 2 with $N = 31$. It is clear to see that the indices of OHC is a subset of RHC. Furthermore, we list in Table II the number of indices for $N = 31$ with dimension ranging from 2 to 5.

It is well-known that the abscissas of Hermite polynomials are non-nested, except the origin. It will lead to a larger number of points than those nested quadratures, such as Chebyshev

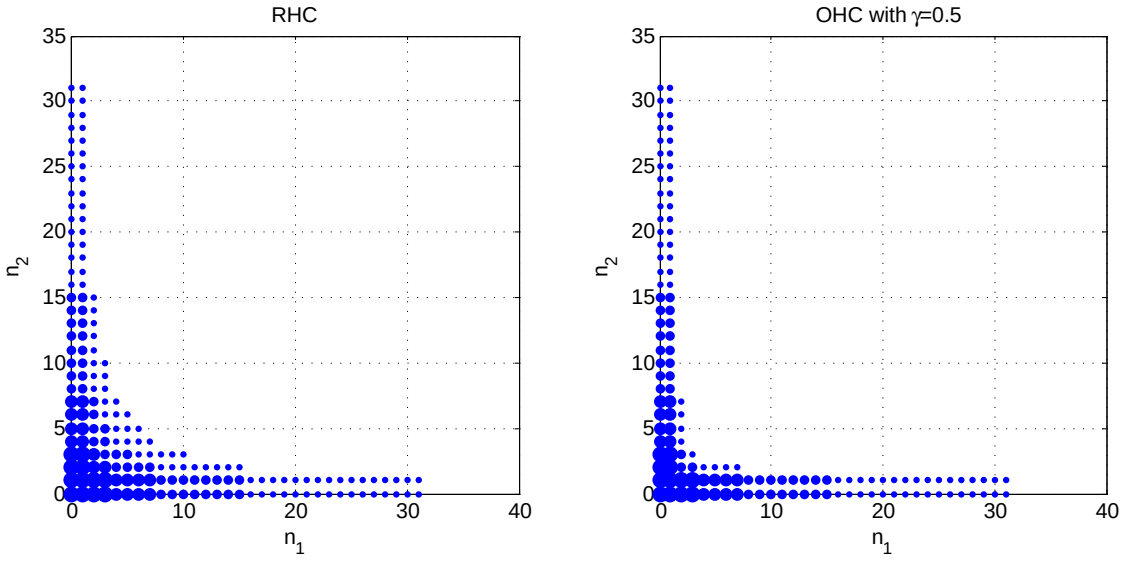

Figure 8. For $d = 2$, $N = 31$. Left: the index set $\mathbf{\Omega}_N$ of RHC. Right: the index set $\mathbf{\Omega}_{N,\gamma}$ of OHC with $\gamma = 0.5$.

polynomials. However, the number is still dramatically reduced, compared to the full grids. We list in Table III the abscissas of RHC, OHC and full grid for $N = 31$ with the dimension ranging from 2 to 4. It is clear that the abscissas in RHC/OHC are much fewer than thoses in the full grid.

6.3.2 HSM with sparse grid

Although the HC approximation is theoretically feasible, it is not suitable for practical implementations, due to the unclear "combining effecting" of the product rules, i.e., how to determine the weights from different combinations of 1-D Gauss-Hermite quadratures. Thus,

dim	2	3	4
# of abscissas in OHC ($\gamma = 0.5$)	108	3348	28944
# of abscissas in RHC	298	6612	82704
# of abscissas in full grid	961	29791	923521

TABLE III

THE NUMBER OF ABSCISSAS OF RHC, OHC AND FULL GRID OF $N = 31$ WITH THE DIMENSION RANGING FROM 2 TO 4.

in this subsection, we use the Smolyak's algorithm (54) to test the accuracy of high-dimensional HSM applied to linear parabolic PDEs.

