The Dissertation Committee for Jiachuan He
certifies that this is the approved version of the following dissertation:

Data-driven uncertainty quantification for predictive subsurface flow and transport modeling

Committee:

Clint Dawson, Supervisor

Tan Bui-Thanh

Omar Ghattas

Chad Landis

Data-driven uncertainty quantification for predictive subsurface flow and transport modeling

by

Jiachuan He

DISSERTATION

Presented to the Faculty of the Graduate School of

The University of Texas at Austin

in Partial Fulfillment

of the Requirements

for the Degree of

DOCTOR OF PHILOSOPHY

THE UNIVERSITY OF TEXAS AT AUSTIN

December 2018

Dedicated to my parents.

Acknowledgments

First and foremost, I want to express my deepest appreciation to my advisor, Prof. Clint N. Dawson, for his unremitting guidance, unwavering support and continuous encouragement over the past years. He offered me great opportunities to broaden my knowledge in engineering science and to purse the research that interested me. This work benefits from his profound insights in the areas of subsurface flow modeling and uncertainty quantification. I truly enjoyed the individual weekly discussions with him which helped me develop my research topics. I am grateful for the invaluable encouragement and help he gave me to complete the work near the end of my study.

I would like to thank Professors Tan Bui-Thanh, Omar Ghattas, and Chad Landis for serving my dissertation committee. Reading Professor Tan Bui-Thanh's notes, "A Gentle Tutorial on Statistical Inversion using the Bayesian Paradigm", and taking Professor Omar Ghattas' Computational and Variational Methods for Inverse Problem class helped me broaden and deepen my knowledge in various aspects of solving inverse problems and quantifying uncertainty.

I would also like to thank Dr. Steven Mattis and Professor Troy Butler for their guidance and discussions on measure-theoretic stochastic inverse problems. I wish to thank Dr. Velimir Vesselinov, my mentor during my summer internship at Los Alamos National Laboratory, for sparking my interest in applying machine learning methods to solve subsurface flow problems.

I am thankful for every member in the Computational Hydraulics Group. My thanks are extended to my friends in Austin, in particular, Jing Ping, Ju Liu, Wei Du, Peng Wang, Yimin Zhong, Hailong Xiao, Bin Wang and Zhen Tao.

I am indebted to my parents for always believing in me. Nothing is possible without their unconditional love.

Data-driven uncertainty quantification for predictive subsurface flow and transport modeling

Publication No. ____________________

Jiachuan He, Ph.D.
The University of Texas at Austin, 2018

Supervisor: Clint Dawson

Specification of hydraulic conductivity as a model parameter in groundwater flow and transport equations is an essential step in predictive simulations. It is often infeasible in practice to characterize this model parameter at all points in space due to complex hydrogeological environments leading to significant parameter uncertainties. Quantifying these uncertainties requires the formulation and solution of an inverse problem using data corresponding to observable model responses. Several types of inverse problems may be formulated under various physical and statistical assumptions on the model parameters, model response, and the data. Solutions to most types of inverse problems require large numbers of model evaluations. In this study, we incorporate the use of surrogate models based on support vector machines to increase the number of samples used in approximating a solution to an inverse problem at a relatively low computational cost. To test the global capabilities of this type of surrogate model for quantifying uncertainties, we use a framework for constructing pullback and push-forward probability measures to study the data-to-parameter-to-prediction propagation of uncertainties under minimal statistical assumptions. Additionally, we demonstrate that it is possible to build a support vector machine using relatively low-dimensional representations of the hydraulic conductivity to propagate distributions. The numerical examples further demonstrate that we can make reliable probabilistic predictions of contaminant concentration at spatial locations corresponding to data not used in the solution to the inverse problem.

This dissertation is based on the article entitled *Data-driven uncertainty quantification for predictive flow and transport modeling using support vector machines* by Jiachuan He, Steven Mattis, Troy Butler and Clint Dawson [32]. This material is based upon work supported by the U.S. Department of Energy Office of Science, Office of Advanced Scientific Computing Research, Applied Mathematics program under Award Number DE-SC0009286 as part of the DiaMonD Multifaceted Mathematics Integrated Capability Center.

Table of Contents

List of Tables

List of Figures

Chapter 1

Introduction[1]

1.1 Motivation

In the past century, demand for clean groundwater has soared due to population growth and pollution of surface water. In some areas of the world, groundwater has become the main drinking water supply or even the sole source of water. Unfortunately, contrary to the popular impression that pumping groundwater from wells and spring water is untainted, we find contamination of aquifers and groundwater a serious problem in many parts of the world [21, 50, 60].

Several years ago, a hexavalent chromium plume was present above the New Mexico groundwater standard of 50 parts per billion in 4 monitoring wells in the regional aquifer beneath Los Alamos National Laboratory (LANL). There is an urgent need for migration control of the chromium plume and best cleanup method assessment. Many theoretical and computational frameworks have been developed to model subsurface flow and contaminant transport [4, 22, 63, 65]. A lot of research also has been done to advance critical decision-

[1] This chapter is based on the article entitled *Data-driven uncertainty quantification for predictive flow and transport modeling using support vector machines* by Jiachuan He, Steven Mattis, Troy Butler and Clint Dawson [32].

making related to remediation strategies with uncertainties in models [8, 31, 37, 54]. Overall, prediction and remediation of subsurface require us to solve a series of mathematical problems among which we first wish to estimate unknown parameter field, e.g., hydraulic conductivity, that characterizes a model of the system. In other words, given experimental or observable data, we need to solve an inverse problem.

1.2 Background

Mathematical models for groundwater contaminant transport simulation often contain parameters that cannot be directly measured. Instead, we must often infer parameter values by formulating and solving an inverse problem using data corresponding to observable model responses. However, a main theoretical difficulty is that most inverse problems are not well-posed in the sense of existence, uniqueness, and stability of the solution. Complicating matters further is the practical issue that solving inverse problems is often computationally intensive.

The high computational cost in solving an inverse problem may arise from many sources including the thousands or millions of forward simulations required, inverting large dense operators, or strong nonlinearities in the parameter-to-observable map even when the forward problem is linear. A number of recently developed methods have focused on constructing surrogate models for improved computational efficiency. A surrogate model can be regarded as a response surface approximation of the parameter-to-observables map defined by the composition of an observation operator with the solution operator to the model. There are many ways to construct surrogate models. For example, some popular surrogate models are based on Polynomial Chaos Expansions [46, 68], the Probabilistic Collocation Method [62, 71], Kriging [6], or Radial Basis Functions [57, 58], to name just a few. In this work, we incorporate the use of a Support Vector Machine (SVM) which has found a wide range of applications in the fields of classification and regression analysis [5, 30]. There are also a few applications of SVM in hydrology [1, 27, 38, 43, 69]. We apply SVM to approximate the parameter-to-observable map using sets of input-output pairs of sampled model parameters and simulated model observables, where the sampled model parameters define a parameterization of an unknown hydraulic conductivity field and model observables correspond to a sparse set of spatially sampled contaminant concentrations.

Using an SVM, like any surrogate, to quantify uncertainties, represents a trade-off in errors where stochastic sources of error (e.g., due to finite sampling) are reduced while deterministic sources of error (e.g., due to approximation errors) may be significantly increased. It is therefore important to study how accurately any surrogate can be used in quantifying uncertainties in both inverse and forward uncertainty quantification (UQ) problems. In general, we first solve inverse UQ problems to quantify uncertainties in model parameters, which are subsequently used to inform forward UQ problems to quantify uncertainties in model predictions. Thus, we focus first on the ability of the SVM to solve a data-to-parameter (i.e., inverse) UQ problem.

In the hydrology community, many inversion methods have been proposed and developed independently [72] for characterizing hydraulic conductivity K [19, 36, 52], which is often the most dominant hydraulic property. The earliest method is the so-called direct method, which is relatively straightforward and has been widely used [48, 49]. Assuming hydraulic head is known, one can substitute head into the forward problem and then solve the inverse problem in terms of a partial differential equation in K. However, this approach has two main shortcomings. First, this method requires the information of hydraulic head over the entire domain. Although values of head can be achieved through interpolation of observations, it inevitably introduces smoothing of the data and errors. Second, this method is unstable due to the ill-posedness of inverse problems that small errors in head may result in large changes in the solution. To overcome these problems, indirect methods were developed to handle limited numbers of observations. Optimal parameters are found by minimizing an objective function which includes a regularization term to ensure stability of the optimization problem. Recently, Bayesian inversion methods have gained popularity in hydrologic studies [7, 24, 41, 51, 64, 67]. This approach allows a flexible integration of prior knowledge about parameters into the solution. Such a probabilistic approach is often preferable in practical problems since its solution also quantifies the uncertainties in the reconstruction. However, such an approach requires additional statistical assumptions (e.g., the specification of a prior and the likelihood function, etc.), which may influence solutions, sometimes in undesirable ways [56]. Moreover, these approaches are generally focused on parameter estimation under uncertainty, and surrogate models generally need to be point-wise accurate only near a nominal parameter value (e.g., the maximum likelihood parameter) in order to obtain accurate posterior distributions, e.g., see [46].

In this work, we consider a general framework for constructing pullback and push-forward probability measures. Since constructing these measures requires global accuracy in the SVM, this serves as a robust test of the ability of an SVM to quantify uncertainties for other types of inverse problems. This framework is based upon the general measure-theoretic framework for the formulation and solution to stochastic inverse problems studied in [10]. The methodology has been successfully applied to a variety of UQ problems in storm surge modeling [28], subsurface contaminant transport [47], and structural damage of vibrating beams[14]. In this study, we specify a probability measure on the observable contaminant concentration data, and through global sampling of the parameter space, we construct a pullback probability measure on the parameters defining hydraulic conductivity. We can verify that a pullback measure was accurately computed by using the parameter-to-observable map to compute its push-forward measure and comparing it to the specified probability measure on these observables. Then, we may use this measure to construct other push-forward measures for quantities of interest (QoI) to be predicted by the model, e.g., contaminant levels at spatial locations not used in the construction of the pullback measure.

The rest of the dissertation is organized as follows. In Chapter 2 and Chapter 3, we describe the groundwater flow and contaminant transport model, and the parameterization of the hydraulic conductivity field which is the unknown parameter in the model. We provide a brief description of the fundamental principle of SVM for constructing surrogate models in Chapter 4 followed by details for constructing the SVM used in this particular work. The UQ framework for constructing pullback and push-forward probability measures is summarized in Chapter 5. We present numerical examples in Chapter 6 to demonstrate the effectiveness of the proposed methodology. Finally, some concluding remarks are provided in Chapter 7.

Chapter 2

Groundwater Flow Model

Groundwater often refers to the water held underground in soils or pores that are fully saturated. Since the natural subsurface system cannot be analyzed directly because of the complex hydrogeological environment, scientist and engineers often use models to describe it. In this chapter, we present the mathematical model that governs groundwater flow based on mass conservation and Darcy's law. To solve the model, hydraulic properties including specific storage and hydraulic conductivity, which can be highly variable sometimes, need to be assigned. However, in practice it's infeasible to have direct measurements of the whole hydraulic property field. Many techniques have been proposed which can be categorized into two main approaches: Empirical and Experimental. The empirical approach is based on the correlation between hydraulic conductivity and known soil properties from other studies. It calculates the hydraulic conductivity using empirical formulae such as Kozeny-Carman equation [2, 18], Hazen equation [17, 39], Breyer equation [53], etc. The experimental approach determines hydraulic conductivity through hydraulic experiments, e.g., laboratory tests and field tests. Having some sparse observations from the tests, we try to infer the hydraulic properties such that the mathematical model reproduces the observed behavior. In this work, assuming covariance functions characterizing the hydraulic conductivity field are obtained from measurements, we treat the conductivity field as a random function and decompose it with a Karhunen-Loève Expansion (KLE). Eigenvalues and eigenfunctions in KLE can be derived analytically in some special case or computed numerically more generally. We discretize and solve the groundwater flow model by a mixed finite element method. After solving the set of equations, the hydraulic heads and flow rates can be obtained and further coupled with transport models to study contaminant transport problems.

2.1 Mass Conservation

The law of conservation of mass states that for a saturated porous medium the net mass flow rate of fluid into a control volume along with sources or sinks inside is equal to the change in fluid mass storage for a given increment of time. The resulting continuity equation can be written as

$$\frac{\partial(\rho\phi)}{\partial t} = -\frac{\partial(\rho q_x)}{\partial x} - \frac{\partial(\rho q_y)}{\partial y} - \frac{\partial(\rho q_z)}{\partial z} + \rho g, \tag{2.1}$$

where q_x, q_y and q_z are components of flux $\boldsymbol{q}$ in three dimensions, ρ is density of fluid, ϕ is porosity, and g is sources or sinks. The left-hand side of Equation (2.1), $\frac{\partial(\rho\phi)}{\partial t}$, can be expanded as the sum of $\phi\frac{\partial\rho}{\partial t}$ and $\rho\frac{\partial\phi}{\partial t}$. These two terms represent the produced mass rate of fluid caused by a change in hydraulic head that leads to fluid density change and the porosity change of the porous medium, respectively. The first term is determined by the compressibility of the fluid while the second term is controlled by the compressibility of the porous media. To simplify $\phi\frac{\partial\rho}{\partial t}$ on the left of Equation (2.1), we define the specific storage, S_s, as the volume of fluid produced under unit decline in head due to the fact that both fluid density and porosity changes are caused by the change in hydraulic head. Therefore, the mass rate of fluid produced can be written as $\rho S_s\frac{\partial h}{\partial t}$, and Equation (2.1) becomes

$$\rho S_s\frac{\partial h}{\partial t} = -\nabla\cdot(\rho\boldsymbol{q}) + \rho g. \tag{2.2}$$

By the chain rule, the first term on the right-hand side is the sum of $-\nabla\rho\cdot\boldsymbol{q}$ and $-\rho\nabla\cdot\boldsymbol{q}$. Since the magnitude of the variation in density of fluid is negligible compared to the flux divergence term, Equation (2.2) can be simplified to

$$S_s\frac{\partial h}{\partial t} = -\nabla\cdot\boldsymbol{q} + g. \tag{2.3}$$

2.2 Darcy's Law

Darcy's law is an empirical law that describes flow through a porous media. The experiment on water filtration through sand beds carried out by Darcy in 1856 showed that the gradient of hydraulic head drives the fluid from high hydraulic head to low hydraulic head. Darcy's law can be written in differential form as

$$\boldsymbol{q} = -K\nabla h, \tag{2.4}$$

where $\boldsymbol{q}$ is the Darcy flux, K is hydraulic conductivity, and h is hydraulic head.

2.3 Hydraulic Conductivity Field

Hydraulic conductivity is a property of a porous medium that describes how easily a fluid can move through it. For example, hydraulic conductivity has higher values for sand or gravel compared with that for clay. In a heterogeneous geologic formation, hydraulic conductivity is a function of position. We treat the hydraulic conductivity, K, in Equation (2.4) as a random function. In other words, nominally, the parameter K belongs to an infinite-dimensional space.

Truncating a Karhunen-Loève Expansion (KLE) is a classical option for deriving finite-dimensional parameterizations for $\ln K$. Here, we summarize some of the pertinent details and refer the interested reader to [26, 44] for more information. Constructing the KLE first requires specification of a covariance function. This may be obtained, for instance, assuming a stationary random field and using a variogram on available data from a sparse set of boreholes. To ensure positive definiteness of the hydraulic conductivity, we often construct the KLE of $Y(\boldsymbol{x}, \omega)$ where $Y(\boldsymbol{x}, \omega) := \ln[K(\boldsymbol{x}, \omega)]$, $\boldsymbol{x}$ is the position vector defined over the domain $\boldsymbol{D}$, and ω belongs to the space of random events $\boldsymbol{\Omega}$. Let $\bar{Y}(\boldsymbol{x})$ denote the expected value of $Y(\boldsymbol{x}, \omega)$ over all possible realizations of the process, and $C(\boldsymbol{x}_1, \boldsymbol{x}_2)$ denote its covariance function (not to be confused with the contaminant concentration $c(x, t)$ in Equation (3.2)). Being an autocovariance function, $C(\boldsymbol{x_1}, \boldsymbol{x_2})$ is bounded, symmetric, and positive definite. Thus, it has the spectral decomposition

$$C(\boldsymbol{x_1}, \boldsymbol{x_2}) = \sum_{n=1}^{\infty} \lambda_n f_n(\boldsymbol{x_1}) f_n(\boldsymbol{x_2}) \tag{2.5}$$

where λ_n and $f_n(\boldsymbol{x})$ are the solutions to the homogeneous Fredholm integral equation of the second kind:

$$\int_{\boldsymbol{D}} C(\boldsymbol{x_1}, \boldsymbol{x_2}) f_n(\boldsymbol{x_1}) d\boldsymbol{x_1} = \lambda_n f_n(\boldsymbol{x_2}). \tag{2.6}$$

The eigenfunctions are orthogonal and form a complete set. They can be normalized according to the following criterion

$$\int_{\boldsymbol{D}} f_n(\boldsymbol{x}) f_m(\boldsymbol{x}) = \delta_{nm}. \tag{2.7}$$

Hence, $Y(\boldsymbol{x}, \omega)$ can be written as

$$Y(\boldsymbol{x}, \omega) = \bar{Y}(\boldsymbol{x}) + \sum_{n=1}^{\infty} \xi_n(\omega) \sqrt{\lambda_n} f_n(\boldsymbol{x}), \tag{2.8}$$

where λ_n and $f_n(\boldsymbol{x})$ are determined by $C(\boldsymbol{x}_1, \boldsymbol{x}_2)$, and $\{\xi_n(\omega)\}$ is a set of random variables that can be inferred from observations. The KLE of a Gaussian field has the further property that $\xi_n(\omega)$ are independent standard normal random variables [40]. Truncating the series in Equation (2.8) at the Nth term gives the finite-dimensional approximation

$$Y(\boldsymbol{x}, \omega) \approx \bar{Y}(\boldsymbol{x}) + \sum_{n=1}^{N} \xi_n(\omega) \sqrt{\lambda_n} f_n(\boldsymbol{x}). \tag{2.9}$$

The uncertain log hydraulic conductivity field is represented as weighted sums of predefined spatially variable basis functions. The truncated KLE provides a flexible and effective method for describing a spatially distributed hydraulic conductivity field. It reduces redundancy while capturing the most important features of the field.

2.3.1 Analytical Solution to KL expansion

The integral eigenvalue problem can be solved analytically for some special types of covariance functions defined on domains of simple geometric shape. Here we consider a one-dimensional random field characterized by an exponential covariance function. If we choose a separable covariance function, the following method can be extended to multidimensional rectangular domains as Equation (2.6) can be solved in each dimension independently.

Assuming the covariance function has the form of:

$$C(x_1, x_2) = \sigma_Y^2 e^{-\frac{|x_1 - x_2|}{\eta}}, \tag{2.10}$$

Equation (2.6) becomes

$$\sigma_Y^2 \int_0^L e^{-\frac{|x_1 - x_2|}{\eta}} f(x_2) dx_2 = \lambda f(x_1). \tag{2.11}$$

After differentiating Equation (2.11) with respect to x_1 by Leibniz rule, we have

$$-\frac{\sigma_Y^2}{\eta} \int_0^{x_1} e^{\frac{x_2 - x_1}{\eta}} f(x_2) dx_2 + \frac{\sigma_Y^2}{\eta} \int_{x_1}^L e^{\frac{x_1 - x_2}{\eta}} f(x_2) dx_2 = \lambda f'(x_1). \tag{2.12}$$

Taking the derivative with respect to x_1 again, we obtain the following equation:

$$f''(x_1) + \frac{2\eta\sigma_Y^2 - \lambda}{\lambda \eta^2} f(x_1) = 0. \tag{2.13}$$

To find the boundary condition of Equation (2.13), we let $x_1 = 0$ in Equation (2.11) and Equation (2.12). It is then obvious that

$$\eta f'(0) = f(0). \tag{2.14}$$

Similarly, we can determine the other boundary condition at $x_1 = L$

$$\eta f'(L) = -f(L). \tag{2.15}$$

The general solution of Equation (2.13) has the form of

$$f(x) = acos(\beta x) + bsin(\beta x), \quad \text{where } \beta^2 = \frac{2\eta\sigma_Y^2 - \lambda}{\lambda\eta^2}. \tag{2.16}$$

The boundary conditions require that

$$a - \eta\beta b = 0 \tag{2.17}$$

$$[-\beta\eta sin(\beta L) + cos(\beta L)]a + [\beta\eta cos(\beta L) + sin(\beta L)]b = 0 \tag{2.18}$$

The homogeneous system of linear equations has a unique trivial solution if and only if the determinant of the coefficient matrix is non-zero. In order for non-trivial solutions to exist, the determinant vanishes,

$$(\eta^2\beta^2 - 1)sin(\beta L) = 2\eta\beta cos(\beta L). \tag{2.19}$$

There are infinitely many solutions, $\beta_n, n = 1, 2, 3...$ to Equation (2.19) in increasing order. The corresponding eigenvalues are

$$\lambda_n = \frac{2\eta\sigma_Y^2}{\eta^2\beta_n^2 + 1}. \tag{2.20}$$

Since f_n are normalized eigenfunctions and Equation (2.16) holds, we can compute a_n and b_n:

$$a_n = \eta\beta_n \frac{1}{\sqrt{(\eta^2\beta_n^2 + 1)L/2 + \eta}}, \tag{2.21}$$

$$b_n = \frac{1}{\sqrt{(\eta^2\beta_n^2 + 1)L/2 + \eta}}. \tag{2.22}$$

2.3.2 Numerical Solution to KL expansion

More often, the integral equation can't be solved analytically due to a complex geometry or a more general covariance function. Therefore, we need numerical methods for the solution of the integral eigenvalue problem. The quadrature method and Galerkin's method are two very commonly used methods. The resulting KL expansion takes the form as:

$$\hat{Y}(\boldsymbol{x}, \omega) = \bar{Y}(\boldsymbol{x}) + \sum_{n=1}^{N} \hat{\xi}_n(\omega)\sqrt{\hat{\lambda}_n}\hat{f}_n(\boldsymbol{x}), \tag{2.23}$$

where $\hat{\lambda}_n$ and $\hat{f}_n(\boldsymbol{x})$ are approximations to the true eigenvalue and eigenfunctions. $\hat{\xi}_n(\omega)$ are standard uncorrelated random variables.