Let us recall that Smolyak's algorithm is given by

$$\mathcal{I}(L,d) = \sum_{L-d+1 \le |\boldsymbol{i}|_1 \le L} (-1)^{L-|\boldsymbol{i}|_1} \binom{d-1}{L-|\boldsymbol{i}|_1} (\mathcal{U}^{i_1} \otimes \cdots \otimes \mathcal{U}^{i_d}),$$

where $\mathcal{U}^i$ is an indexed family of 1D quadrature, i is the 1D level; $\boldsymbol{i} = (i_1, \cdots, i_d)$ is the level vector, L is the max level. The sparse grid is formed by weighted combinations of those product rules whose product level $|\boldsymbol{i}|_1$ falls between $L - d + 1$ and L.

In Figure 9, we display the abscissas of the Hermite functions and the index set with level L ranging from 2 to 4 in $d = 2$.

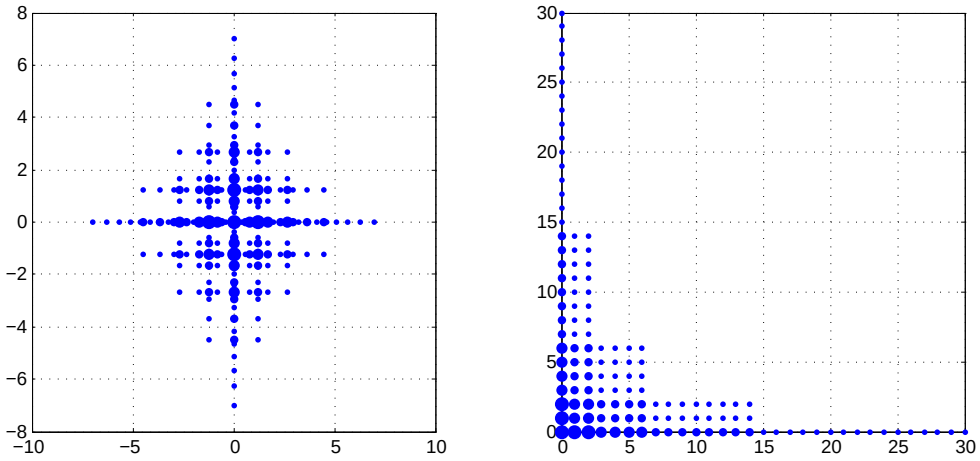

Figure 9. In $d = 2$, level L ranging from 2 to 4. Left: the abscissas of Hermite functions. Right: the indices in the index set. The larger the dot is, the lower the level.

Let us test the accuracy with the following linear parabolic PDE

$$\begin{cases} \partial_t u = \triangle u - \sum_{i=1}^{d} x_i^2 u + f(\boldsymbol{x}, t) \\ u(\boldsymbol{x}, 0) = \left(\sum_{i=1}^{d} x_i \right) e^{-\frac{1}{2}(x_1^2 + \cdots + x_d^2)} \end{cases},$$

where $\triangle$ is the Laplacian operator, and

$$f(\boldsymbol{x}, t) = \left[\cos t + d \sin t + (d+2) \sum_{i=1}^{d} x_i \right] e^{-\frac{1}{2}(x_1^2 + \cdots x_d^2)}.$$

By direct computations, the exact solution to this PDE is

$$u(\boldsymbol{x},t)=\left(\sum_{i=1}^{d}x_i+\sin t\right)e^{-\frac{1}{2}(x_1^2+\cdots+x_d^2)}.$$

It is known from (39) that the best scaling factor is $\boldsymbol{\alpha}=\mathbf{1}$ in this case, since the first two Hermite functions will resolve the exact solution perfectly only with the round-off errors (around 10^{-16} on my computer). To make the convergence rate observable with respect to the level L, we shall choose the scaling factor $\boldsymbol{\alpha}$ to be $1.01\times\mathbf{1}$.