2.3.2.1 Quadrature Method

We discretize the integral on the left-hand side of Equation (2.6) as needed for computations. The integral is approximated by numerical integration:

$$\sum_{l=1}^{M} w_l Cov(\boldsymbol{x}, \boldsymbol{x}_l)\hat{f}_n(\boldsymbol{x}_l), \tag{2.24}$$

where $x_l, l = 1, ..., M$ are a finite set of M quadrature points in the domain, w_l is the corresponding integration weight, and $\hat{f}_n, n = 1, ..., N$ are approximations to the true eigenfunctions f_n. Therefore, Equation (2.6) can be written as:

$$\sum_{l=1}^{M} w_l Cov(\boldsymbol{x}, \boldsymbol{x}_l)\hat{f}_n(\boldsymbol{x}_l) = \hat{\lambda}_n\hat{f}_n(\boldsymbol{x}), \tag{2.25}$$

If we solve Equation (2.24) at the quadrature points, a set of equations can be formulated as:

$$\sum_{l=1}^{M} w_l Cov(\boldsymbol{x}_m, \boldsymbol{x}_l)\hat{f}_n(\boldsymbol{x}_l) = \hat{\lambda}_n \hat{f}_n(\boldsymbol{x}_m), \quad m = 1, ..., M \tag{2.26}$$

They can be expressed in matrix form:

$$\boldsymbol{C}\boldsymbol{W}\hat{\boldsymbol{f}}_n = \hat{\lambda}_n \hat{\boldsymbol{f}}_n, \tag{2.27}$$

where $\boldsymbol{C}$ is an $M \times M$ symmetric positive semi-definite matrix in which $c_{ml} = Cov(\boldsymbol{x}_m, \boldsymbol{x}_l)$, $\boldsymbol{W}$ is a diagonal matrix with nonnegative elements $w_{ll} = w_l$, and $\hat{\boldsymbol{f}}_n = (\hat{f}_n(\boldsymbol{x}_1), ..., \hat{f}_n(\boldsymbol{x}_M))'$ is an M dimensional vector. We can then solve Equation (2.25) for the interpolation formula of the eigenfunction $\hat{f}_n(\boldsymbol{x})$:

$$\hat{f}_n(\boldsymbol{x}) = \frac{1}{\hat{\lambda}_n} \sum_{l=1}^{M} w_l \hat{f}_n(\boldsymbol{x}_l) C(\boldsymbol{x}, \boldsymbol{x}_l). \tag{2.28}$$

The KL expansion of the random field is approximated as:

$$\hat{Y}(\boldsymbol{x}, \omega) = \bar{Y}(\boldsymbol{x}) + \sum_{n=1}^{N} \frac{\hat{\xi}_n(\omega)}{\sqrt{\hat{\lambda}_n}} \sum_{l=1}^{M} w_l \hat{f}_n(\boldsymbol{x}_l) C(\boldsymbol{x}, \boldsymbol{x}_l) \tag{2.29}$$

2.3.2.2 Galerkin Method

Galerkin methods can also be used to solve the integral equation. We let $\varphi_l(\boldsymbol{x})$ be a finite set of basis functions, and expand $f_n(\boldsymbol{x})$ with respect to this basis as:

$$f_n(\boldsymbol{x}) \approx \hat{f}_n(\boldsymbol{x}) = \sum_{l=1}^{M} d_l^n \varphi_l(\boldsymbol{x}). \tag{2.30}$$

Therefore, the residue of Equation (2.6) resulting from the truncated approximation of the eigenfunctions in Equation (2.30) is

$$r = \sum_{l=1}^{M} d_l^n [\int_{\boldsymbol{D}} C(\boldsymbol{x}_1, \boldsymbol{x}_2)\varphi_l(\boldsymbol{x}_2) d\boldsymbol{x}_2 - \lambda_n \varphi_l(\boldsymbol{x}_1)]. \tag{2.31}$$

According to Galerkin orthogonality, it yields a set of equations:

$$(r, \varphi_l(\boldsymbol{x})) = 0, \quad l = 1, ..., M. \tag{2.32}$$

Equivalently, they can be written in matrix form:

$$\boldsymbol{G}\boldsymbol{d}^n = \lambda_n \boldsymbol{B}\boldsymbol{d}^n, \tag{2.33}$$

where $\boldsymbol{G}$ is an $M \times M$ matrix in which

$$G_{ml} = \int_{\boldsymbol{D}} \int_{\boldsymbol{D}} C(\boldsymbol{x}_1, \boldsymbol{x}_2)\varphi_l(\boldsymbol{x}_2) d\boldsymbol{x}_2 \varphi_m(\boldsymbol{x}_1) d\boldsymbol{x}_1, \tag{2.34}$$

$\boldsymbol{B}$ is $M \times M$ with elements

$$B_{ml} = \int_{\boldsymbol{D}} \varphi_m(\boldsymbol{x})\varphi_l(\boldsymbol{x}) d\boldsymbol{x}. \tag{2.35}$$

We solve the generalized eigenvalue problem Equation 2.33 for $\boldsymbol{d}^n$ and λ_n. Next, $\boldsymbol{d}^n$ can be substituted into Equation 2.30 to obtain the approximated eigenfunctions of the covariance kernel.

2.4 Groundwater Flow Equation

Combining mass conservation and Darcy's law, the groundwater flow model can be written as

$$S_s \frac{\partial h}{\partial t} + \nabla \cdot \boldsymbol{q} = g \tag{2.36}$$

$$\boldsymbol{q} = -K\nabla h \tag{2.37}$$

subject to initial and boundary conditions, where S_s is specific storage, h is hydraulic head, $\boldsymbol{q}$ is flux, g is source or sink, and K is hydraulic conductivity.

We consider steady groundwater flow over domain Ω with boundary, $\partial\Omega = \Gamma$, that is decomposed into two parts in an incompressible saturated aquifer. The model can be simplified as:

$$\nabla \cdot \boldsymbol{q} = g \quad \text{in } \Omega \tag{2.38}$$

$$\boldsymbol{q} = -K\nabla h \quad \text{in } \Omega \tag{2.39}$$

$$h = h_D \quad \text{on } \Gamma_D \tag{2.40}$$

$$\boldsymbol{q} \cdot \boldsymbol{n} = f \quad \text{on } \Gamma_N \tag{2.41}$$

$$\Gamma = \bar{\Gamma}_D \cup \bar{\Gamma}_N, \Gamma_D \cap \Gamma_N = \emptyset, \Gamma_D \neq \emptyset \tag{2.42}$$

2.4.1 Variational Formulation

We define a Hilbert space

$$H(div) = H(div, \Omega) = \{\boldsymbol{\tau} \in L^2(\Omega; \mathbb{R}^2) \mid \nabla \cdot \boldsymbol{\tau} \in L^2(\Omega)\}. \tag{2.43}$$

We let

$$\Sigma_g = \{\boldsymbol{\tau} \in H(div) \mid \boldsymbol{\tau} \cdot \boldsymbol{n} = g \text{ on } \Gamma_N\}, \tag{2.44}$$

$$V = L^2(\Omega). \tag{2.45}$$

Multiplying Equation (2.38) by a scalar test function v and Equation (2.39) by a vector-valued test function τ, and then integrating over the domain Ω, we obtain a weak formulation: find $\boldsymbol{q} \in \Sigma_g$ and $h \in V$ such that

$$\int_{\Omega} \nabla \cdot \boldsymbol{q} v d\boldsymbol{x} = \int_{\Omega} g v d\boldsymbol{x} \quad \forall v \in V, \tag{2.46}$$

$$\int_{\Omega} K^{-1}\boldsymbol{q} \cdot \boldsymbol{\tau} d\boldsymbol{x} - \int_{\Omega} h \nabla \cdot \boldsymbol{\tau} d\boldsymbol{x} = -\int_{\Gamma_D} h_D \boldsymbol{\tau} \cdot \boldsymbol{n} ds \quad \forall \boldsymbol{\tau} \in \Sigma_0. \tag{2.47}$$

The boundary condition for the flux is now an essential boundary condition and should be enforced in the function space, while the other boundary condition becomes a natural boundary condition, which is applied to the variational form.

2.4.2 Mixed Finite Element Method

We choose finite dimensional subspaces $\Sigma^h \subset \Sigma$ and $V^h \subset V$, and the statement of the problem becomes: Find $\boldsymbol{q}^h \in \Sigma^h_g, h^h \in V^h$ such that

$$\int_{\Omega} \nabla \cdot \boldsymbol{q}^h v^h d\boldsymbol{x} = \int_{\Omega} g v^h d\boldsymbol{x} \quad \forall v^h \in V^h, \tag{2.48}$$

$$\int_{\Omega} K^{-1}\boldsymbol{q}^h \cdot \boldsymbol{\tau}^h d\boldsymbol{x} - \int_{\Omega} h^h \nabla \cdot \boldsymbol{\tau}^h d\boldsymbol{x} = -\int_{\Gamma_D} h_D \boldsymbol{\tau}^h \cdot \boldsymbol{n} ds \quad \forall \boldsymbol{\tau}^h \in \Sigma^h_0. \tag{2.49}$$

Several mixed finite element spaces may be considered, including the RTN spaces, BDM spaces, BDFM spaces, BDDF spaces, or CD spaces, to obtain a stable method.

2.4.3 Convergence Test

We consider the problem on a square domain, $\Omega = (0, 10) \times (0, 10)$. We construct a triangular mesh of $\boldsymbol{D}$ with n elements in each direction. We let

$$K(x, y) = 3.0 + sin(\pi x) + cos(2\pi y), \tag{2.50}$$

$$
\begin{aligned}
f = & -\frac{1}{4}\pi^2 sin(\frac{\pi}{2}y)(8cos(\pi x)cos(2\pi x) \\
& - 17sin(2\pi x)sin(\pi x) \\
& - 8sin(2\pi x)cos(\pi y) \\
& - 21sin(2\pi x)cos(2\pi y) \\
& - 55sin(2\pi x)).
\end{aligned} \tag{2.51}
$$

We impose Dirichlet conditions of $h_D = 0$ on the left and right boundaries. On the top and bottom boundaries

$$\boldsymbol{q} \cdot \boldsymbol{n} = 0 \tag{2.52}$$

and

$$\boldsymbol{q} \cdot \boldsymbol{n} = -\frac{\pi}{2}(3.0 + sin(\pi x) + cos(2\pi y))sin(2\pi x), \tag{2.53}$$

respectively. The exact solutions to this simple case are:

$$h_e = sin(2\pi x)sin(\frac{\pi}{2}y), \tag{2.54}$$

and $\boldsymbol{q}_e = K\nabla h_e$.

We choose Raviart-Thomas elements of order 1 for Σ^h, and piecewise constant for V^h. The numerical solution is obtained by using the FEniCS package. We plot the computed head and flux in Figure 2.1 and Figure 2.2. Figure 2.3 shows the discretization errors in L^2 as a function of the mesh size h. We observe that the numerical results are consistent with the finite element convergence theory that

$$\|\boldsymbol{q}_e - \boldsymbol{q}^h\|_{L^2} \leq Ch, \tag{2.55}$$

$$\|h_e - h^h\|_{L^2} \leq Ch. \tag{2.56}$$

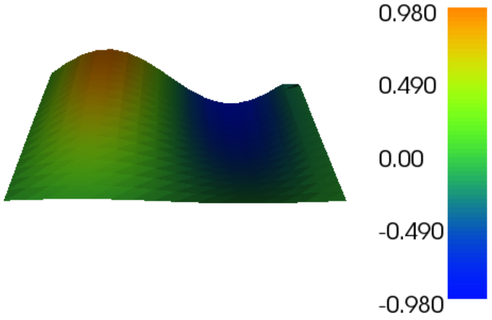


Figure 2.1: Hydraulic head

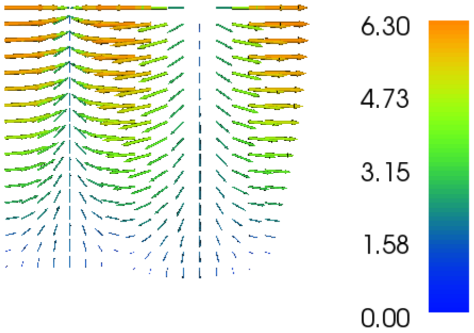


Figure 2.2: Flux

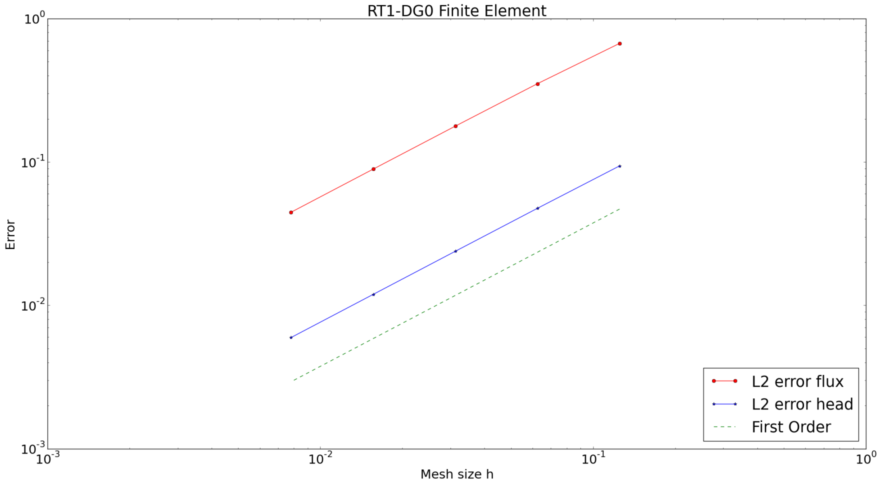


Figure 2.3: Error against mesh size

Chapter 3

Transport Model

Non-reactive subsurface contaminant transport in a single fluid phase can be described by a simple scalar advection-diffusion equation. However, the numerical solution to the model is still a challenge when advection is dominant. Many methods have been developed to avoid spurious oscillations. In this chapter, we first use the streamline upwind Petrov Galerkin (SUPG) method which stabilizes the numerical solution but still exhibits local oscillations in crosswind directions when gradients of the contaminant concentration are large. A nonlinear crosswind dissipation is then added to the SUPG formulation as an additional stabilization. The resulting nonlinear scheme can be solved by using linearizion through simple iteration. We show a numerical example to demonstrate the additional crosswind diffusion damps the overshoots of the SUPG solution.

3.1 Advection Diffusion Equation

Transport of solutes in porous medium can be described by conservation of mass. It states that the net rate of change of mass of solute within a control volume equals sum of the net flux of solute into the control volume and sources/sinks inside the control volume. Advection and diffusion are two components of solute movement. The former is the transport of solute caused by the flowing groundwater that carries the solute. The latter describes the process of dispersion due to molecular diffusion. Mathematical descriptions of solute transport can be written as

$$\boldsymbol{u} = \frac{\boldsymbol{q}}{\phi} \tag{3.1}$$

$$\frac{\partial c}{\partial t} + \nabla \cdot (\boldsymbol{u} c) - \nabla \cdot (\boldsymbol{\kappa} \nabla c) = f \quad \text{in } \Omega, \tag{3.2}$$

where c is the solute concentration, $\boldsymbol{u}$ is the velocity field, $\boldsymbol{\kappa}$ is the diffusivity and f is the source. We assume the following boundary conditions associated with Equation (3.2)

$$c = g \quad \text{on } \Gamma_D, \tag{3.3}$$

$$\boldsymbol{\kappa} \nabla c \cdot \boldsymbol{n} = 0 \quad \text{on } \Gamma_N, \tag{3.4}$$

where g is a given function, and $\boldsymbol{n}$ is the unit normal vector at the boundary. The initial condition is imposed as:

$$c(\boldsymbol{x}, 0) = c_0(\boldsymbol{x}) \quad \text{in } \Omega. \tag{3.5}$$

3.1.1 Semi-Discrete Galerkin Method

We define the space of trial solutions S and the space of weighting functions V as:

$$S = \{c(\cdot, t) \in H^1(\Omega) \mid c = g \text{ on } \Gamma_D\} \tag{3.6}$$

$$V = \{w \in H^1(\Omega) \mid w = 0 \text{ on } \Gamma_D\} \tag{3.7}$$

Multiplying Equation (3.2) by a test function w and integrating by parts, we have the variational formulation of Equation (3.2): Find $c \in S$, such that

$$\int_\Omega w\frac{\partial c}{\partial t}d\Omega - \int_\Omega \nabla w \cdot \boldsymbol{u}c + \nabla w \cdot (\boldsymbol{\kappa}\nabla c)d\Omega = \int_\Omega wfd\Omega \quad \forall w \in V \tag{3.8}$$

Assume we have a finite element partition of the domain Ω. Let $S^h \subset S$ and $V^h \subset V$ be finite-dimensional trial solution and test function spaces.

$$S^h = \{c^h(\cdot, t) \in H^1(\Omega) \mid c^h = g \text{ on } \Gamma_D\} \tag{3.9}$$

$$V^h = \{w^h \in H^1(\Omega) \mid w^h = 0 \text{ on } \Gamma_D\} \tag{3.10}$$

The Galerkin approximation formulation of Equation (3.8) can be stated as: Find $c^h \in S^h$, such that

$$\begin{aligned} &\int_\Omega w^h\frac{\partial c^h}{\partial t}d\Omega - \int_\Omega \nabla w^h \cdot \boldsymbol{u}^h c^h + \nabla w^h \cdot (\boldsymbol{\kappa}^h\nabla c^h)d\Omega \\ &= \int_\Omega w^h f^h d\Omega \quad \forall w^h \in V^h \end{aligned} \tag{3.11}$$

or, in an abstract compact form,

$$(w^h, c_t^h) + B_G(w^h, c^h) = L(w^h) \quad \forall w^h \in V^h \tag{3.12}$$

where

$$B_G(w^h, c^h) := -\int_\Omega \nabla w^h \cdot \boldsymbol{u}^h c^h + \nabla w^h \cdot (\boldsymbol{\kappa}^h\nabla c^h)d\Omega \tag{3.13}$$

$$L(w^h) := \int_\Omega w^h f^h d\Omega \tag{3.14}$$

The trial solution and weighting function are continuous functions written as:

$$c^h(\boldsymbol{x}, t) = \sum_{i=1}^{N} c_i(t) N_i(\boldsymbol{x}), \tag{3.15}$$

$$w^h(\boldsymbol{x}) \sum_{i=1}^{N} w_i N_i(\boldsymbol{x}), \tag{3.16}$$

where N_i is the standard nodal basis of S^h.

The problem above can be formulated in matrix form:

$$\boldsymbol{M}\dot{\boldsymbol{c}}(t) + \boldsymbol{K}\boldsymbol{c}(t) = \boldsymbol{f}(t), \tag{3.17}$$

where the dot represents the time derivative. $\boldsymbol{c}$ is the vector of time-dependent nodal values of c^h.

$$M_{ij} = (N_i, N_j), \tag{3.18}$$

$$K_{ij} = B_G(N_i, N_j), \tag{3.19}$$

$$f_i = L(N_i). \tag{3.20}$$

Various numerical schemes can be applied to solve the above ordinary differential equation.