The corresponding spectral scheme (cf. Equation 6.77, Equation 6.78) is as follows:

$$\begin{cases}\langle\partial_t u_N(t),\varphi\rangle=-\langle\nabla u_N,\nabla\varphi\rangle-\displaystyle\sum_{i=1}^{d}\langle x_i^2u_N,\varphi\rangle+\langle f,\varphi\rangle\\[2ex] u_N(0)=P_Nu_0,\end{cases}\tag{6.92}$$

for all $\varphi\in X_N$. Here, we choose $X_N=X_N^{\alpha,\beta}=\operatorname{span}\{\mathcal{H}_{\boldsymbol{n}}^{\boldsymbol{\alpha},\boldsymbol{\beta}}:\boldsymbol{\Omega}_N\}$, from Smolyak (54). Thus, we can write the numerical solution as

$$u_N(\boldsymbol{x},t)=\sum_{\boldsymbol{n}\in\boldsymbol{\Omega}_N}a_{\boldsymbol{n}}(t)\mathcal{H}_{\boldsymbol{n}}^{\boldsymbol{\alpha},\boldsymbol{\beta}}(\boldsymbol{x}),$$

Taking $\varphi(\boldsymbol{x}) = \mathcal{H}_{\boldsymbol{n}}^{\boldsymbol{\alpha},\boldsymbol{\beta}}(\boldsymbol{x})$ in Equation 6.92, and due to Equation 6.5, Equation 6.4 and Equation 6.13, we arrive at the ODEs

$$\begin{cases} \dfrac{d}{dt} a_{\boldsymbol{n}} = A a_{\boldsymbol{n}} + \hat{f}_{\boldsymbol{n}} \\ a_{\boldsymbol{n}}(0) = (\hat{u}_0)_{\boldsymbol{n}} \,. \end{cases} \tag{6.93}$$

Here $\hat{f}_{\boldsymbol{n}}$ (resp. $(\hat{u}_0)_{\boldsymbol{n}}$) is the Hermite coefficients of f (resp. u_0) and the matrix A comes from the Laplacian operator and the potential. In Figure 10, we display the nonzero entries of the matrix A for dimension 3 and 4 with level= 4.

We use a central difference scheme to solve Equation 6.93 with $T = 0.1$, $dt = 10^{-5}$, $\boldsymbol{\alpha} = 1.01 \times \mathbf{1}$ and $\boldsymbol{\beta} = \mathbf{0}$. Figure 11 shows the L^2-norm of $(u_N - u_{\text{exact}})$ with respect to the level in dimensions ranging from 2 to 4. It is exactly what we expect, as in the semi-log plot the error goes down almost along a straight line, which indicates that the convergence rate is nearly exponential. However, when the dimension grows, the error becomes slightly larger. This reveals that the convergence rate still slightly deteriorates with increasing dimension.

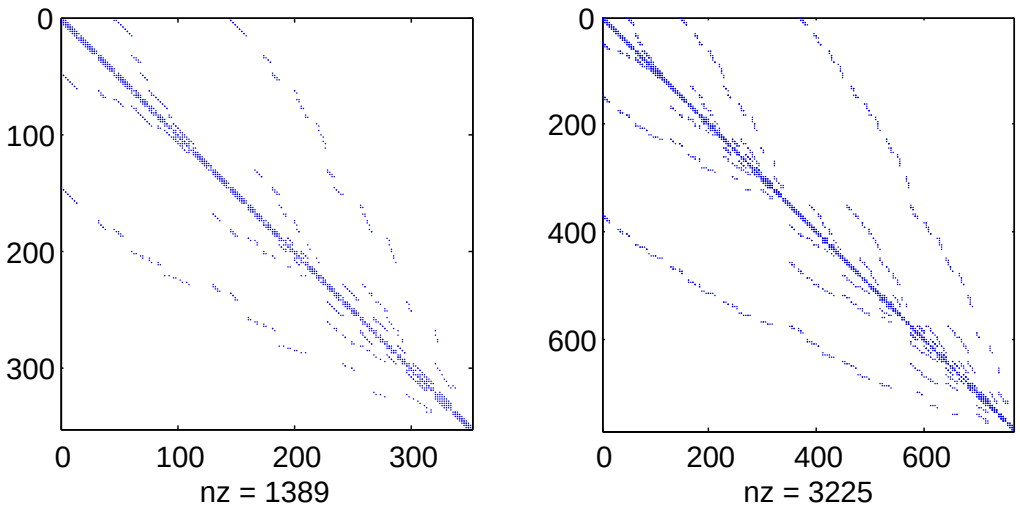

Figure 10. The nonzero entries in the matrix A (cf. Equation 6.93) are displayed with level= 4. Left: $d = 3$, Right: $d = 4$.

level/dim	2	3	4
2	2.24E-03	7.99E-03	n/a
3	3.99E-04	544E-03	2.10E-02
4	4.75E-06	1.93E-03	1.14E-02
5	2.72E-07	2.66E-04	4.11E-03

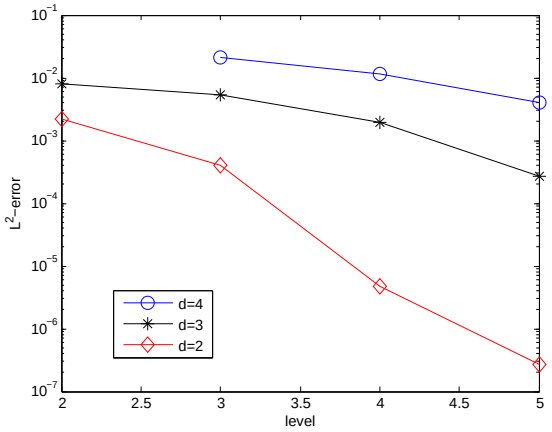

Figure 11. The L^2 error of u_N with respect to the level in $d = 2, 3$ and 4 is drawn.

CITED LITERATURE

1. Arulampalam, M. S., Maskel, S., Gordon, N. and Clapp, T.: A tutorial on particle filters for online nonlinear/non-Gaussian Bayesian tracking. IEEE Trans. Signal Process., 50(2):174–188, 2002.

2. Bain, A. and Crisan, D.: Fundamentals of Stochastic Filtering. Stochastic Modelling and Applied Probability, Vol. 60, Springer, 2009.

3. Baras, J., Blankenship, G. L. and Hopkins, W.: Existence, uniqueness, and asymptotic behavior of solutions to a class of Zakai equations with unbounded coefficients. IEEE Trans. Automat. Control, 28(2):203-214, 1983.

4. Barthelmann, V., Novak, E. and Ritter, K.: High dimensional polynomial interpolation on sparse grids. Adv. Comput. Math., 12:273-288, 2000.

5. Besala, P.: On the existence of a fundamental solution for a parabolic differential equation with unbounded coefficients. Ann. Polonici Math., 29:403-409, 1975.

6. Boyd, J.: The rate of convergence of Hermite function series. Math. Comp., 35:1039-1316, 1980.

7. Boyd, J.: Asymptotic coefficients of Hermite function series. J. Comput. Phys., 54:382-410, 1984.

8. Boyd, J.: Chebyshev and Fourier Spectral Methods. 2d. edition, Dover, New York, 2001

9. Bensoussan, A.: Some existence results for stochastic partial differential equations, in *Stochastic Partial Differential Equations and Applications.* Pitman Res. Notes Math., vol. 268, Longman Scientific and Technical, Harlow, UK, 1992, pp. 37-53.

10. Bensoussan, A., Glowinski, R. and Rascanu, A.: Approximation of the Zakai equation by the splitting up method. SIAM J. Control Optim., 28:1420-1431, 1990.