3.1.2 Semi-Discrete Stabilized Method

When the Peclet number increases, the flow becomes advection dominated. Solving the advection-diffusion equation by the standard Galerkin method results in unphysical oscillation of the numerical solution. To remedy the spurious oscillations, we use Steamline Upwind Petrov-Galerkin (SUPG) method to solve the equation. In SUPG, artificial diffusion is added over element interiors along the steamline direction to increase the stability of the solution. The resulting scheme can be written as: Find $c^h \in S^h$, such that

$$\begin{aligned}
&\int_{\Omega} w^h \frac{\partial c^h}{\partial t} d\Omega - \int_{\Omega} \nabla w^h \cdot \boldsymbol{u} c^h + \nabla w^h \cdot (\boldsymbol{\kappa} \nabla c^h) d\Omega \\
&+ \sum_{e=1}^{N_{el}} \int_{\Omega^e} \tau_{SUPG} \boldsymbol{u} \cdot \nabla w^h R(c^h) d\Omega \\
&= \int_{\Omega} w^h f d\Omega \quad \forall w^h \in V^h
\end{aligned} \tag{3.21}$$

or,

$$(w^h, c_t^h) + B_G(w^h, c^h) + \sum_{e=1}^{N_{el}} (\tau_{SUPG} \boldsymbol{u} \cdot \nabla w^h, R(c^h))_{\Omega^e} = L(w^h) \quad \forall w^h \in V^h, \tag{3.22}$$

where

$$R(c^h) := \frac{\partial c^h}{\partial t} + \nabla \cdot (\boldsymbol{u} c^h) - \nabla \cdot (\boldsymbol{\kappa} \nabla c^h) - f, \tag{3.23}$$

$$\tau_{SUPG} = \frac{\alpha h}{2 |\boldsymbol{u}|} \tag{3.24}$$

$$\alpha = coth\gamma - \frac{1}{\gamma} \tag{3.25}$$

$$\gamma = \frac{|\boldsymbol{u}| h}{2\kappa} \tag{3.26}$$

This method has strong consistency as the terms added to the standard Galerkin method vanish for all sufficiently smooth solutions.

The semidiscrete equation is a system of ODE's

$$\boldsymbol{M} \dot{\boldsymbol{c}}(t) + \boldsymbol{K} \boldsymbol{c}(t) = \boldsymbol{F}(t) \tag{3.27}$$

where

$$M_{ij} = (N_i, N_j) + \sum_{e=1}^{N_el} (\tau_{SUPG} \boldsymbol{u} \cdot \nabla N_i, N_j), \tag{3.28}$$

$$K_{ij} = B_G(N_i, N_j) + \sum_{e=1}^{N_el} (\tau_{SUPG} \boldsymbol{u} \cdot \nabla N_i, \nabla \cdot (\boldsymbol{u} N_j) - \nabla \cdot (\boldsymbol{\kappa} \nabla N_j)), \tag{3.29}$$

$$f_i = L(N_i) + (\tau_{SUPG} \boldsymbol{u} \cdot \nabla N_i, f). \tag{3.30}$$

3.1.3 Time Integration

There are many numerical methods available to solve the following systems of ordinary differential equations by advancing transient solutions step-by-step,

$$\boldsymbol{M} \dot{\boldsymbol{c}}(t) + \boldsymbol{K} \boldsymbol{c}(t) = \boldsymbol{f}(t). \tag{3.31}$$

Linear multistep methods and Runge-Kutta methods are two main categories of numerical methods for solving first-order initial value problem. Furthermore, we can divide them into two groups that are explicit or implicit. For example, Adams-Moulton methods and backward differentiation methods (BDF) are implicit linear multistep methods, whereas diagonally implicit Runge-Kutta (DIRK), singly diagonally implicit runge kutta (SDIRK), and Gauss-Radau (based on Gaussian quadrature) numerical methods are implicit Runge-Kutta methods. Explicit linear multistep methods include the Adams-Bashforth methods. The most well known member of the Runge-Kutta family, RK4, and a generalization of the RK4 method are explicit methods. In this work, we use the standard $\theta-$ method to fully discretize the Equation (3.31) into a linear system of algebraic equations as

$$(\boldsymbol{M} + \theta \Delta t \boldsymbol{K}) \boldsymbol{c}_{n+1} = \theta \Delta t \boldsymbol{f}_{n+1} + (1-\theta)\Delta t \boldsymbol{f}_n + (\boldsymbol{M} - (1-\theta)\Delta t \boldsymbol{K}) \boldsymbol{c}_n \quad (3.32)$$

where $\Delta t = t_{n+1} - t_n$ is the time step, and $0 \leq \theta \leq 1$ is a real parameter. When $\theta = 0$ and $\theta = 1$, the scheme becomes the explicit forward Euler and implicit backward Euler scheme, respectively, which both give the first-order accuracy. For $\theta = \frac{1}{2}$, it is the second-order unconditionally stable Crank-Nicolson method.

3.1.4 Crosswind-dissipation

3.1.4.1 Discontinuity-capturing crosswind-dissipation

In some cases where the solution has sharp gradients, the SUPG formulation alone does not completely remove the oscillations. The discontinuity-capturing technique, also known as the shock-capturing, is proposed to circumvent this problem by introducing more numerical diffusion into the system besides the streamline diffusion. In the literature, various researchers have developed several shock-captureing methods [33, 42, 55, 61]. In this work, we use the method proposed in [20] which is less diffusive than other such methods. It keeps the artificial diffusion the same as that in the SUPG formulation along the direction of the steamlines, and adds extra modified crosswind diffusion properly. Specifically, we let $\boldsymbol{u}_{\|}$ be the projection of $\boldsymbol{u}$ onto ∇c_h, which is defined as

$$\boldsymbol{u}_{\|} = \frac{\boldsymbol{u} \cdot \nabla c_h}{|\nabla c_h|^2} \nabla c_h \quad (3.33)$$

when $|\nabla c_h|$ is nonzero. The corresponding element Peclet number can be computed as

$$\gamma_{\|}^{e} = \frac{|\boldsymbol{u}_{\|}|\, h}{2\kappa} \tag{3.34}$$

$\gamma_{\|}^{e}$ is small in the regions where $|\boldsymbol{u} \cdot \nabla c_h|$ is small.

The crosswind diffusion added to the left-hand side of Equation SUPG can be described as

$$A_{SC}(c^h; w^h, c^h) := \sum_{e=1}^{N_{el}} \int_{\Omega^e} \frac{1}{2} \alpha_c^e h^e \frac{|R(c_h)|}{|\nabla c_h|} \nabla w_h \cdot (\boldsymbol{I} - \frac{1}{|\boldsymbol{u}|^2} \boldsymbol{u} \otimes \boldsymbol{u}) \cdot \nabla c_h d\Omega \tag{3.35}$$

where $\boldsymbol{I}$ is the unit tensor. The function α_c^e is defined as

$$\alpha_c^e = max\{0, C - \frac{1}{\gamma_{\|}^{e}}\}, \tag{3.36}$$

where C is an empirical constant which is often set to be 0.7 in $2D$ problems for linear elements. It is obvious that the crosswind diffusion is proportional to the residual defined within each element. Therefore, the consistency property still holds. Moreover, when $|\boldsymbol{u} \cdot \nabla c_h|$ is small, α_c^e will take a value close or equal to 0. That means less or no crosswind diffusion will be add to the regions where the convective term of the residual is small, which improves the accuracy of this method. In Figure 3.1, we show the artificial diffusion added in the streamline and crosswind directions in a 2D case.

The crosswind diffusion defined in Equation (3.35) is nonlinear. Nonlinear methods like Newton-GMRES can be applied to solve the resulting nonlinear algebraic system.

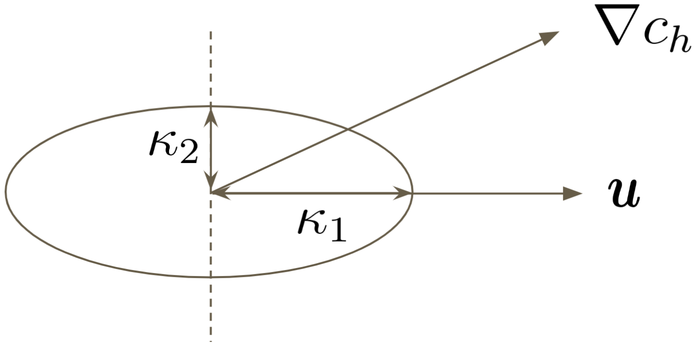


Figure 3.1: A schematic figure showing streamline diffusion and crosswind diffusion where $\kappa_1 = \kappa + \frac{1}{2}\alpha h\,|\boldsymbol{u}|$ and $\kappa_2 = \kappa + \frac{1}{2}\alpha_c^e h\,|R(c_h)|\,/\,|\nabla c_h|$.

3.1.4.2 Linearization of the nonlinear problem

As noted, the crosswind diffusion defined in Equation (3.35) is nonlinear. As an alternative to the nonlinear methods used in [61], we use a simple two-iteration method to solve the nonlinear equation at a low computational cost. At each time step, we first solve the transport equation by SUPG for c^h_{SUPG}:

$$\begin{aligned}
&(w^h, \frac{\partial c^h_{SUPG}}{\partial t}) + B_G(w^h, c^h_{SUPG}) \\
+&\sum_{e=1}^{N_{el}} (\tau_{SUPG}\boldsymbol{u}\cdot\nabla w^h, R(c^h_{SUPG}))_{\Omega^e} \\
=&L(w^h) \quad \forall w^h \in V^h
\end{aligned} \tag{3.37}$$

In the second iteration, we determine the magnitude of the crosswind diffusion based on the solution c^h_{SUPG} and solve the linearized equation for c^h:

$$\begin{aligned}&(w^h, \frac{\partial c^h}{\partial t}) + B_G(w^h, c^h) \\ &+\sum_{e=1}^{N_{el}}(\tau_{SUPG}\boldsymbol{u} \cdot \nabla w^h, R(c^h))_{\Omega^e} + A_{SC}(c^h_{SUPG}; w^h, c^h) \\ &=L(w^h) \quad \forall w^h \in V^h\end{aligned} \tag{3.38}$$

where

$$A_{SC}(c^h_{SUPG}; w^h, c^h) = \sum_{e=1}^{N_{el}} \int_{\Omega^e} \frac{1}{2}\alpha^e_c h^e \frac{\left|R(c^h_{SUPG})\right|}{\left|\nabla c^h_{SUPG}\right|} \nabla w^h \cdot (\boldsymbol{I} - \frac{1}{|\boldsymbol{u}|^2}\boldsymbol{u} \otimes \boldsymbol{u}) \cdot \nabla c^h d\Omega \tag{3.39}$$

3.1.4.3 Numerical example

We consider a transport problem in $\Omega = (0,1)\times(0,1)$ with homogeneous Dirichlet boundary conditions. We assume the solute concentration is zero everywhere at the initial time. The model parameters are taken as $\boldsymbol{u} = (0,1)$, $\kappa = 10^{-8}$, and $f = 1$. We solve the problem within FEniCS using continuous piecewise linear elements for spatial discretization on a uniform triangular mesh of 65×65 and the backward Euler method with a uniform time step of 10^{-2} for time integration. We integrate in time until $T = 0.3$.

Figure 3.2 and Figure 3.3 show contours of the solutions at $T = 0.3$ which are computed using SUPG with and without crosswind diffusion. From the figures, it is observed that in the SUPG solution there are localized oscillations near the boundaries where the gradient of the solution is sharp. In

Figure 3.4 and Figure 3.5, we see that those spurious oscillations are suppressed with the application of the crosswind diffusion.

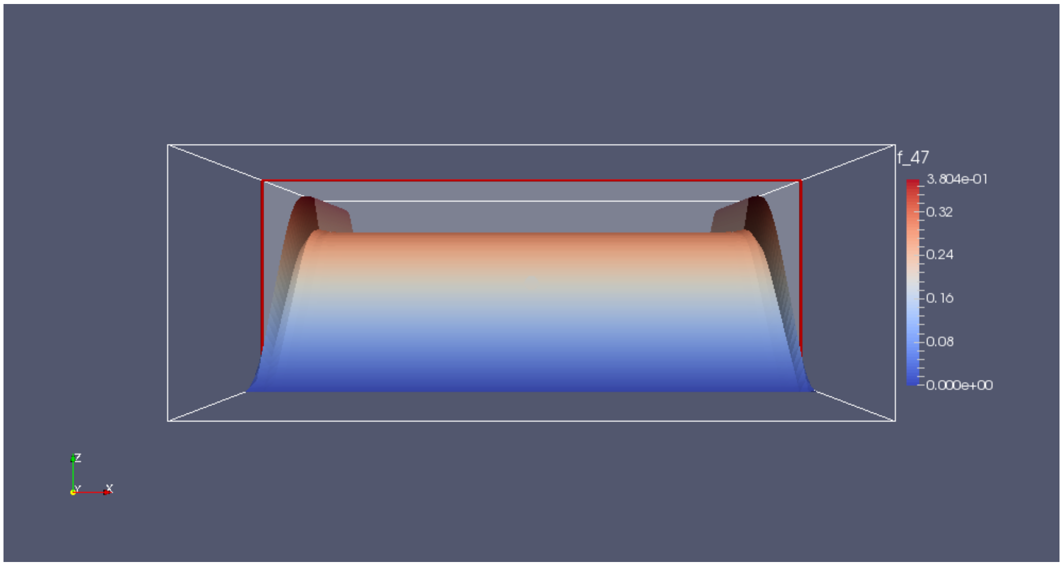


Figure 3.2: SUPG solution at $T = 0.3$

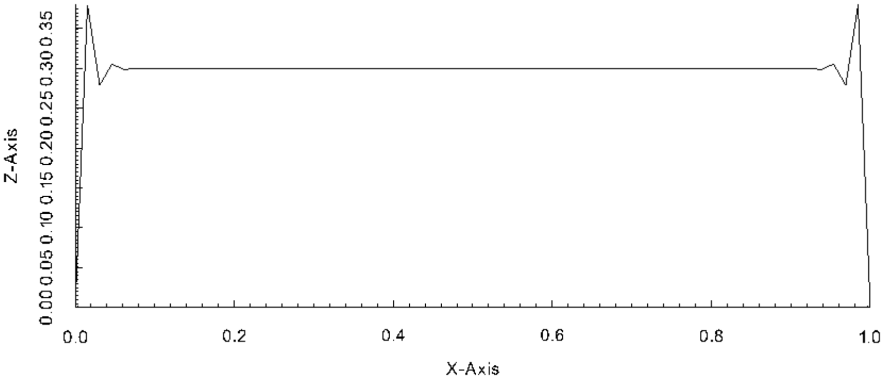


Figure 3.3: Slice of SUPG solution at $y = 0.5$ at $T = 0.3$

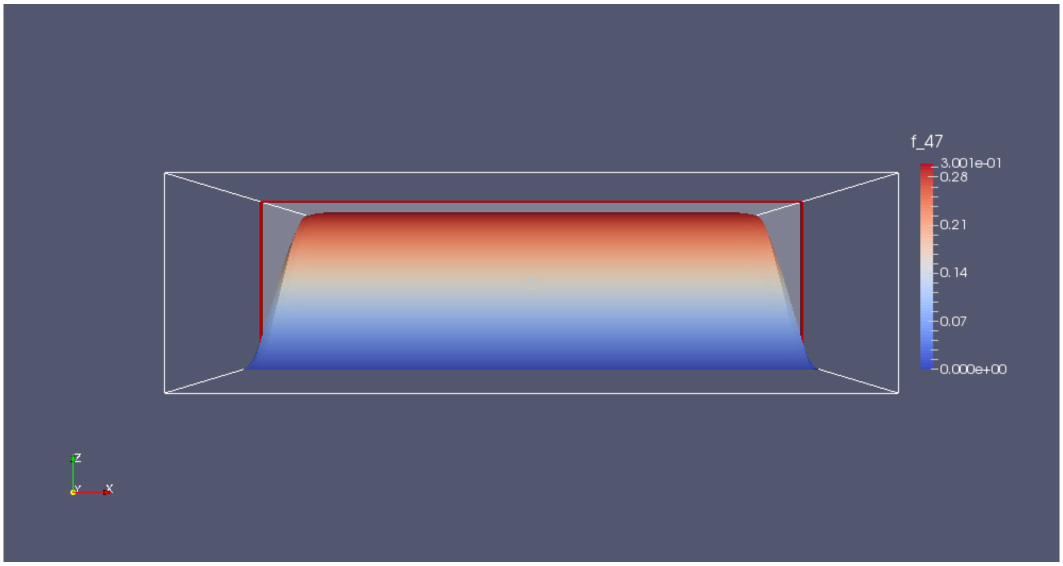


Figure 3.4: SUPG with crosswind diffusion solution at $T = 0.3$

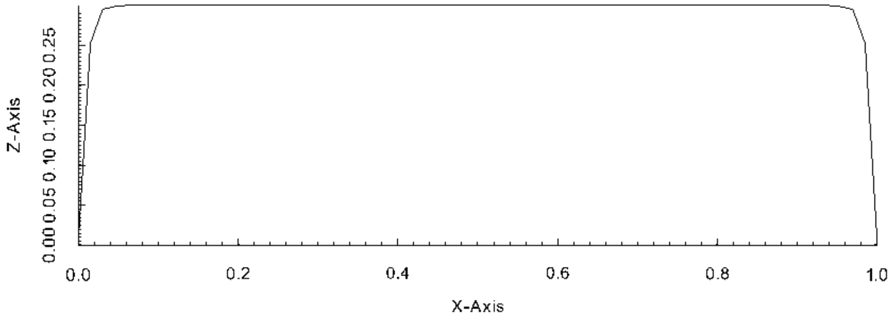


Figure 3.5: Slice of SUPG with crosswind diffusion solution at $y = 0.5$ at $T = 0.3$

Chapter 4

Surrogate Model[1]

It is often computationally expensive to solve an inverse problem. One of the high computational cost may come from thousands of forward simulation evaluations. To improve the efficiency, several types of surrogate models have been studied to replace expensive physics model simulations. In this chapter, we use Support Vector Machine (SVM) to build surrogate models to approximate a response surface between model parameters and a quantity of interest. Based on a small set of sampled data obtained by solving the forward problem with randomly chosen inputs, SVM surrogate models can be built to predict model output (quantity of interest) of an unseen input (model parameters). Compared with a true model solve, computational cost associated with a surrogate model evaluation is negligible.

4.1 Surrogate Modeling with Support Vector Machines

The theory of SVM was developed based on statistical learning theory for the purpose of classification, and later extended for regression [66]. Suppose

[1] This chapter is based on the article entitled *Data-driven uncertainty quantification for predictive flow and transport modeling using support vector machines* by Jiachuan He, Steven Mattis, Troy Butler and Clint Dawson [32].

we are given a set of l training points, $\{(\boldsymbol{x}_1, y_1), ..., (\boldsymbol{x}_l, y_l)\}$, where $\boldsymbol{x}_i \in \mathbf{R}^n$ is an input vector and $y_i \in \mathbf{R}^1$ is the target output. The solution to the regression problems define the SVM to approximate the relation between the input vector and the output, thereby estimating the values of the output at unsampled points in the space of the input domain.

As a first step, the input vector $\boldsymbol{x}$ is mapped to a higher dimensional feature space by a map, $\Phi(\boldsymbol{x})$. Then, the regression tries to find a function $f(\boldsymbol{x})$ that is within an error tolerance of ε away from the given outputs in the feature space. The regression function takes the general form:

$$f(\boldsymbol{x}) = \langle \boldsymbol{w}, \Phi(\boldsymbol{x}) \rangle + b, \tag{4.1}$$

where $\boldsymbol{w}$ is a vector in the feature space. In this case, the norm of $\boldsymbol{w}$ indicates the flatness of the function. The regression problem can be mathematically expressed in terms of the following optimization problem:

$$\begin{gathered} \min \frac{1}{2} \|\boldsymbol{w}\|^2 + P \sum_{i=1}^{l} (\zeta_i + \zeta_i^*) \\ \text{subject to} \begin{cases} y_i - f(\boldsymbol{x}_i) \leq \varepsilon + \zeta_i \\ f(\boldsymbol{x}_i) - y_i \leq \varepsilon + \zeta_i^* \\ \zeta_i, \zeta_i^* \geq 0 \\ i = 1, ..., l \end{cases} \end{gathered} \tag{4.2}$$

where the positive constant P determines the trade-off between the flatness of f and the amount up to which deviations larger than ε are tolerated. In other words, P determines how much large deviations are penalized in the regression. The slack variables ζ and ζ^* are described as

$$\zeta_i, \zeta_i^* = \begin{cases} 0, & \text{if } |y_i - f(\boldsymbol{x_i})| \le \varepsilon \\ |y_i - f(\boldsymbol{x_i})| - \varepsilon, & \text{otherwise} \end{cases} \tag{4.3}$$

In other words, points inside the margin (dotted lines in Figure 4.1) do not contribute to the cost function.