11. Bensoussan, A., Glowinski, R. and Rascanu, A.: Approximation of some stochastic differential equations by the splitting up methods. Appl. Math. Optim., 25:81-106, 1992.

12. Burgartz, H.-J. and Griebel, M.: A note on the complexity of solving Poisson's equation for spaces of bounded mixed derivatives. J. Complexity, 15:167-199, 1999.

13. Burgartz, H.-J. and Griebel, M.: Sparse grids. Acta Numer., 13:147-269, 2004.

14. Burkardt, J.: The Sparse Grid Interpolant. http://people.sc.fsu.edu/~jburkardt/presentations /sparse_interpolant.pdf, 2012.

15. Duncan, T.: Probability density for diffusion processes with applications to nonlinear filtering theory. Ph. D. thesis, Stanford University, Stanford, CA, 1967.

16. Fleming, W. and Mitter, S.: Optimal control and nonlinear filtering for nondegenerate diffusion processes. Stochastics, 8:63-77, 1982.

17. Fok, J., Guo, B. and Tang, T.: Combined Hermite spectral-finite difference method for the Fokker-Planck equation. Math. Comp., 71(240):1497-1528, 2001.

18. Friedman, A.: Partial differential equations of parabolic type. Prentice-Hall, Englewood Cliffs, NJ, 1964.

19. Funaro, D. and Kavian, O.: Approximation of some diffusion evolution equation in unbounded domains by Hermite function. Math. Comp., 37:597-619, 1991.

20. Gerstner, T. and Griebel, M.: Numerical integration using sparse grids. Numer. Algorithms, 18:209-232, 1998.

21. Gottlieb, D. and Orszag, S.: Numerical analysis of spectral methods: theory and applications. Soc. In. and Appl. Math., Philadelphia, 1977.

22. Gradinaru, V.: Fourier transform on sparse grids: Code design and the time dependent Shrödinger equation. Computing, 80:1-22, 2007.

23. Griebel, M.: Adaptive sparse grid multilevel methods for elliptic PDEs based on finite differences. Computing, 61(2):151-179, 1998.

24. Griebel, M. and Hamakekers, J.: Sparse grid for the Schrödinger equation. M2AN Math. Model. Numer. Anal., 41:215-247, 2007.

25. Griebel, M. and Oeltz, D.: A sparse grid space-time discretization scheme for parabolic problems. Computing, 81(1):1-34, 2007.

26. Griebel, M., Oswald, P. and Schiekofer, T.: Sparse grids for boundary integral equations. Numer. Mathematik 83(2):279-312, 1999.

27. Guo, B.-Y., Shen, J. and Xu, C.-L.: Spectral and pseudospectral approximation using Hermite function: Application in Dirac equation. Adv. Comput. Math., 19:35-55, 2003.

28. Gyongy, I. and Krylov, N.: On the splitting-up method and stochastic partial differential equation. Ann. Probab., 31:564-591, 2003.

29. Hemker, P.: Sparse-grid finite-volume multigrid for 3D-problems. Adv. Comput. Math., 4:83-110, 1995.

30. Hopkins, W. E., Jr.: Nonlinear filtering of nondegnerate diffusions with unbounded coefficients. Ph. D. thesis, Dept. Elec. Eng., Univ. Maryland, College Park, Nov. 1982.

31. Ito, K.: Approximation of the Zakai equation for nonlinear filtering. SIAM J. Control Optim., 34:620-634, 1996.

32. Ito, K. and Rozovskii, B.: Approximation of the Kushner equation for nonlinear filtering. SIAM J. Control Optim., 38:893-915, 2000.

33. Kalman, R. E.: A new approach to linear filtering and prediction problems. ASME Trans., J. Basic Eng., ser. D., 82:35-45, 1960.

34. Kalman, R. E. and Bucy, R. S.: New results in linear prediction and filtering theory. ASME Trans., J. Basic Eng., ser. D., 83:95-108, 1961.

35. Klimke, A. and Wohlmuth, B.: Algorithm 847: Spinterp: piecewise multilinear hierarchical sparse grid interpolation in MATLAB. ACM Trans. Math. Softw., 31(4):561-579, 2005.

36. Knapek, S.: Hyperbolic cross approximation of integral operators with smooth kernel. Tech. Report 665, SFB 256, Univ. Bonn, 2000.

37. Le Gland, F.: Splitting-up approximation for SPDEs and SDEs with application to nonlinear filtering. Lecture Notes in Control and Inform. Sci., Vol. 176, Springer, New York, 1992, pp. 177-187.

38. Lototsky, S., Mikulevicius, R. and Rozovskii, B.: Nonlinear filtering revisited: a spectral approach. SIAM J. Control Optim., 35(2):435-461, 1997.

39. Luo, X. and Yau, S. S.-T.: Complete real time solution of the general nonlinear filtering problem without memory. accepted for publication by IEEE Trans. Automat. Control., 2013. arXiv:1208.0962

40. Luo, X. and Yau, S. S.-T.: Hermite spectral method to 1D forward Kolmogorov equation and its application to nonlinear filtering problems. accepted for publication by IEEE Trans. Automat. Control., 2013 arXiv:1301.1403.

41. Mortensen, R.: Optimal control of continuous time stochastic systems. Ph. D. dissertation, University of California, Berkeley, CA, 1966.

42. Nagase, N.: Remarks on nonlinear stochastic partial differential equations: An application of the splitting-up method. SIAM J. Control Optim., 33:1716-1730, 1995.

43. Nobile, F., Tempone, R. and Webster, C.: A sparse grid stochastic collocation method for partial differential equations with random input data. SIAM J. Numer. Anal., 46(5):2309-2345, 2008.

44. Pardoux, E.: Stochastic partial differential equations and filtering of diffusion processes. Stochastics, 3:127-167, 1979.

45. Pardoux, E.: Equations du filtrage nonlinéaire, de la prédiction et du lissage. Publication de mathématique Appliquées, 80(6), Univérsite de Provence.

46. Rao, C.: Nonlinear filtering and evolution equations: fast algorithms with applications to target tracking. Ph. D. thesis, Univ. Southern California, Los Angeles, CA, 1998.

47. Rozovsky, B.: Stochastic partial differential equations arising in nonlinear filtering problems. Usp. Mat. Nauk., 27:213-214, 1972.

48. Schumer, J. W. and Holloway, J. P.: Vlasov simulations using velocity-scaled Hermite representations. J. Comp. Phys, 144:626-661, 1998.

49. Schwab, C. and Stevenson, R.: Adaptive wavelet algorithms for elliptic PDE's on product domains. Math. Comp., 77:71-92, 2008.

50. Schwab, C. and Todor, R.: Sparse finite elements for elliptic problems with stochastic loading. Numer. Math., 95(4):707-734, 2003.

51. Shen, J. and Wang, L.-L.: Sparse spectral approximations of high-dimensional problems based on hyperbolic cross. SIAM J. Numer. Anal., 48(3):1087-1109, 2010.

52. Shen, J. and Yu, H.: Efficient spectral sparse grid methods and applications to high-dimensional elliptic problems. SIAM J. Sci. Comput., 32(6):3228-3250, 2010.

53. Smolyak, S. A.: Quadrature and interpolation formulas for tensor products of certain classes of functions. Dokl. Akad. Nauk SSSR, 4:240-243, 1963.

54. Sobolev, S. L.: Applications of functional analysis in mathematical physics. Tran. Math. Monographs, vol. 7, AMS, Providence, RI, 1963.

55. Tang, T.: The Hermite spectral method for Gaussian-type functions. SIAM J. Sci. Comput., 14(3):594606, 1993.

56. von Petersdorff, T. and Schwab, C.: Numerical solution of parabolic equations in high dimensions. M2AN Math. Model. Numer. Anal., 38(1):93127, 2004.