The above optimization problem is usually solved in its Lagrangian dual form:

$$\max -\frac{1}{2}\sum_{i,j=1}^{l}(\alpha_i - \alpha_i^*)(\alpha_j - \alpha_j^*)\langle\Phi(\boldsymbol{x_i}), \Phi(\boldsymbol{x_j})\rangle - \varepsilon\sum_{i=1}^{l}(\alpha_i + \alpha_i^*) + \sum_{i=1}^{l} y_i(\alpha_i - \alpha_i^*)$$

$$\text{subject to } \begin{cases} \sum_{i=1}^{l}(\alpha_i - \alpha_i^*) = 0, \\ 0 \le \alpha_i, \alpha_i^* \le P, \\ i = 1, ..., l, \end{cases} \tag{4.4}$$

where α_i and α_i^* are Lagrange multipliers. In the derivation of Equation (4.4), by setting the derivatives of the Lagrangian with respect to $\boldsymbol{w}$ to zero, we have

$$\boldsymbol{w} = \sum_{i=1}^{l}(\alpha_i - \alpha_i^*)\Phi(\boldsymbol{x_i}). \tag{4.5}$$

Thus, the regression function can be rewritten as

$$f(\boldsymbol{x}) = \sum_{i=1}^{l}(\alpha_i - \alpha_i^*)k(\boldsymbol{x_i}, \boldsymbol{x}) + b, \tag{4.6}$$

where $k(\boldsymbol{x_i}, \boldsymbol{x}) = \langle\Phi(\boldsymbol{x}_i), \Phi(\boldsymbol{x})\rangle$ is the kernel function (not to be confused with the hydraulic conductivity K in Equation (2.37)). The values of $\{\alpha_i\}_{i=1}^{l}$ and $\{\alpha_i^*\}_{i=1}^{l}$ are obtained by solving the dual problem, and b can be computed by exploiting the so called Karush-Kuhn-Tucker (KKT) conditions. Also, from the KKT condition, it follows that for all points inside the margin, the corresponding α_i and α_i^* vanish. In general, when the dimensionality of $\boldsymbol{w}$ is higher than the number of data points, it is easier to solve the optimization problem in its dual formulation. Once the dual problem is solved, the function value at any unsampled point depends only on the inner product between $\Phi(\boldsymbol{x})$ and the points in the training set with non-zero α_i values. Moreover, working with the dual problem enables us to perform the kernel trick method. Rather than mapping the input vectors through an explicit Φ and working in the enlarged feature space, it is sufficient to know $k(\boldsymbol{x}_i, \boldsymbol{x}_j)$. This is important because in many applications of SVMs, the dimensionality of the feature space is so high that it can easily become computationally infeasible. By using kernels, one only needs to compute $K(\boldsymbol{x}_i, \boldsymbol{x}_j)$ for all $\binom{l}{2}$ distinct pairs i, j in Equation (4.4). Therefore, the dimensionality of the feature space does not affect the computation. Note that $\boldsymbol{w}$ is no longer given explicitly this way. Algorithmic details for computing the values needed to evaluate Equation (4.6) are discussed by [23].

Some commonly used kernel types in SVM are linear, polynomial, sigmoid and radial basis functions; see [35] for more information. The penalty parameter P and kernel parameters are then often determined using grid search with cross-validation.

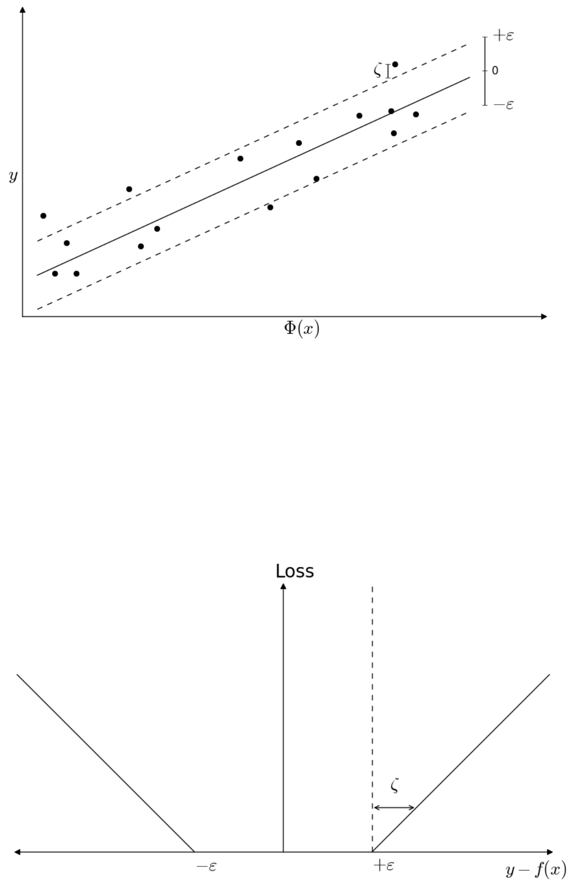


Figure 4.1: A schematic of ε-insensitive loss function

4.2 SVM for Subsurface Flow Models

In this section, a two-dimensional model of saturated flow is used to construct the SVM surrogate models used in this study. This model is also used for numerical examples in Chapter 6.

We consider steady groundwater flow over domain $\boldsymbol{D} = (0, 10) \times (0, 10)[L^2]$ in heterogeneous porous media; see Figure 4.2 for an illustration. We impose Dirichlet conditions of $h_L = 15[L]$ and $h_R = 10[L]$ on the left and right boundaries, respectively. On the top and bottom boundaries, $\boldsymbol{q} \cdot \boldsymbol{n} = 0[LT^{-1}]$, where $\boldsymbol{n}$ denotes the outward directed boundary normal. We assume, for simplicity, that $Y(\boldsymbol{x}, \omega) = \ln[K(\boldsymbol{x}, \omega)]$ is Gaussian [25, 34] with zero mean and a separable exponential covariance function,

$$C(\boldsymbol{x_1}, \boldsymbol{x_2}) = C(x_1, y_1; x_2, y_2) = \sigma_Y^2 e^{[-\frac{|x_1 - x_2|}{\eta_1} - \frac{|y_1 - y_2|}{\eta_2}]}, \tag{4.7}$$

where $\sigma_Y^2 = 2$, $\eta_1 = 10[L]$ and $\eta_2 = 4[L]$ are the variance and the correlation lengths of the random field. Consequently, $\ln K$ can be expanded with the form of Equation (2.8) where eigenvalues and the corresponding eigenfunctions can be analytically determined in this case according to [70]. We use the FEniCS package [3, 45] to solve the groundwater flow and transport models for the concentration. Equation (2.36) is solved using the Raviart-Thomas mixed method on a 64 × 64 mesh with triangular elements. For Equation (3.2), suppose we have geophysically reasonable parameters $\phi = 0.1, D = 5[L^2T^{-1}]$. There is no contaminant in the domain at the initial time. The concentration is prescribed on the left boundary as the contaminant source, $C_L = 50[ML^{-3}]$,

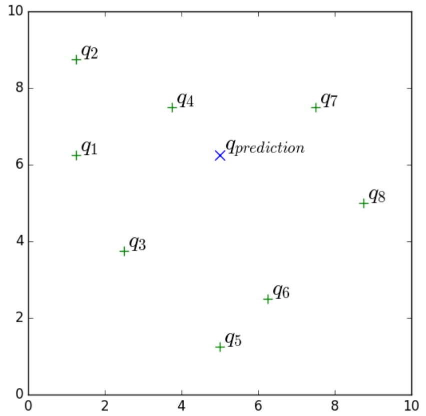


Figure 4.2: The flow domain. 8 green + are the only measurement locations where contaminant concentrations are available; Blue × is the prediction location where concentration is predicted.

and no-flow (i.e., zero Neumann) otherwise. The system is discretized in time using the Crank-Nicolson method with a time step of $dt = 0.05[T]$, and then solved by the streamline upwind Petrov Galerkin method on the same mesh.

We use the inverse transformation method (see Chapter 2 in [59]) to transform the $\mathcal{N}(0,1)$ distributed random variables, $\xi_i(\omega)$, to $\mathcal{U}(0,1)$ distributed random variables, so that we can define the parameter domain as the unit hypercube $\Xi = [0,1]^9$. For the sake of notational simplicity, we also let ξ_i denote the transformed uniform random variables. Then, any inverse transformation of a point $\boldsymbol{\xi} = (\xi_1, \xi_2, ..., \xi_9) \in \Xi$ realizes a log hydraulic conductivity field via Equation (2.9), and the nine-dimensional domain is mapped to an eight-dimensional output space via the parameter-to-observables map

$$Q(\boldsymbol{\xi}) = [q_1(\boldsymbol{\xi}), q_2(\boldsymbol{\xi}), q_3(\boldsymbol{\xi}), q_4(\boldsymbol{\xi}), q_5(\boldsymbol{\xi}), q_6(\boldsymbol{\xi}), q_7(\boldsymbol{\xi}), q_8(\boldsymbol{\xi})],$$

which involves solving the flow and contaminant transport models with the corresponding K and calculating the solution at the eight observation locations in the physical domain (see Figure 4.2) at $T = 2[T]$. We draw 5000 independent identically distributed (i.i.d.) sample points in Ξ, and compute the corresponding concentrations in $\mathcal{D}$.

We use the open-source software LIBSVM [16] to construct a response approximation between $\boldsymbol{\xi}$ (input) and the contaminant concentration at each observation/prediction location, $q_i(\boldsymbol{\xi})$ (output) for $i = 1, ..., 8$, using the 5000 model evaluations as a training set. We let $\varepsilon = 0.1$ in the loss function Equation (4.2).

Since the dimension of input parameters is low, the RBF kernel,

$$k(\boldsymbol{x}_i, \boldsymbol{x}_j) = e^{-\gamma\|\boldsymbol{x}_i - \boldsymbol{x}_j\|^2}, \tag{4.8}$$

is naturally a good choice since it can handle the nonlinear relation between input and output with less hyper-parameters compared with other kernels, e.g., polynomial kernel, see [35] for more information regarding choosing kernels. The feature space in this case is implicitly defined and infinite-dimensional. Two hyper-parameters, the penalty parameter P and γ in the RBF kernel, must be determined to construct the SVM. We use a straightforward two-step grid-search method to find the optimal hyper-parameter pair (P, γ). A coarse and fine grid search with 10-fold cross-validation are performed on a subset of size 1000 from the training set to determine the optimal hyper-parameter pair. Using an appropriate subset of data that has similar range and distribution of target outputs as the larger training set can drastically speed up the process of hyper-parameters tuning through cross-validation. In Figure 4.3, we show plots of the range and distribution of $q_1(\boldsymbol{\xi})$ from 1000 sample points and the whole training set. Scatter plots of contaminant concentration observations at q_i are shown in Figure 4.5.

Specifically, we first consider various pairs of (P, γ) values in which $P = 2^{-5}, 2^{-3}, ..., 2^{15}$, and $\gamma = 2^{-5}, 2^{-3}, ..., 2^{15}$ (see Figure 4.4) on a coarse grid. For each (P, γ), we quantify its quality by performing a 10-fold cross-validation. The 1000 sample points are divided into 10 subsets of equal size. We train the model based on 9 subsets, and treat the remaining subset as an "unknown" set. The mean square error (MSE) can be computed on the "unknown" set to measure the quality of the prediction,

$$\text{MSE} = \frac{1}{\bar{l}} \sum_{i=1}^{\bar{l}} (f(\boldsymbol{x}_i) - y_i)^2, \tag{4.9}$$

where $\bar{l}$ is the number of samples in the "unknown" set. The procedure is repeated 10 times until each subset has been predicted once. The average of the 10 resulting MSE estimates indicates how accurate the model can predict unknown data. The pair that leads to the highest cross-validation accuracy (the smallest value of the average MSE) is found in Table 4.1. We then repeat the search process on a fine grid in the neighborhood of the optimal (P, γ) obtained

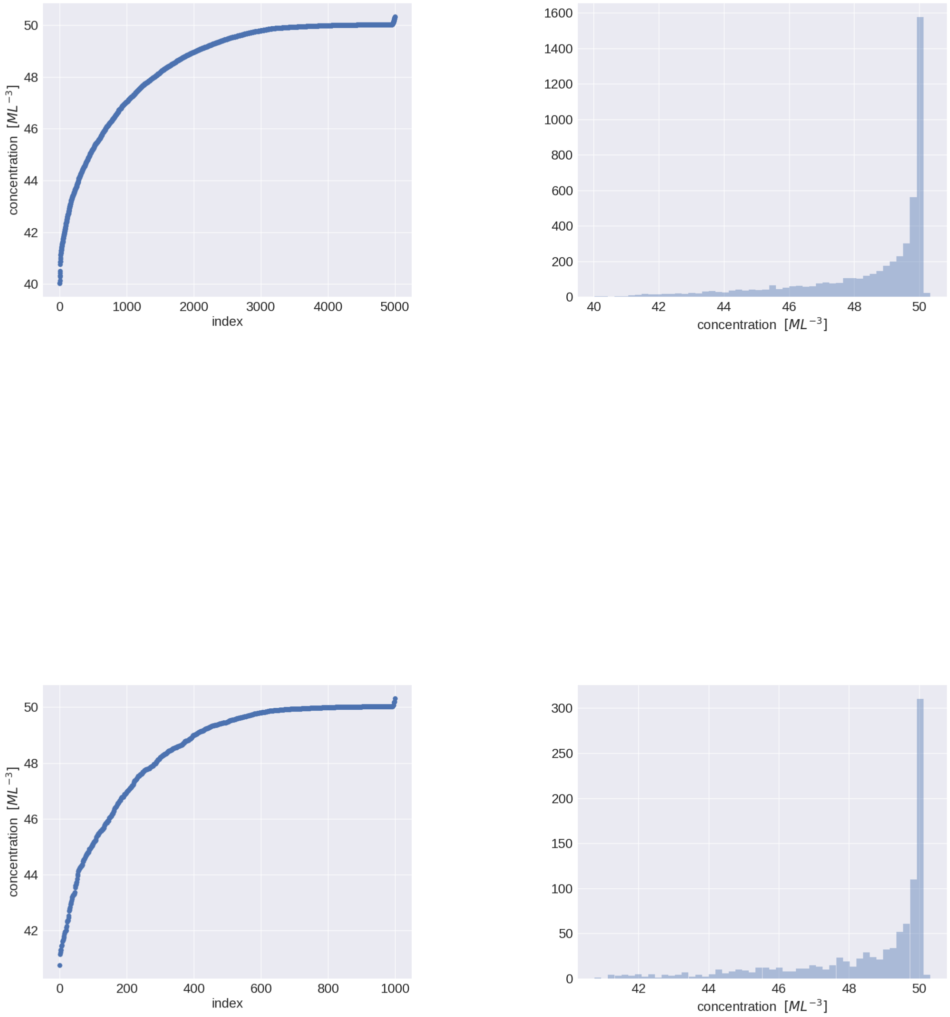


Figure 4.3: Range and distribution of q_1. Dataset of 5000 samples (top) and dataset of 1000 samples (bottom).

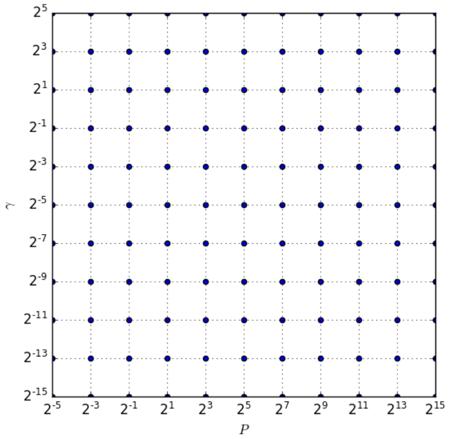
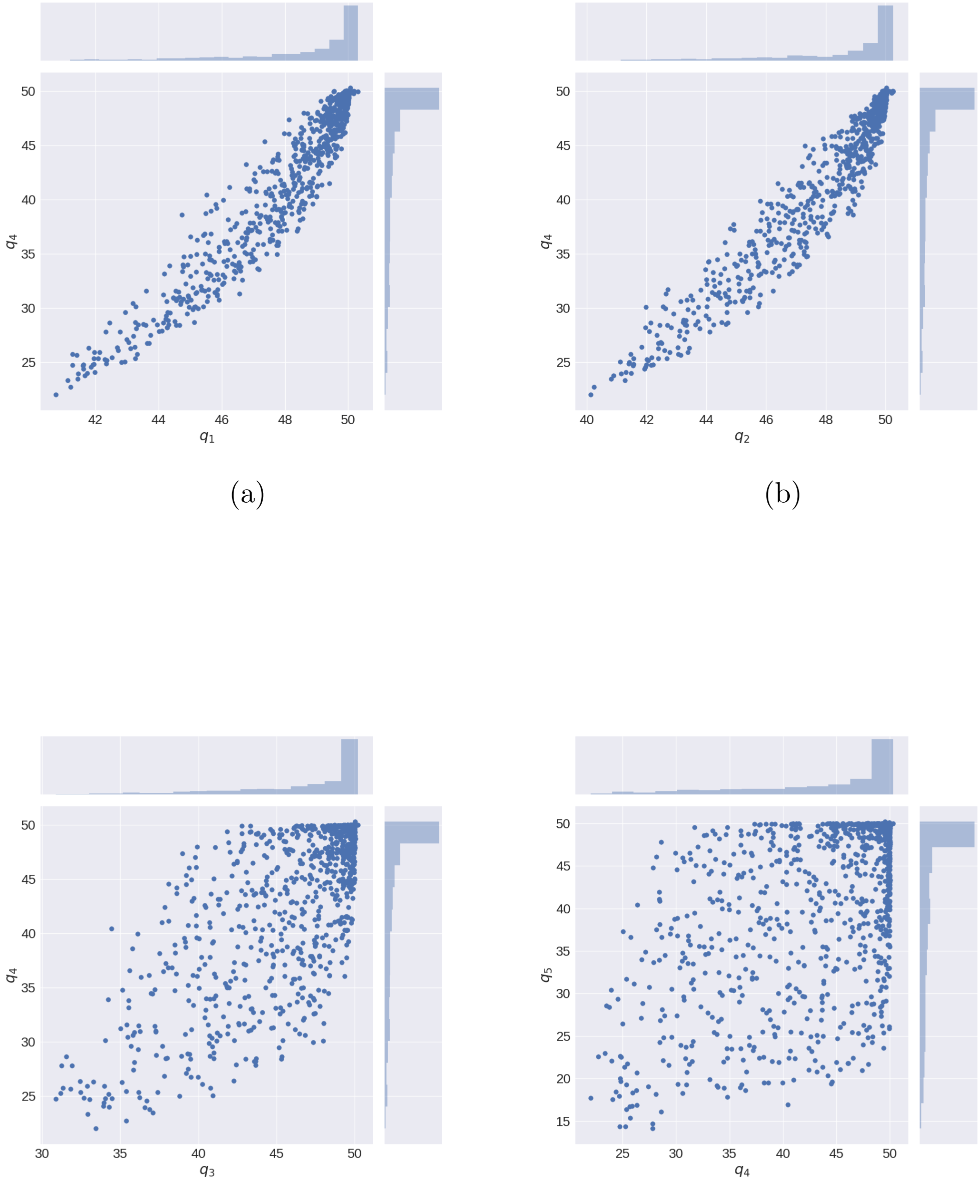


Figure 4.4: Illustration of grid search on (P, γ). Pairs of values at the grid nodes are tried, and the one with the best cross-validation accuracy is picked to train the whole training set.

from the previous coarse grid search. For example, the grid-search is performed in the region of $P = 2^{11}, 2^{11.5}, ..., 2^{13}, ..., 2^{15}$, and $\gamma = 2^{-5}, 2^{-4.5}, ..., 2^{-3}, ..., 2^{-1}$ to determine the optimal parameters for the SVM to approximate the relation between $q_1(\boldsymbol{\xi})$ and $\boldsymbol{\xi}$. The optimal hyper-parameters are listed in Table 4.2.

Before we build surrogate models on the whole training set including 5000 data points, we plot learning curves for a sanity check on the training set size. Learning curves plot the prediction accuracy on training and validation set against the training set size to show how the model improves at predicting the target output as we increase the number of sample points in the training set. This helps diagnose whether the model suffers from high bias or variance, and tells whether more training points will help in improving the model performance on prediction. If two curves converge at a low accuracy, a predictive

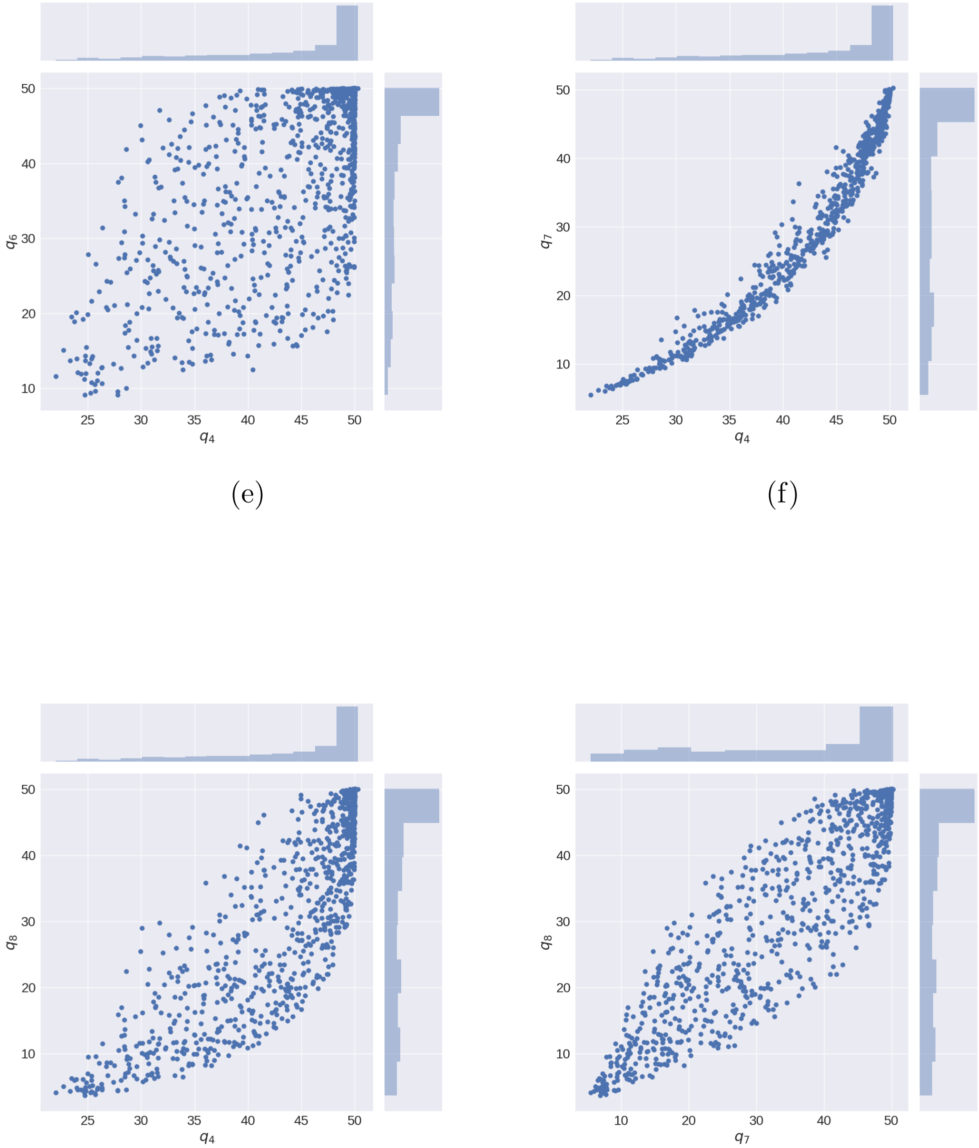

q_4
50
45
40
35
30
25
42
44
46
48
50
q_1
q_4
50
45
40
35
30
25
40
42
44
46
48
50
q_2
q_4
50
45
40
35
30
25
30
35
40
45
50
q_3
q_5
50
45
40
35
30
25
20
15
25
30
35
40
45
50
q_4

(a)

(b)

(c)

(d)



(e)

(f)

(g)

(h)

Figure 4.5: Scatter plots of computed concentrations at pairs of q_i

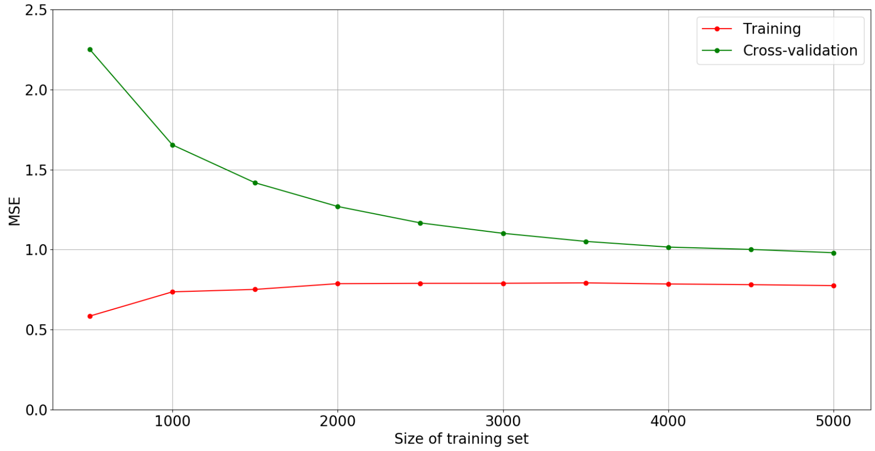


Figure 4.6: Learning curves of the SVM surrogate for $q_4(\boldsymbol{\xi})$.

model is underfitting and is unable to capture the relationship between the input and target output. Adding more training data is not helpful in this case. A more complex model is needed. If the model performs well on the training data but poorly on the validation set, i.e., there is a large gap between the two learning curves, it is overfitting. In other words, the model memorizes the data it has seen but doesn't generalize for unseen data. We shuffle and split the whole dataset 10 times into training and validation data in the ratio of 4 to 1. Subsets of the training set with varying sizes are used to train the SVM with the hyper-parameters in Table 4.2, and MSE for each training subset size and the validation set are computed. The MSE is then averaged over all 10 runs for each training subset size. In Figure. 4.6, we show the learning curves of the surrogate model for q_4. When the training set is small, the training

Table 4.1: Optimum (P, γ) for RBF kernels from coarse grid search

	P	γ	MSE	R^2
q_1	2^{13}	2^{-3}	0.185761369	0.957232
q_2	2^{13}	2^{-3}	0.242837426	0.952064
q_3	2^{13}	2^{-3}	0.628202705	0.966536
q_4	2^{7}	2^{-1}	1.86252441	0.964607
q_5	2^{9}	2^{-1}	3.88174606	0.957161
q_6	2^{13}	2^{-3}	4.27258069	0.967153
q_7	2^{9}	2^{-1}	5.50957803	0.971784
q_8	2^{13}	2^{-3}	5.25627116	0.975476
$q_{prediction}$	2^{13}	2^{-3}	2.27758075	0.975397

error is small too. As the training set size grows, the training error slowly increases but still remains low. On the other hand, the validation error is high due to overfitting when the model is trained on a small training set and does not generalize. Also, the large gap between training error and validation error indicates that the model trained with the given hyper-parameters in Table 4.2 exhibits high variance. In this situation, using more training points is helpful to reduce high variance. However, the validation error starts to level off when the training set size is around 4000. Including more points for training will further reduce the validation error slightly at the expense of longer training time. Therefore, in this work, we use 5000 training points to create reliable surrogate models in low computational time.

Table 4.2: Optimum (P, γ) for RBF kernels from fine grid search

	P	γ	MSE	R^2
q_1	2^{12}	2^{-3}	0.179665488	0.960749
q_2	2^{12}	$2^{-2.5}$	0.238587029	0.953566
q_3	$2^{14.5}$	$2^{-3.5}$	0.626073890	0.968131
q_4	$2^{8.5}$	$2^{-1.5}$	1.76103827	0.967289
q_5	$2^{9.5}$	2^{-1}	3.69977373	0.961253
q_6	$2^{14.5}$	$2^{-3.5}$	4.23652222	0.969461
q_7	2^{8}	2^{-1}	5.22852799	0.973601
q_8	2^{12}	2^{-3}	5.18621775	0.976414
$q_{prediction}$	2^{12}	$2^{-2.5}$	2.20208775	0.976066

Chapter 5

Measure-Theoretic Framework[1]

Different types of inverse problems may be formulated under various physical and statistical assumptions on model parameters. In this chapter, we use a set-approximation method to solve the stochastic inverse problem which is formulated within a measure-theoretic framework. We consider a deterministic model where the dimension of the observable output is smaller than that of the model input parameters. The corresponding inverse problem then has set-valued solutions. We present a numerical method to approximate the set-valued solutions of probability measure of model input, given an assumed probability distribution on the observations.

As mentioned in Chapter 1, we focus on constructing pullback and push-forward probability measures through the surrogate defined by the SVM. By not assuming prior distributions or likelihoods, the quality of computing such probability measures is solely dependent upon the global accuracy of the SVM and its ability to propagate probabilistic events accurately.

[1] This chapter is based on the article entitled *Data-driven uncertainty quantification for predictive flow and transport modeling using support vector machines* by Jiachuan He, Steven Mattis, Troy Butler and Clint Dawson [32].

5.1 Pullback and push-forward measures

We briefly describe pullback and push-forward probability measures using the notation of the previous sections. For a more thorough discussion of pullback measures including a discussion of existence and uniqueness, we direct the interested reader to *A Measure-Theoretic Computational Method for Inverse Sensitivity Problems III: Multiple Quantities of Interest* [10] and the references therein.

Let $\mathcal{D} = Q(\Xi)$ denote the range of the parameter-to-observable map and $P_{\mathcal{D}}$ a probability measure defined on $\mathcal{D}$. In practice, this probability measure may be obtained by either a statistical analysis of measured data, engineering knowledge of the uncertainty in measured data, or imposed as part of an engineering design (e.g., representing worst-case scenario analysis or desired responses assuming some level of control/intervention of the model parameters Ξ). Once $P_{\mathcal{D}}$ is specified, a pullback measure P_{Ξ} on Ξ is any measure satisfying the (consistency) condition,

$$P_{\Xi}(Q^{-1}(A)) = P_{\mathcal{D}}(A), \tag{5.1}$$

for every event A in $\mathcal{D}$. Oftentimes, these probability measures are described as densities ρ_{Ξ} and $\rho_{\mathcal{D}}$ on Ξ and $\mathcal{D}$, respectively, and consistency takes the form of

$$P_{\Xi}(Q^{-1}(A)) = \int_{Q^{-1}(A)} \rho_{\Xi} d\mu_{\Xi} = \int_{A} \rho_{\mathcal{D}} d\mu_{\mathcal{D}} = P_{\mathcal{D}}(A), \tag{5.2}$$

for every event A in $\mathcal{D}$, where μ_{Ξ} and $\mu_{\mathcal{D}}$ describe (volume) measures on Ξ and $\mathcal{D}$, respectively.

In general, there is not a unique pullback measure P_Ξ since the consistency condition only requires specification of this measure on events $Q^{-1}(A)$ within Ξ. Thus, unless Q is a bijection between Ξ and $\mathcal{D}$, for any event A in $\mathcal{D}$, we are free to make certain choices on how P_Ξ is evaluated on subsets of $Q^{-1}(A)$. In Figure 5.1, we use a general two-to-one map as an example. If Q is a mapping from $\Lambda \subset \mathbf{R}^2$ to $\mathcal{D} \subset \mathbf{R}^1$, then through the inverse map there is a set of values, $Q^{-1}(Q(\lambda))$, in Ξ that are associated to a given value $Q(\lambda)$ where $\lambda \in \Lambda$. We call this inverse set a generalized contour. Any two points in the same generalized contour are equivalent (not distinguishable) as they correspond to the same value in $\mathcal{D}$. The space of equivalence classes imposed by Q^{-1} in Λ is denoted by $\mathcal{L}$, so that each point in $\mathcal{L}$ identifies a generalized contour. Therefore, Q^{-1} defines a bijection map between $\mathcal{L}$ and $\mathcal{D}$. To compute the probability measure of any event in Ξ, we can use the Disintegration Theorem to decompose it into measures in $\mathcal{L}$ and along generalized contours corresponding to points in $\mathcal{L}$. However, the latter is not available by inverting Q. An Ansatz needs to be incorporated to specify the probability measures along the contours. In order to test the global accuracy of the SVM defining Q, we use the standard Ansatz (see [10]) to proportion probabilities uniformly in directions of Ξ not informed by the map Q, hence resulting in a unique pullback measure P_Ξ.

Given any probability measure P_Ξ on Ξ, we may use the map Q to define a push-forward of this measure on $\mathcal{D}$ defined by

$$P_{\mathcal{D}}^{Q(\Xi)}(A) := P_\Xi(Q^{-1}(A)), \tag{5.3}$$

for every event A in $\mathcal{D}$. Comparing Equation (5.1) and Equation (5.3), we observe that we can easily check if a pullback measure P_Ξ was constructed by comparing $P_{\mathcal{D}}^{Q(\Xi)}$ with $P_{\mathcal{D}}$ on $\mathcal{D}$. Moreover, by considering other maps Q (e.g., corresponding to QoI to be predicted), we can use a pullback measure to easily construct other push-forward measures quantifying uncertainties in predictions.

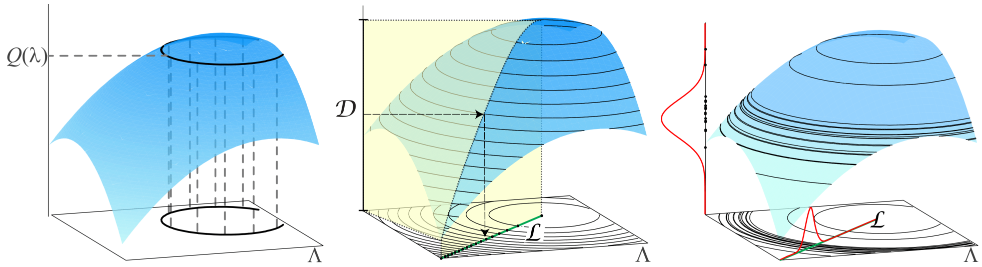


Figure 5.1: Illustrations of the inverse problem for a general two-to-one map Left: The set-valued inverse of a single output value. Middle: The representation of $\mathcal{L}$ as a transverse parameterization. Right: A probability measure described as a density on $\mathcal{D}$ maps uniquely to a probability density on $\mathcal{L}$. Figures adopted from [9]

5.2 Numerical construction of pullback and push-forward measures

Forward UQ problems involving the construction of push-forward measures are well-studied and the measures are typically approximated using Monte Carlo or other sampling schemes. In Algorithm 1, we summarize a basic sampling scheme for approximating a pullback measure with the standard Ansatz first introduced in [10]. The output of Algorithm 1 is an array of probabilities $\{p_{\Xi,j}\}_{j=1}^{N}$ associated with each sample $\{\boldsymbol{\xi}^{(j)}\}_{j=1}^{N} \in \Xi$. Using this array of probabilities, we can approximate the probability of any event A in Ξ using a counting measure

$$P_{\Xi}(A) \approx P_{\Xi,N}(A) := \sum_{\boldsymbol{\xi}^{(j)} \in A} p_{\Xi,j}. \tag{5.4}$$

Thus, we obtain an approximation to the pullback probability measure on Ξ. This algorithm is implemented within the BET software package [29]. BET stands for Butler Estep Tavener method.

In Algorithm 1, we approximate events, implicitly, with finite collections of Voronoi tessellations of Ξ. The error of implicit Voronoi approximations of $Q^{-1}(D_k)$ in Step 5 of the algorithm due to finite sampling effects the counting measure estimates. Increasing the number of samples is one of the approaches to reduce the error as shown in Figure 5.2, and with a sufficiently large number of i.i.d. samples, we often use the Monte Carlo approximation in Step 7 of the algorithm that $V_j = \mu_{\Xi}(\Xi)/N$ (i.e., each Voronoi cell is approximated to have the same volume). However, errors in the SVM can lead

Algorithm 1: Numerical Approximation of a Pullback Measure

```
1. Choose samples $\{\boldsymbol{\xi}^{(j)}\}_{j=1}^{N} \in \Xi$ implicitly defining a Voronoi tessellation $\{\mathcal{V}_j\}_{j=1}^{N} \subset \Xi$.
2. Evaluate $Q^{(j)} = Q(\boldsymbol{\xi}^{(j)})$ for all $\boldsymbol{\xi}^{(j)}$, $j = 1, .., N$.
3. Choose a partitioning of $\mathcal{D}$, $\{D_k\}_{k=1}^{M} \subset \mathcal{D}$. Refer to each $D_k$ as a bin.
4. Compute $p_{\mathcal{D},k} \approx P_{\mathcal{D}}(D_k)$ for $k = 1, ..., M$.
5. Let $\mathcal{C}_k = \{j | Q^{(j)} \in D_k\}$ for $k = 1, ..., M$ denote a pointer indicating the subset of $\{\mathcal{V}_j\}_{j=1}^{N}$ approximating $Q^{-1}(D_k)$.
6. Let $\mathcal{O}_j = \{k | Q^{(j)} \in D_k\}$, for $j = 1, .., N$ denote a pointer indicating where sample $Q^{(j)}$ is binned in $\mathcal{D}$.
7. Let $V_j$ be the approximate volume of $\mathcal{V}_j$, i.e. $V_j \approx \int_{\mathcal{V}_j} d\mu_{\Xi}(\mathcal{V}_j)$ for $j = 1, .., N$.
8. Set $p_{\Xi,j} = (V_j / \sum_{i \in \mathcal{C}_{\mathcal{O}_j}} V_i) p_{\mathcal{D},\mathcal{O}_j}$, $j = 1, .., N$.
```

to incorrect binning of samples in Step 6 of the algorithm. These errors subsequently impact both pointer $\mathcal{C}_k$ and $\mathcal{O}_j$ in Steps 5 and 6 of the algorithm, respectively. Such errors propagate directly to the array of computed probabilities in Step 8 of the algorithm.

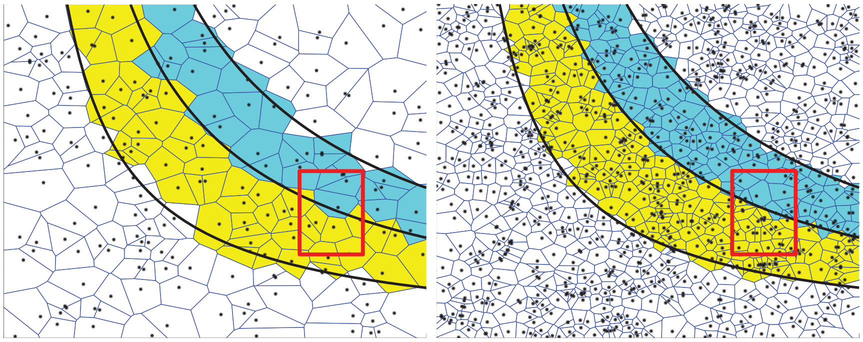


Figure 5.2: The error in the $\mu_\Xi-$volume of a Voronoi coverage of $Q^{-1}(D_k)$ affects P_Ξ estimation. For any fixed partitioning of $\mathcal{D}$, $P_{\Xi,N}$ converges to P_Ξ as $N \to \infty$.

5.3 Comparison to Bayesian posterior and computational complexity

Here, we use some simplifying assumptions in order to provide a reasonable comparison between the solutions and computational complexity for inverse problems formulated in either the measure-theoretic or Bayesian frameworks. We first assume that there are no hyperparameters used in the definitions of the prior distribution for the Bayesian formulation and that this prior is also used to formulate the Ansatz for distributing probabilities along the contour events in the measure-theoretic formulation. If we further assume that the likelihood function is defined in such a way that it matches the distribution we use to invert for the same parameter-to-observables map, then the solutions to either problem formulation are probability measures on the (same) parameter space that have the same conditional probability distributions on the generalized contours of the parameter-to-observables map. However, the probability measures will still be different in directions (locally) orthogonal to the generalized contours. This difference is due to the influence of the prior distribution in the Bayesian setting in all directions of parameter space including those directions informed by the data, which is not the case in the measure-theoretic approach. See [11] for a simple 1-D example highlighting this difference, which emphasizes the fact that the Bayesian formulation is not attempting to construct a pullback measure. In other words, even when the setup of the problems are effectively identical in either formulation, the actual problem being solved is based on fundamentally different perspectives so that the solutions have different structures.