57. Xiang, X.-M. and Wang, Z.-Q.: Generalized Hermite spectral method and its applications to problems in unbounded domains. SIAM J. Numer. Anal., 48(4):1231-1253, 2010.

58. Yau, S. and Yau, S. S.-T.: Existence and uniqueness and decay estimates for the time dependent parabolic equation with application to Duncan-Mortensen-Zakai equation. Asian J. Math., 2:1079-1149, 1998.

59. Yau, S.-T. and Yau, S. S.-T.: Real time solution of nonlinear filtering problem without memory II. SIAM J. Control Optim., 47(1):230-243, 2008.

60. Zakai, M.: On the optimal filtering of diffusion processes. Z. Wahrscheinlichkeitstheorie und Verw. Gebiete, 11:230-243, 1969.

61. Zenger, C.: Sparse grids in Parallel Algorithms for Partial Differential Equations. Notes Numer. Fluid Mech. 31, W. Hackbusch, ed., Vieweg, Braunschweig, 241-251, 1991.

VITA

PERSONAL DETAILS

Birth March, 1982, Chengdu, Sichuan, P. R. China

Gender Female

Address Department of Mathematics, Statistics and Computer Science,
University of Illinois at Chicago,
851 S. Morgan Street, SEO 322
Chicago, IL, 60607, USA

Email xluo6@uic.edu

Homepage http://www.math.uic.edu/∼xluo

EDUCATION

Ph. D. in Applied Mathematics 2010-2013(expected)

University of Illinois at Chicago (UIC)

M. S. in Applied Mathematics 2010-2011

UIC

Visiting scholar in Mathematics 2009-2010

UIC

Graduate student research fellow in Mathematics 2008-2009

University of Connecticut

B. S. in Mathematics 2000-2004

East China Normal University (ECNU)

RESEARCH INTERESTS

Analysis of partial differential equations; Nonlinear filtering theory; Numerical analysis of spectral methods; Sparse grid algorithms; Fluid mechanics.

HONORS AND AWARDS

2013 Student Presenter Award, UIC

2013 Chancellor's Graduate Research Fellowship, UIC

2012 GSC Travel Award, UIC

2012 LAS PhD Student Travel Award, UIC

2011 Victor Twersky Memorial Scholarship, UIC

2008 One-year fellowship, China Scholarship Council (CSC)

2005 Excellent Graduate Student Scholarship, ECNU

INVITED TALKS

Dec, 2012 *51st IEEE Conference on Decision and Control*, Hawaii, US.

CONFERENCES AND WORKSHOPS

Nov, 2012 *Madison Autumn Analysis and PDE Workshop*, University of Wisconsin-Madison

Apr, 2012 *69th Midwest Partial Differential Equations Seminar*, UIC

Nov, 2010 *66th Midwest Partial Differential Equations Seminar*, UIC

Jul, 2008 *The 5th East China Partial Differential Equations*, Nanjing University, China

Jun, 2007 *The 4th East China Partial Differential Equation*, Yantai University, China

PUBLICATIONS AND PREPRINTS

1. *Hermite Spetral Method with Hyperbolic Cross Approximations to High-Dimensional Parabolic PDEs* (with S. S.-T. Yau), 23 pp. in ms., accepted for publication by SIAM J. Numer. Anal., 2013. arXiv:1306.3207
2. *Hermite Spectral Method to 1D Forward Kolmogorov Equation and its Application to Nonlinear Filtering Problems* (with S. S.-T. Yau), 13 pp. in ms., accepted for publication by IEEE Trans. Automat. Control., 2013. arXiv:1301.1403
3. *A Sharp Estimate of Dickman-De Bruijin Function and a Sharp Polynomial Estimate of Positive Integral Points in 4-dimensional Tetrahedron* (with S. S.-T. Yau and H. Zuo), 20 pp. in ms., accepted for publication by Math. Nachr., 2013.
4. *Complete Real Time Solution of the General Nonlinear Filtering Problem without Memory* (with S. S.-T. Yau),15 pp. in ms., accepted for publication by IEEE Trans. Automat. Control., 2013. arXiv:1208.0962