To compare the computational complexities, assume that the goal of generating samples from the Bayesian posterior distribution is to approximate probabilities of events. If a Monte Carlo sampling scheme is used in a Bayesian framework, then the convergence of the Monte Carlo estimates of probabilities of events is subject to the well-known Central Limit Theorem. Algorithm 1 can be interpreted as a Monte Carlo approximation to probabilities of events if samples in Step 1 are drawn in the parameter space according to a prior density (although a prior density is not necessary to apply Algorithm 1). The convergence then follows from results in stochastic geometry that rely on the

Strong Law of Large Numbers [12], which is a key result used in proving the Central Limit Theorem. In other words, the rates of convergence for either method using similar statistical tools for generating random samples should generally be similar in practice. Since the measure-theoretic approach involves estimation of the discretized contour events, it is possible to define non-random sets of samples to reduce errors in probability for any contour event to a desired level of accuracy [13]. We note that much work has been done over the last decade in accelerating sampling of the Bayesian posterior using methods primarily based upon Markov Chain Monte Carlo sampling. Leveraging such sampling approaches for the measure-theoretic inversion is the topic of future research.

Chapter 6

Numerical Examples[1]

In this chapter, a probability measure on the observable data is specified to account for measurement uncertainty, and a pullback probability measure on the parameters that characterizes hydraulic conductivity is constructed. In the first three examples, contaminant concentration is used as observation. In Example 4 and 5, we derive hydraulic conductivity by conditioning on hydraulic head data. The estimated hydraulic conductivity is then used in models to predict head or contaminant concentration values where measurements are not available. In these examples, we draw samples from the parameter domain and solve the physics-based models with hydraulic conductivity fields characterized by those samples to obtain training sets. The evaluations of physics-based models are replaced by SVMs which are learned on the training sets.

[1] This chapter is based on the article entitled *Data-driven uncertainty quantification for predictive flow and transport modeling using support vector machines* by Jiachuan He, Steven Mattis, Troy Butler and Clint Dawson [32].

6.1 Construction of a pullback measure from concentration observation

Example 1: In this example, we construct and analyze a pullback measure with the SVM given in Chapter 4. We describe the impact of Steps 5 and 6 in Algorithm 1 in identifying highly probable, but spatially disparate, hydraulic conductivity fields. This demonstrates how the pullback measure is defined globally on Ξ in terms of inverse sets that may stretch across large portions of Ξ.

A realization of 9-term truncated KLE is randomly chosen to be the reference field considered as the true underlying hydraulic conductivity field (see Figure 6.1). We solve the flow and contaminant transport models based on the reference field using the discretization described in Chapter 4. The simulated contaminant concentration at time $T = 2$ at the measurement locations marked q_i for $i = 1, 2, \ldots, 8$ in Figure 4.2 yields the reference output data (observations) $Q_{obs} = [\hat{q}_1, \hat{q}_2, \ldots, \hat{q}_8]$ given by the vector

$$[44.95, 45.37, 37.58, 31.35, 24.39, 16.79, 11.96, 8.06],$$

where $\hat{q}_i$ denotes the concentration datum at the observation location q_i for $i = 1, 2, ..., 8$. A probability measure $P_{\mathcal{D}}$ on $\mathcal{D}$ is then defined in terms of a multivariate normal probability density function $\rho_{\mathcal{D}}$ centered at Q_{obs} with a standard deviation of $0.01 \times Q_{obs}$ to reflect the measurement uncertainty in the observation data. In other words, we take the measurement uncertainty as a Gaussian distribution over all possible values that could be attributed to

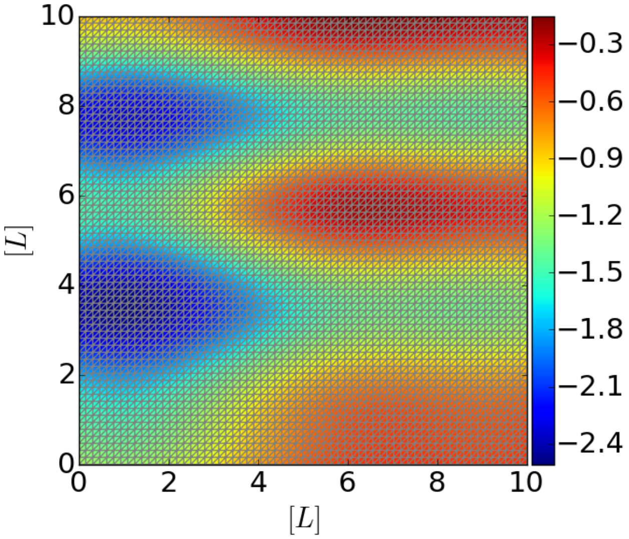
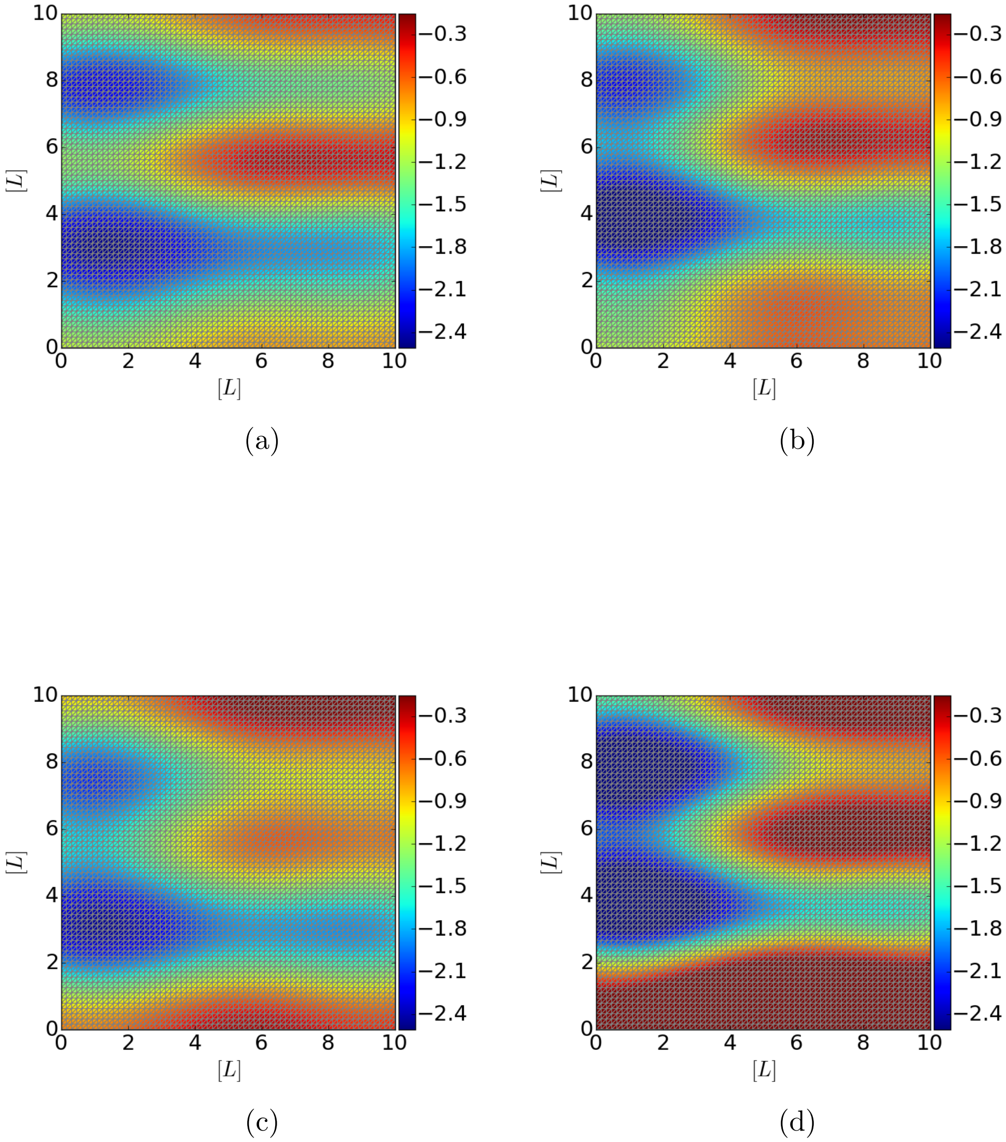


Figure 6.1: The reference $\ln K$ field approximated by a truncated KLE with 9 terms.

the uncertain concentration measurement. The relative measurement uncertainty, the measurement uncertainty divided by the single absolute value of the measured concentration, is 0.01 here.

Given $\boldsymbol{\xi}$, $q_i(\boldsymbol{\xi})$ for $i = 1, 2, ...8$ can be efficiently evaluated by the SVM using Equation (4.6) based on 5000 training points. Then, 5×10^5 points in Ξ are evaluated using the SVMs based on the hyper-parameters in Table 4.2. We apply the 5×10^5 samples in Algorithm 1 to approximate P_Ξ. We order all of the samples by probability that is associated to each implicitly defined Voronoi cell. In Figure 6.2 and Figure 6.3, we show some samples from the region of the highest probability in the parameter space. It is clear that sample (a) in Figure 6.2 yields a hydraulic conductivity field that exhibits a pattern very similar to the reference field. Since the solution to the deterministic inverse problem



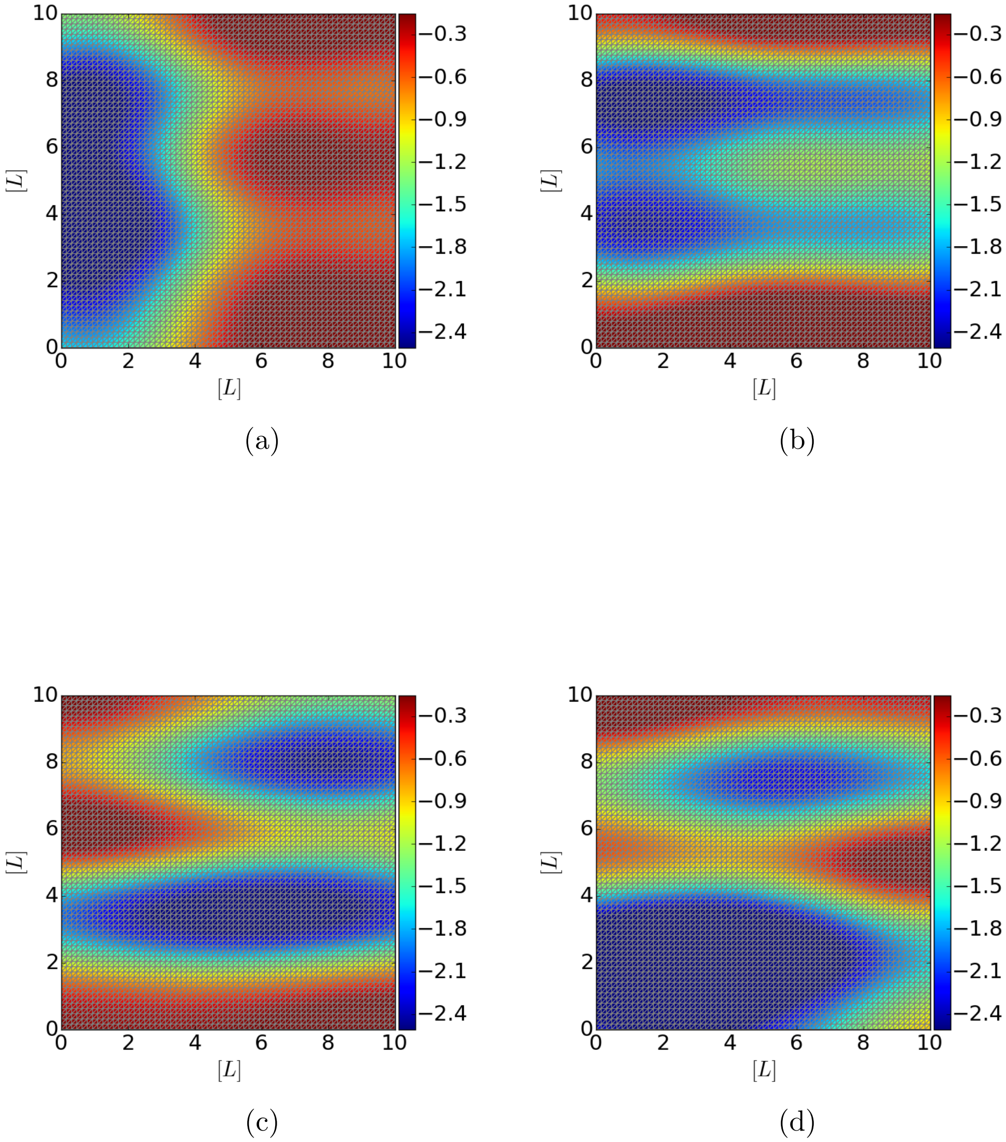
(a)

(b)

(c)

(d)

Figure 6.2: Samples of $\ln K$ from Voronoi cells with highest probability that are qualitatively similar to the reference $\ln K$



(a)

(b)

(c)

(d)

Figure 6.3: Samples of $\ln K$ from Voronoi cells with highest probability that are qualitatively different from the reference $\ln K$

is inherently set-valued, many hydraulic conductivity fields result in the same or similar simulated values of measured contaminant concentrations. As a result, although samples shown in Figure 6.3 completely misrepresent the truth, they correspond to $Q^{(j)}$ samples that are binned in a region of high probability according to Algorithm 1. However, we in general have limited knowledge of what the truth is, and unless more data or domain specific knowledge can rule out the fields represented by samples in Figure 6.3, they should be used in constructing push-forward measures and constructing conservative predictions. The rest of the samples in Figure 6.2 capture the major features of the reference profile, but they over/under predict the conductivity in some area. It may be possible to use this non-parametric probability measure to define a physically informed prior density in a Bayesian setting to further localize the probability to small ranges, but this is beyond the scope of this work.

6.2 Validation with push-forward measure

Example 2: In this example, we verify the convergence of the pullback measures and validate these results by assessing how well the pullback measures can be used to predict unknown data by leave-one-out validation. Specifically, we use the observation from seven wells to "predict" the concentration at the remaining observation well. A realization of a 100-term truncated KLE is used to represent a more realistic reference hydraulic conductivity field (Figure 6.5) for practical applications. However, to demonstrate how well the SVM constructed on low-dimensional representations of hydraulic conductivity performs, we use the same SVM based on the 9-term truncated KLE described in Chapter 4, which only retains about 70% of the reference variance in the infinite-term expansions of the hydraulic conductivity fields. Thus, with the same model setup as the previous example except for the reference K, we solve the flow and contaminant transport models to generate the observable contaminant concentrations at $T = 2$, $Q_{obs} = [\hat{q}_1, \hat{q}_2, \hat{q}_4, \hat{q}_5, \hat{q}_6, \hat{q}_7, \hat{q}_8]$ given by the vector

$$[49.71, 49.77, 47.40, 45.88, 42.76, 40.08, 35.88],$$

in the seven wells (note that $\hat{q}_3$ is not used). We again assume that ρ_D is a multivariate normal distribution with mean at Q_{obs} and standard deviation of $0.01 \times Q_{obs}$. To verify the convergence of the pullback measures, we construct surrogate models for $q_i(\boldsymbol{\xi})$ based on training set of N_t training points, where $N_t = 1000, 2000, 3000, 4000, 5000$. We use 5×10^5 i.i.d. sample points in Algorithm 1 to compute the corresponding pullback probability measure on the 9-dimensional parameter space, $P^{N_t}_{\Xi, 5\times10^5}$. We estimate the change in the probability measures in terms of total variation by

$$d(P^{N_t}_{\Xi,5\times10^5}, P^{N'_t}_{\Xi,5\times10^5}) = \sum_{i=1}^{5\times10^5} |P^{N_t}_{\Xi}(\boldsymbol{\xi}^{(i)}) - P^{N'_t}_{\Xi}(\boldsymbol{\xi}^{(i)})|. \tag{6.1}$$

The total variation is a metric that ranges from 0 to 2. It is 0 if two probability measures are idential; it is 2 if the probability measures have disjoint supports.

In Table 6.1, we verify the convergence of the pullback probability measures computed from SVMs with increasing training set sizes by computing

Table 6.1: Total variations of pullback probability measures for example 2.

$d(P^{1000}_{\Xi,5\times10^5}, P^{5000}_{\Xi,5\times10^5})$	0.9673
$d(P^{2000}_{\Xi,5\times10^5}, P^{5000}_{\Xi,5\times10^5})$	0.7291
$d(P^{3000}_{\Xi,5\times10^5}, P^{5000}_{\Xi,5\times10^5})$	0.5776
$d(P^{4000}_{\Xi,5\times10^5}, P^{5000}_{\Xi,5\times10^5})$	0.4221

the total variation between these measures and the pullback probability measure constructed for the SVM with the full 5000 training samples. Note that the total variation monotonically decreases as the number of samples in the training set monotonically increases.

We next validate the results by constructing the push-forward probability measure on the space of predictions, $q_3(\boldsymbol{\xi})$. This prediction of the push-forward probability measure on q_3 is obtained by weighting simulated concentrations of $q_3(\boldsymbol{\xi})$ using the SVM surrogate at the 5×10^5 samples of Ξ where the weights come from the pullback probability measure $P^{5000}_{\Xi,5\times10^5}$. We approximate the probability density function of the contaminant concentration at q_3 with the weighted histogram in Figure 6.4. The "true" observation (corresponding to the full 100-term reference KLE of the hydraulic conductivity field) lies within the highest probability region of the predicted push-forward distribution. This indicates that the low-dimensional representation of hydraulic conductivity can be used to define accurate SVM surrogates for both probabilistic inversion and prediction.

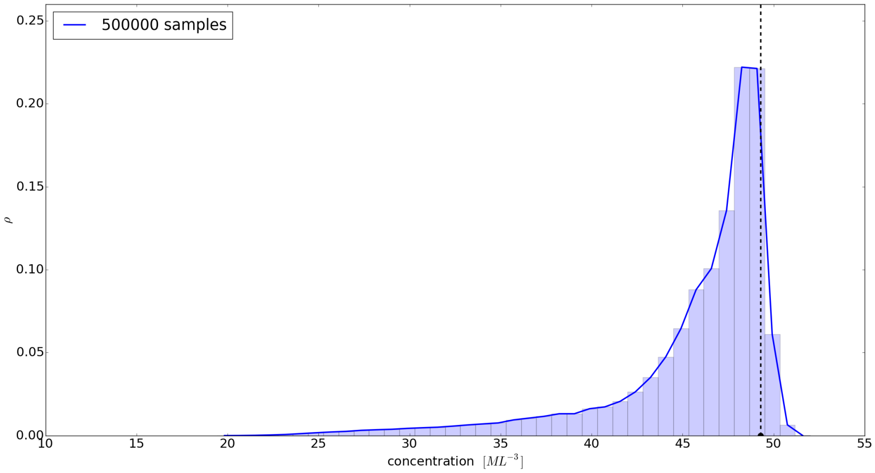


Figure 6.4: Predicted probability density function of concentration at q_3 using 500000 samples. The true observation at q_3 is illustrated by a black dot on the x-axis.

6.3 Concentration data-to-parameter-to-concentration prediction

Example 3: In this example, we consider the goal of estimating the contaminant concentration in the domain where measurements are unavailable, which is often needed for subsurface contaminant remediation or resources management. The model setup is the same as the previous example except we use the full observation vector $Q_{obs} = [\hat{q}_1, \hat{q}_2, \ldots, \hat{q}_8]$ given by

$$[49.71, 49.77, 49.28, 47.34, 45.88, 42.76, 40.08, 35.88],$$

along with the entire set of observations as the QoI map, i.e., $Q(\boldsymbol{\xi}) = [q_1(\boldsymbol{\xi}), q_2(\boldsymbol{\xi}), \ldots, q_8(\boldsymbol{\xi})]$ for all $\boldsymbol{\xi} \in \Xi$. The same 5×10^5 i.i.d. sample points are used in Algorithm 1 to compute the probability measure on the 9-dimensional pa-

Table 6.2: Total variations of pullback probability measures for example 3.

$d(P^{1000}_{\Xi,5\times10^5}, P^{5000}_{\Xi,5\times10^5})$	0.8708
$d(P^{2000}_{\Xi,5\times10^5}, P^{5000}_{\Xi,5\times10^5})$	0.5584
$d(P^{3000}_{\Xi,5\times10^5}, P^{5000}_{\Xi,5\times10^5})$	0.4718
$d(P^{4000}_{\Xi,5\times10^5}, P^{5000}_{\Xi,5\times10^5})$	0.3345

rameter space. Table 6.2 shows the how the total variation of this probability measure converges numerically as the number of training points is increased. We then propagate the probability measure on Ξ to define a push-forward probability measure at $q_{prediction}$ in Figure 4.2. Specifically, we draw the same training set of 5000 sample points in Ξ to solve the groundwater flow and transport models for the concentrations $q_{prediction}(\boldsymbol{\xi})$ at the prediction location in Figure 4.2. An SVM surrogate model is constructed on $q_{prediction}$ using this training set. The remaining 5×10^5 samples are evaluated using the SVM. We approximate the probability density of the contaminant concentration at the prediction location with the weighted histogram in Figure 6.6 where we agian show the reference "true" value of $q_{prediction}$ based on the reference 100-term KLE representation of hydraulic conductivity. The prediction quality of the push-forward measure suggests the SVM can be used to both construct reasonably accurate pullback measures and push-forward measures corresponding to prediction QoI.

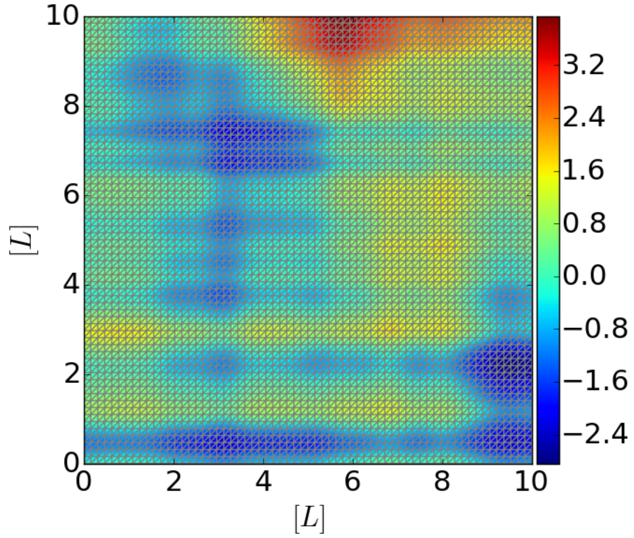


Figure 6.5: Contours of the more realistic reference $\ln K$ field approximated by a truncated KLE with 100 terms

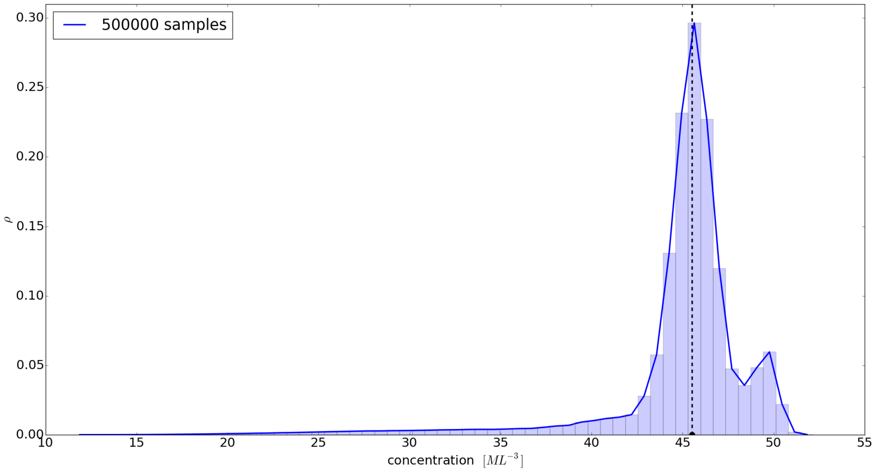


Figure 6.6: Predicted probability density of concentration at the prediction location using 500000 samples. The reference concentration is illustrated by a black dot on the x-axis.

6.4 Head data-to-parameter-to-head prediction

Example 4: Hydrogeologists often use field tests, e.g., pumping tests, to characterize an aquifer, evaluate well performance and identify aquifer boundaries. From the tests, the hydraulic conductivity can be estimated from the inversion of hydraulic head observations. In this example, we employ the measured hydraulic head values to characterize the saturated hydraulic conductivity field in the form of a 9-term KLE and use the computed measure on the parameter space to construct a push-forward measure for head where measurements are unavailable.

Assume instead of contaminant concentration we have hydraulic head measurements at the wells. Specifically, we use the same reference hydraulic conductivity field in Example 2 and Example 3 and generate hydraulic head observation by solving the groundwater flow model. Evaluating the solution at the well locations, we have $Q_{obs}^{head} = [\hat{q_1}, \hat{q_2}, \ldots, \hat{q_8}] = [14.44, 14.09, 13.80, 11.73, 12.27, 11.54, 10.63, 10.62]$. Similarly, we build SVMs to approximate the functions from $\boldsymbol{\xi}$ (input) to hydraulic head at each observation/prediction location, $q_i(\boldsymbol{\xi})$ (output) for $i = 1, \ldots, 8$. We use the 5000 training examples in the previous examples, except now the target values being simulated are hydraulic head data. To tune the hyper-parameters in the SVMs, we perform two level grid-search with 10-fold cross-validations on a subset of 1000 training examples. In Figure 6.8, we show plots of the range and distribution of hydraulic head values from 1000 sample points and the whole training set. Similar ranges and distributions of target outputs are observed. In Figure 6.7,

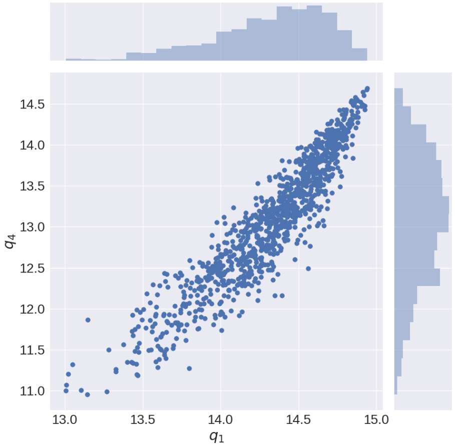
Table 6.3: Optimum (P, γ) for RBF kernels from coarse grid search for example 4

	P	γ	MSE	R^2
q_1	2^{11}	2^{-5}	0.006244	0.952978
q_2	2^{11}	2^{-5}	0.006114	0.951013
q_3	2^{11}	2^{-3}	0.011611	0.971642
q_4	2^{11}	2^{-3}	0.013963	0.976434
q_5	2^{13}	2^{-3}	0.011322	0.983352
q_6	2^{13}	2^{-3}	0.011187	0.981586
q_7	2^{9}	2^{-3}	0.011448	0.969877
q_8	2^{9}	2^{-5}	0.005692	0.954962
$q_{prediction}$	2^{11}	2^{-3}	0.013577	0.978973

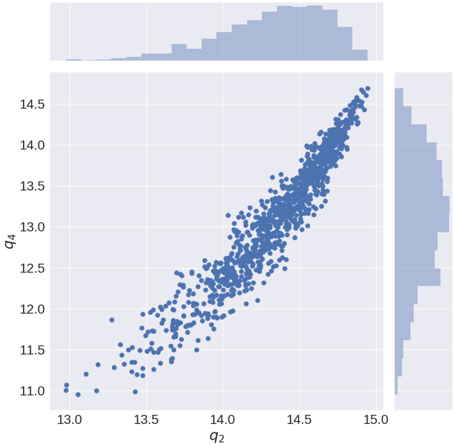
we show some scatter plots of the simulated output, q_i. The optimal hyperparameters obtained from coarse and fine grid search are listed in Table 6.3 and Table 6.4. Learning curves are plotted in Figure 6.9 to make sure 5000 data points are reasonable to form a training set.

Before we use all head observation from eight wells to construct a pullback measure and make head prediction at $q_{prediction}$, we perform verification and leave-one-out validation as before. $\hat{q}_3$ is left out for validation; assumptions that ρ_D is a multivariate normal distribution with mean at $Q_{obs} := [\hat{q}_1, \hat{q}_2, \hat{q}_4, \hat{q}_5, \hat{q}_6, \hat{q}_7, \hat{q}_8]$ and standard deviation of $0.01 \times Q_{obs}$ are made. In Table 6.5, we show the convergence of the pullback probability measures computed from SVMs trained on 1000, 2000, 3000, 4000 and 5000 training points. The total variation decreases as more training examples are used to build the SVMs.

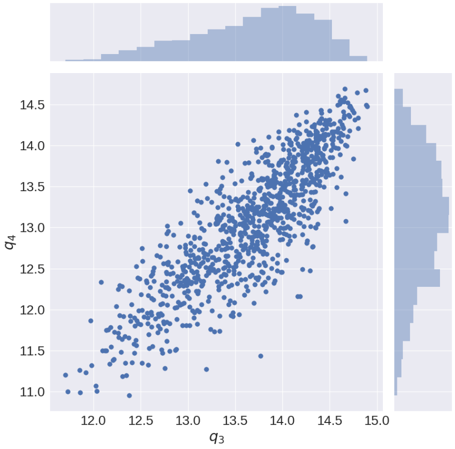
We approximate the probability density of the hydraulic head at $\hat{q}_3$


14.5
14.0
13.5
13.0
12.5
12.0
11.5
11.0
q_4
13.0
13.5
14.0
14.5
15.0
q_1

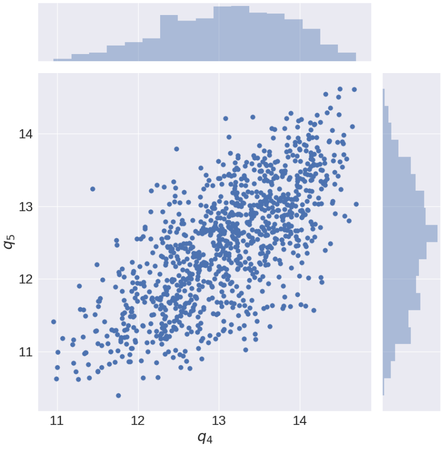
(a)


14.5
14.0
13.5
13.0
12.5
12.0
11.5
11.0
q_4
13.0
13.5
14.0
14.5
15.0
q_2

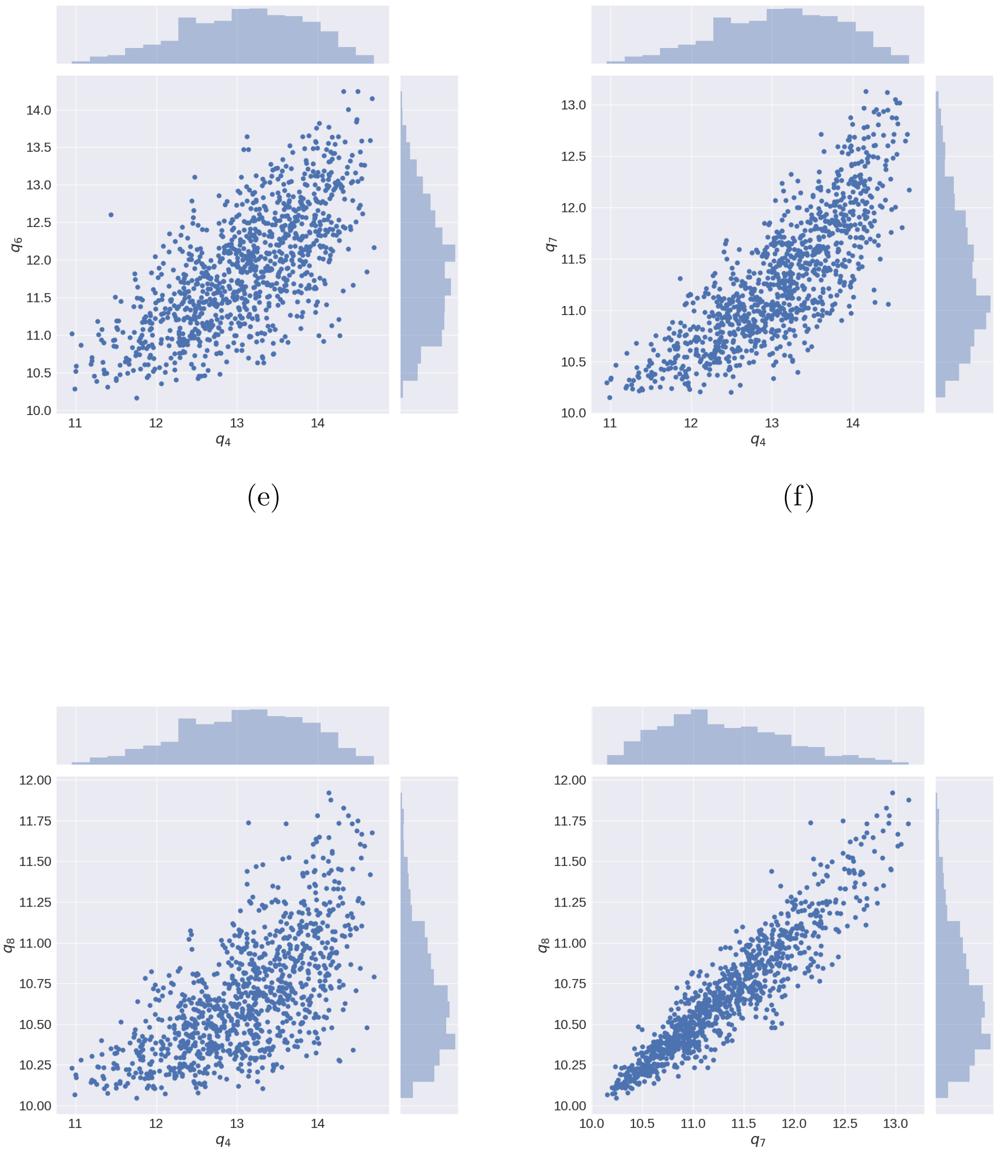
(b)


14.5
14.0
13.5
13.0
12.5
12.0
11.5
11.0
q_4
12.0
12.5
13.0
13.5
14.0
14.5
15.0
q_3

(c)


14
13
12
11
q_5
11
12
13
14
q_4

(d)



(e)

(f)

(g)

(h)

Figure 6.7: Scatter plots of computed concentrations at pairs of q_i

Table 6.4: Optimum (P, γ) for RBF kernels from fine grid search for example 4

	P	γ	MSE	R^2
q_1	2^{13}	$2^{-5.5}$	0.006168	0.953555
q_2	2^{11}	$2^{-6.5}$	0.006078	0.951237
q_3	$2^{12.5}$	$2^{-3.5}$	0.011276	0.972449
q_4	2^{13}	$2^{-3.5}$	0.013374	0.977410
q_5	2^{15}	2^{-4}	0.010575	0.984424
q_6	2^{15}	2^{-4}	0.010160	0.983221
q_7	2^{10}	$2^{-3.5}$	0.011093	0.970713
q_8	2^{11}	$2^{-5.5}$	0.005682	0.955030
$q_{prediction}$	2^{13}	2^{-4}	0.012480	0.980667

Table 6.5: Total variations of pullback probability measures for validation

$d(P^{1000}_{\Xi,5\times10^5}, P^{5000}_{\Xi,5\times10^5})$	0.7330
$d(P^{2000}_{\Xi,5\times10^5}, P^{5000}_{\Xi,5\times10^5})$	0.5985
$d(P^{3000}_{\Xi,5\times10^5}, P^{5000}_{\Xi,5\times10^5})$	0.4680
$d(P^{4000}_{\Xi,5\times10^5}, P^{5000}_{\Xi,5\times10^5})$	0.3606

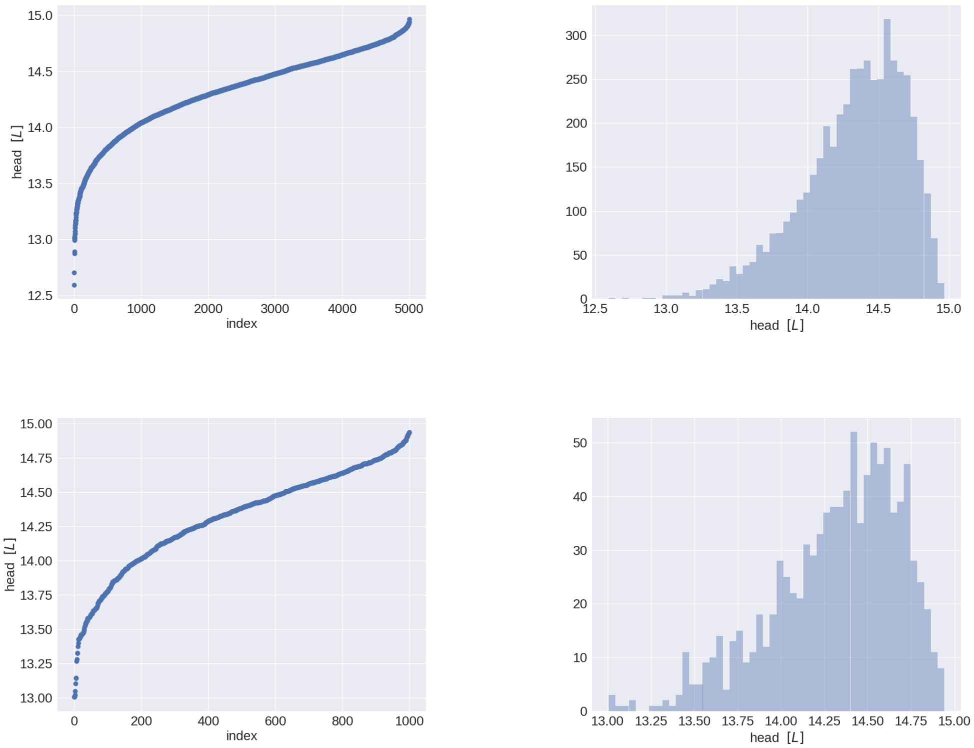


Figure 6.8: Range and distribution of hydraulic head at q_1. Dataset of 5000 samples (top) and dataset of 1000 samples (bottom).

with the weighted histogram in Figure 6.10. Given the observation reference value lies within the area close to the high probability region, we show that propagating the computed measure derived from head observations via the parameter-to-observable map offers solid prediction ability for hydraulic head too. However, the prediction quality is not as good as that for concentration in Example 2. One possible reason is hydraulic heads observed from wells that are close to the Dirichlet boundaries (q_1, q_2, q_8) are influenced more by the prescribed head condition than the underlying hydraulic conductivity.

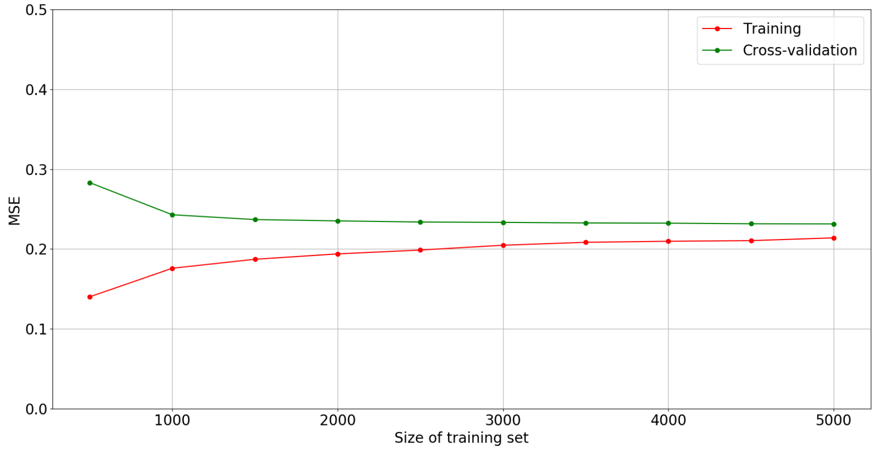


Figure 6.9: Learning curves of the SVM surrogate for hydraulic head data $q_4(\boldsymbol{\xi})$.

Prescribing Dirichlet boundaries leads to smaller sensitivities [15].

Table 6.6 shows the total variation of probability measure constructed from all eight head observations converges as the number of training points is increased to build the SVMs.

The prediction of the push-forward probability measure at $q_{prediction}$ is

Table 6.6: Total variations of pullback probability measures for hydraulic head prediction

$d(P^{1000}_{\Xi,5\times10^5}, P^{5000}_{\Xi,5\times10^5})$	0.8696
$d(P^{2000}_{\Xi,5\times10^5}, P^{5000}_{\Xi,5\times10^5})$	0.5336
$d(P^{3000}_{\Xi,5\times10^5}, P^{5000}_{\Xi,5\times10^5})$	0.4348
$d(P^{4000}_{\Xi,5\times10^5}, P^{5000}_{\Xi,5\times10^5})$	0.3331

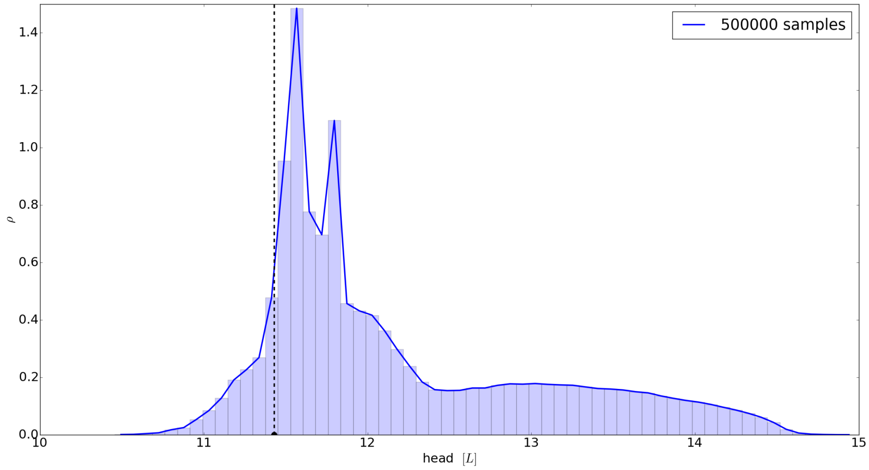


Figure 6.10: Predicted probability density function of head at q_3 using 500000 samples. The true observation at q_3 is illustrated by a black dot on the x-axis.

shown in Figure 6.11.