5. *On Number Theoretic Conjecture of Positive Integral Points in 5-Dimension Tetrahedron and a Sharp Estimate of Dickman-De Bruijn Function* (with K.-P. Lin, S. S.-T. Yau and H. Zuo), 33 pp. in ms., accepted for publication by J. Eur. Math. Soc., 2013.

6. *A Novel Algorithm to Solve the Robust DMZ Equation in Real Time* (with S. S.-T. Yau), Proceedings of the 51st IEEE Conference on Decision and Control, Dec 2012, pp. 606-611.

7. *Regularity of the Extremal Solution for Some Elliptic Problems with Singular Nonlinearity and Advection* (with D. Ye and F. Zhou), J. Differential Equations, Vol. 251, No. 8, 2011, pp. 2082-2099. arXiv:1004.3956

8. *Uniqueness of Weak Extremal Solution to Biharmonic Equation with Logarithmically Convex Nonlinearitie*, J. Partial Differential Equations, Vol. 23, No. 4, 2010, pp. 315-329.

9. *Parameters of Two Special Toric Surface Codes* (with P. Zhang and H. Zuo), Chinese Ann. Math. Ser. A 31, No. 5, 2010, pp. 517-524.

10. *Asymptotic Behavior of Oscillating Radial Solutions to Certain Nonlinear Equations, Part II* (with C. Gui and F. Zhou), Methods and Applications of Analysis, Vol. 16, No. 4, 2009, pp. 459-468.

11. *New suboptimal filter for polynomial filtering problems*, (with Y. Jiao, S. S.-T. Yau and W.-L. Chiou), submitted to IEEE Trans. Automat. Control., 2013.

12. *The quenching behavior of MEMS with fringing field*, in preparation.

13. *Homogeneous solutions for the Euler equation of ideal fluid*, (with R. Shvydkoy), in preparation.

14. *The Dynamics of a Kinetic Activator-Inhibitor System with Positive Production Rate*(with W.-M. Ni and X. Xiang), 25 pp. in ms., preprint.

15. *On Classification of Toric Surface Codes of Low Dimension*(with S. S.-T. Yau and H. Zuo), 20 pp. in ms., preprint.

PROFESSIONAL ACTIVITIES

Member of AMS, SIAM.

Referee for IEEE Trans. Automat. Control., Internat. J. Systems Sci.

EXPERIENCES

Research assistant, UIC 2012-2013

– Supported by Prof. Roman Shvydkoy, NSF grant DMS-1210896

Graduate teaching assistant, UIC 2010-2012

– Holding the discussion and recitation sections for *Finite Mathematics for Business*, *Precalculus*, *Calculus III* and *Calculus for Business*.

– Grader for *Introduction to Proofs*.

SKILLS

Languages **Mandarin** (mother tongue)

English (fluent)

French, Japanese (capable of reading academic papers)

Software MATLAB, LaTeX, HTML, C

REFERENCES

Stephen S.-T. Yau *University of Illinois at Chicago, USA and Tsinghua University, P. R. China*

yau@uic.edu, +86 (10) 62787874

Roman Shvydkoy *University of Illinois at Chicago, USA*

shvydkoy@uic.edu, +1 (312) 413-2967

Dong Ye *Universite Paul Verlaine-Metz, France*

dong.ye@univ-metz.fr, +33 (0)3 87 54 72 90

Changfeng Gui *University of Connecticut, USA*

gui@math.uconn.edu, +1 (860) 486-3203

Feng Zhou *East China Normal University, P. R. China*

fzhou@math.ecnu.edu.cn, +86 (21) 62237325