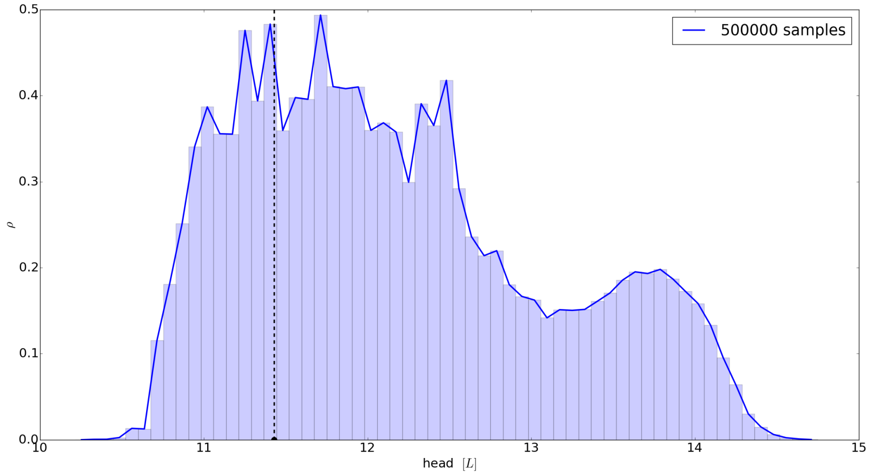


Figure 6.11: Predicted probability density of head at the prediction location using 500000 samples. The reference head is illustrated by a black dot on the x-axis.

6.5 Head data-to-parameter-to-concentration prediction

Example 5: We construct the push-forward probability measure on $q_{prediction}$ by weighting simulated concentrations of $q_{prediction}$ using the SVM surrogate at the 5×10^5 samples of Ξ where the weights are from the pull-back probability measure $P^{5000}_{\Xi,5\times10^5}$ which is now derived based on observable hydraulic head data. We approximate the probability density function of the contaminant concentration at the prediction location in Figure 6.12. It shows useful information that the push-forward probability measure localizes the predicted concentration to small ranges of values and the reference "true" value is within the high probability region. However, the predicted push-forward distribution has a bimodal shape. It suggests samples corresponding to the

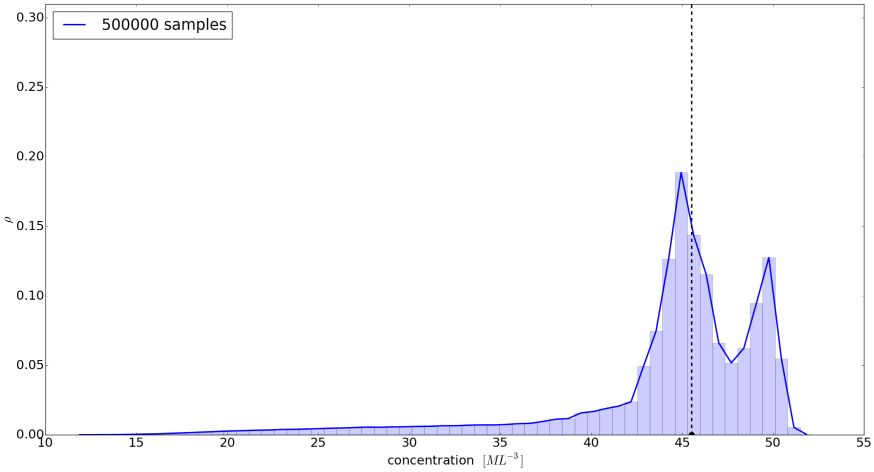


Figure 6.12: Predicted probability density of concentration at the prediction location using 500000 samples. The reference concentration is illustrated by a black dot on the x-axis.

other predicted peak at around $50[ML^{-3}]$ also have high probability due to the fact that evaluating those samples through the groundwater flow model generates head outputs close to the observation too.

Chapter 7

Conclusions[1]

In this work, given measured data from a limited number of observation wells, we used a framework based on intensive global sampling in the parameter space to infer the unknown spatially heterogeneous hydraulic conductivity field and predict the concentration or hydraulic head at other locations. We constructed SVM surrogate models for improved computationally efficiency in sampling parameter-to-observable responses of flow and transport models. The examples demonstrated that the SVM can be constructed on a relatively low-dimensional truncation of a KLE to replace the full flow and transport model solves within the measure-theoretic framework. Useful pullback and push-forward probability measures can be computed to illuminate the unknown model parameter and predict the model state. This suggests the SVM surrogate modeling technique based on statistical learning theory is promising for many UQ problems that involve either global or local propagations of uncertainties.

In this work, a full error analysis was not addressed in favor of a more

[1] This chapter is based on the article entitled *Data-driven uncertainty quantification for predictive flow and transport modeling using support vector machines* by Jiachuan He, Steven Mattis, Troy Butler and Clint Dawson [32].

qualitative analysis of results in terms of inferring correct parameter values or predicting certain ranges of quantities of interest with high probability. A future work will investigate more thoroughly the numerical error in model simulations defining the samples used to construct the SVM and the approximation error of the SVM itself, which impacts the accuracy of all samples. We will investigate adjoint techniques to both estimate and correct the error in individual samples and perform local sensitivity analyses. The gradient of concentration with respect to each input parameter sample can be computed with adjoint techniques to gain more physics-information from the models. This can be potentially used to enhance the local approximation properties of the data-driven surrogate models by incorporating this knowledge into hyper-parameters used in the SVM.

Bibliography

[1] Sajjad Ahmad, Ajay Kalra, and Haroon Stephen. Estimating soil moisture using remote sensing data: A machine learning approach. *Advances in Water Resources*, 33(1):69–80, 2010.

[2] LR Ahuja, DK Cassel, RR Bruce, and BB Barnes. Evaluation of spatial distribution of hydraulic conductivity using effective porosity data. *Soil Science*, 148(6):404–411, 1989.

[3] Martin Alns, Jan Blechta, Johan Hake, August Johansson, Benjamin Kehlet, Anders Logg, Chris Richardson, Johannes Ring, Marie Rognes, and Garth Wells. The FEniCS Project Version 1.5. *Archive of Numerical Software*, 3(100), 2015.

[4] Todd Arbogast, Steve Bryant, Clint Dawson, Fredrik Saaf, Chong Wang, and Mary Wheeler. Computational methods for multiphase flow and reactive transport problems arising in subsurface contaminant remediation. *Journal of Computational and Applied Mathematics*, 74(1-2):19–32, 1996.

[5] Debasish Basak, Srimanta Pal, and Dipak Chandra Patranabis. Support vector regression. *Neural Information Processing-Letters and Reviews*, 11(10):203–224, 2007.

[6] Domenico A Bau and Alex S Mayer. Stochastic management of pump-and-treat strategies using surrogate functions. *Advances in Water Resources*, 29(12):1901–1917, 2006.

[7] Keith Beven and Jim Freer. Equifinality, data assimilation, and uncertainty estimation in mechanistic modelling of complex environmental systems using the glue methodology. *Journal of Hydrology*, 249(1):11–29, 2001.

[8] D Bolster, M Barahona, M Dentz, D Fernandez-Garcia, X Sanchez-Vila, P Trinchero, C Valhondo, and DM Tartakovsky. Probabilistic risk analysis of groundwater remediation strategies. *Water Resources Research*, 45(6), 2009.

[9] J Breidt, T Butler, and D Estep. A measure-theoretic computational method for inverse sensitivity problems i: Method and analysis. *SIAM Journal on Numerical Analysis*, 49(5):1836–1859, 2011.

[10] T Butler, D Estep, S Tavener, C Dawson, and JJ Westerink. A measure-theoretic computational method for inverse sensitivity problems iii: Multiple quantities of interest. *SIAM/ASA Journal on Uncertainty Quantification*, 2(1):174–202, 2014.

[11] T Butler, J Jakeman, and Tim Wildey. Combining push-forward measures and bayes' rule to construct consistent solutions to stochastic inverse problems. *SIAM Journal on Scientific Computing*, 40(2):A984–A1011, 2018.

[12] Troy Butler, Don Estep, Simon Tavener, Timothy Wildey, Clint Dawson, and Lindley Graham. Solving stochastic inverse problems using sigma-algebras on contour maps. *arXiv preprint arXiv:1407.3851*, 2014.

[13] Troy Butler, L Graham, S Mattis, and Scott Walsh. A measure-theoretic interpretation of sample based numerical integration with applications to inverse and prediction problems under uncertainty. *SIAM Journal on Scientific Computing*, 39(5):A2072–A2098, 2017.

[14] Troy Butler, Antti Huhtala, and Mika Juntunen. Quantifying uncertainty in material damage from vibrational data. *Journal of Computational Physics*, 283:414–435, 2015.

[15] Jesus Carrera and Shlomo P Neuman. Estimation of aquifer parameters under transient and steady state conditions: 2. uniqueness, stability, and solution algorithms. *Water Resources Research*, 22(2):211–227, 1986.

[16] Chih-Chung Chang and Chih-Jen Lin. Libsvm: a library for support vector machines. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 2(3):27, 2011.

[17] Robert P Chapuis. Predicting the saturated hydraulic conductivity of sand and gravel using effective diameter and void ratio. *Canadian geotechnical journal*, 41(5):787–795, 2004.

[18] Robert P Chapuis and Michel Aubertin. On the use of the kozeny carman equation to predict the hydraulic conductivity of soils. *Canadian*

Geotechnical Journal, 40(3):616–628, 2003.

[19] Yan Chen and Dongxiao Zhang. Data assimilation for transient flow in geologic formations via ensemble kalman filter. *Advances in Water Resources*, 29(8):1107–1122, 2006.

[20] Ramon Codina. A discontinuity-capturing crosswind-dissipation for the finite element solution of the convection-diffusion equation. *Computer Methods in Applied Mechanics and Engineering*, 110(3-4):325–342, 1993.

[21] National Research Council et al. *Alternatives for managing the nation's complex contaminated groundwater sites.* National Academies Press, 2013.

[22] Gedeon Dagan. Stochastic modeling of groundwater flow by unconditional and conditional probabilities: 2. the solute transport. *Water Resources Research*, 18(4):835–848, 1982.

[23] Rong-En Fan, Pai-Hsuen Chen, and Chih-Jen Lin. Working set selection using second order information for training support vector machines. *Journal of Machine Learning Research*, 6(Dec):1889–1918, 2005.

[24] J Freer and K Beven. Bayesian estimation of uncertainty in runoff prediction and the value of data: An applicaiton of the glue approach. *Water Resources Research*, 32(7):2161–2173, 1996.

[25] R Allan Freeze. A stochastic-conceptual analysis of one-dimensional groundwater flow in nonuniform homogeneous media. *Water Resources Research*, 11(5):725–741, 1975.

[26] Roger G Ghanem and Pol D Spanos. *Stochastic finite elements: a spectral approach*. Courier Corporation, 2003.

[27] M Kashif Gill, Tirusew Asefa, Mariush W Kemblowski, and Mac McKee. Soil moisture prediction using support vector machines. *Journal of the American Water Resources Association*, 42(4):1033–1046, 2006.

[28] Lindley Graham, Troy Butler, Scott Walsh, Clint Dawson, and Joannes J. Westerink. A measure-theoretic algorithm for estimating bottom friction in a coastal inlet: Case study of bay st. louis during hurricane gustav (2008). *Monthly Weather Review*, 145(3), Mar 2017.

[29] Lindley Graham, Steven Mattis, Scott Walsh, Troy Butler, Michael Pilosov, and Damon McDougall. BET: Butler, Estep, Tavener Method v2.0.0, August 2016.

[30] Isabelle Guyon, Jason Weston, Stephen Barnhill, and Vladimir Vapnik. Gene selection for cancer classification using support vector machines. *Machine Learning*, 46(1):389–422, 2002.

[31] Dylan R Harp and Velimir V Vesselinov. Contaminant remediation decision analysis using information gap theory. *Stochastic environmental research and risk assessment*, 27(1):159–168, 2013.

[32] Jiachuan He, Steven A Mattis, Troy D Butler, and Clint N Dawson. Data-driven uncertainty quantification for predictive flow and transport modeling using support vector machines. *Computational Geosciences*, pages 1–15, 2018.

[33] Charles Hirsch. *Numerical computation of internal and external flows: The fundamentals of computational fluid dynamics.* Elsevier, 2007.

[34] Robert J Hoeksema and Peter K Kitanidis. Analysis of the spatial structure of properties of selected aquifers. *Water Resources Research*, 21(4):563–572, 1985.

[35] Chih-Wei Hsu, Chih-Chung Chang, Chih-Jen Lin, et al. A practical guide to support vector classification.

[36] A Jardani, A Revil, and JP Dupont. Stochastic joint inversion of hydrogeophysical data for salt tracer test monitoring and hydraulic conductivity imaging. *Advances in Water Resources*, 52:62–77, 2013.

[37] Gyozo Jordan and Ahmed Abdaal. Decision support methods for the environmental assessment of contamination at mining sites. *Environmental monitoring and assessment*, 185(9):7809–7832, 2013.

[38] Ajay Kalra and Sajjad Ahmad. Using oceanic-atmospheric oscillations for long lead time streamflow forecasting. *Water Resources Research*, 45(3), 2009.

[39] Michael Kasenow. *Determination of hydraulic conductivity from grain size analysis.* Water Resources Publication, 2002.

[40] Olivier Le Maître and Omar M Knio. *Spectral methods for uncertainty quantification: with applications to computational fluid dynamics.* Springer Science & Business Media, 2010.

[41] PC Leube, A Geiges, and W Nowak. Bayesian assessment of the expected data impact on prediction confidence in optimal sampling design. *Water Resources Research*, 48(2), 2012.

[42] Randall J LeVeque. Conservative methods for nonlinear problems. In *Numerical Methods for Conservation Laws*, pages 122–135. Springer, 1990.

[43] Gwo-Fong Lin, Guo-Rong Chen, Ming-Chang Wu, and Yang-Ching Chou. Effective forecasting of hourly typhoon rainfall using support vector machines. *Water Resources Research*, 45(8), 2009.

[44] Michel Loeve. *Probability theory.* Courier Dover Publications, 2017.

[45] Anders Logg, Kent-Andre Mardal, and Garth Wells, editors. *Automated Solution of Differential Equations by the Finite Element Method.* Springer Berlin Heidelberg, 2012.

[46] Youssef M. Marzouk, Habib N. Najm, and Larry A. Rahn. Stochastic spectral methods for efficient bayesian solution of inverse problems. *Journal of Computational Physics*, 224(2):560–586, June 2007.

[47] SA Mattis, TD Butler, CN Dawson, D Estep, and VV Vesselinov. Parameter estimation and prediction for groundwater contamination based on measure theory. *Water Resources Research*, 51(9):7608–7629, 2015.

[48] R Nelson. In-place measurement of permeability in heterogeneous media: 1. theory of proposed method. *Journal of Geophysical Research*, 65(6):1753–1758, 1960.

[49] R William Nelson. In-place measurement of permeability in heterogeneous media: 2. experimental and computational considerations. *Journal of Geophysical Research*, 66(8):2469–2478, 1961.

[50] D Kirk Nordstrom. Worldwide occurrences of arsenic in ground water, 2002.

[51] W Nowak, FPJ De Barros, and Y Rubin. Bayesian geostatistical design: Task-driven optimal site investigation when the geostatistical model is uncertain. *Water Resources Research*, 46(3), 2010.

[52] G Nützmann, M Thiele, S Maciejewski, and K Joswig. Inverse modelling techniques for determining hydraulic properties of coarse-textured porous media by transient outflow methods. *Advances in Water Resources*, 22(3):273–284, 1998.

[53] Justine Odong. Evaluation of empirical formulae for determination of hydraulic conductivity based on grain-size analysis. *Journal of American Science*, 3(3):54–60, 2007.

[54] D O'Malley and VV Vesselinov. Groundwater remediation using the information gap decision theory. *Water Resources Research*, 50(1):246–256, 2014.

[55] Elaine S Oran and Jay P Boris. *Numerical simulation of reactive flow.* Cambridge University Press, 2005.

[56] Houman Owhadi, Clint Scovel, and Timothy John Sullivan. On the brittleness of bayesian inference. *SIAM Review*, 57(4):566–582, 2015.

[57] Rommel G Regis and Christine A Shoemaker. A stochastic radial basis function method for the global optimization of expensive functions. *INFORMS Journal on Computing*, 19(4):497–509, 2007.

[58] Rommel G Regis and Christine A Shoemaker. Combining radial basis function surrogates and dynamic coordinate search in high-dimensional expensive black-box optimization. *Engineering Optimization*, 45(5):529–555, 2013.

[59] C.P. Robert and George Casella. *Monte Carlo Statistical Methods.* Springer, 2004.

[60] Hermann Rügner, Michael Finkel, Arno Kaschl, and Martin Bittens. Application of monitored natural attenuation in contaminated land managementa review and recommended approach for europe. *Environmental science & policy*, 9(6):568–576, 2006.

[61] Yin-Tzer Shih and Howard C Elman. Iterative methods for stabilized discrete convection-diffusion problems. *IMA journal of numerical analysis*, 20(3):333–358, 2000.

[62] Menner A Tatang, Wenwei Pan, Ronald G Prinn, and Gregory J McRae. An efficient method for parametric uncertainty analysis of numerical geophysical models. *Journal of Geophysical Research: Atmospheres*, 102(D18):21925–21932, 1997.

[63] R Therrien, RG McLaren, EA Sudicky, and SM Panday. Hydrogeosphere: A three-dimensional numerical model describing fully-integrated subsurface and surface flow and solute transport. *Groundwater Simulations Group, University of Waterloo, Waterloo, ON*, 2010.

[64] Mads Troldborg, Wolfgang Nowak, Nina Tuxen, Poul L Bjerg, Rainer Helmig, and Philip J Binning. Uncertainty evaluation of mass discharge estimates from a contaminated site using a fully bayesian framework. *Water Resources Research*, 46(12), 2010.

[65] M Th Van Genuchten and RJ Wagenet. Two-site/two-region models for pesticide transport and degradation: Theoretical development and analytical solutions. *Soil Science Society of America Journal*, 53(5):1303–1310, 1989.

[66] Vladimir Naumovich Vapnik and Vlamimir Vapnik. *Statistical learning theory*, volume 1. Wiley New York, 1998.

[67] J Vrugt, C ter Braak, H Gupta, and B Robinson. Equifinality of formal (dream) and informal (glue) bayesian approaches in hydrologic modeling? *Stochastic Environmental Research and Risk Assessment*, 23(7):1011–1026, 2008.

[68] Dongbin Xiu and George Em Karniadakis. The wiener–askey polynomial chaos for stochastic differential equations. *SIAM Journal on Scientific Computing*, 24(2):619–644, 2002.

[69] Pao-Shan Yu, Shien-Tsung Chen, and I-Fan Chang. Support vector regression for real-time flood stage forecasting. *Journal of Hydrology*, 328(3):704–716, 2006.

[70] Dongxiao Zhang and Zhiming Lu. An efficient, high-order perturbation approach for flow in random porous media via karhunen–loeve and polynomial expansions. *Journal of Computational Physics*, 194(2):773–794, 2004.

[71] Yi Zheng, Weiming Wang, Feng Han, and Jing Ping. Uncertainty assessment for watershed water quality modeling: A probabilistic collocation method based approach. *Advances in Water Resources*, 34(7):887–898, 2011.

[72] Haiyan Zhou, J Jaime Gómez-Hernández, and Liangping Li. Inverse methods in hydrogeology: Evolution and recent trends. *Advances in Water Resources*, 63:22–37, 2014.

Vita

Jiachuan He was born in Shanghai, China. He received the Bachelor of Science degree in Engineering Mechanics from Shanghai Jiao Tong University in 2010 and the Master of Science degree in Mechanical Engineering from Carnegie Mellon University in 2012. He started the PhD program in the Department of Aerospace Engineering and Engineering Mechanics at the University of Texas at Austin in August 2012.

Permanent address: jiachuan.he@gmail.com

This dissertation was typeset with LaTeX[†] by the author.

[†] LaTeX is a document preparation system developed by Leslie Lamport as a special version of Donald Knuth's TeX Program.