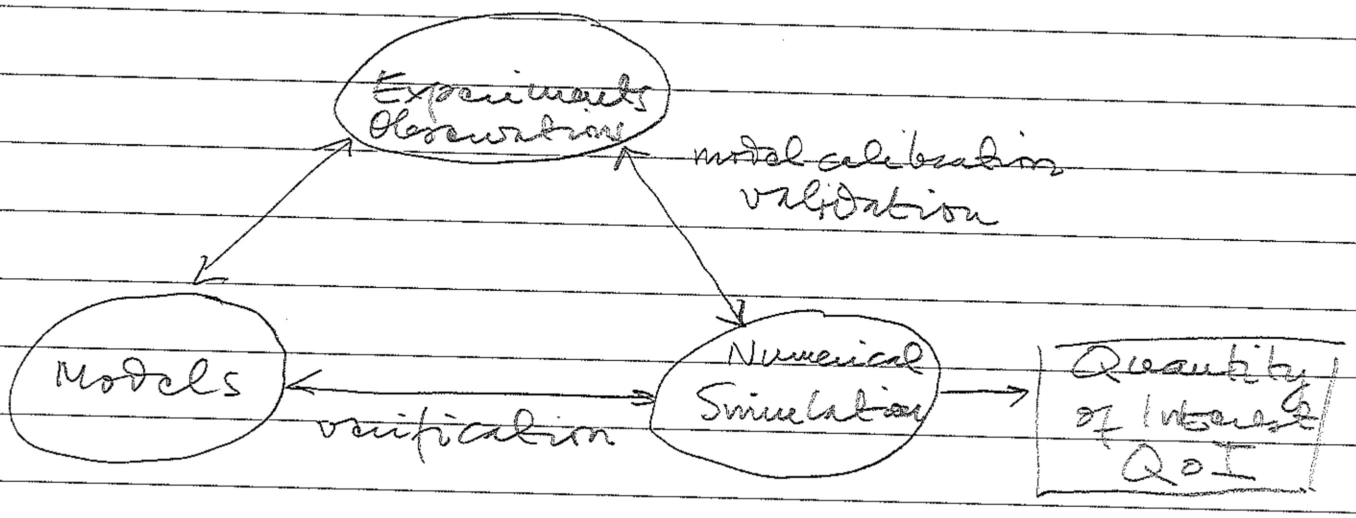
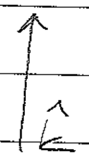
Predictive science



Weather prediction

$$\frac{\partial \rho}{\partial t} + \nabla \cdot (\rho v) = 0$$

$$\frac{\partial v}{\partial t} = -v \cdot \nabla v - \frac{1}{\rho} \nabla p - g\hat{k} - 2\Omega \times v$$



$$\rho c_v \frac{\partial T}{\partial t} + p \nabla \cdot v = -\nabla \cdot F + \nabla \cdot (k \nabla T) + \rho \dot{q}(T, p, \rho)$$

↙ concentration of water in solid, liquid, gaseous phase

$$j = 1, 2, 3: \quad \frac{\partial m_j}{\partial t} = -v \cdot \nabla m_j + S_{m_j}(T, m_j, \chi_j, \rho)$$

↙ aerosol j concentration

$$j = 1 \dots J: \quad \frac{\partial \chi_j}{\partial t} = -v \cdot \nabla \chi_j + S_{\chi_j}(T, \chi_j, \rho)$$

phenomenological models for S_{m_j}, S_{χ_j}

$$S_{m_2} = S_1 + S_2 + S_3 - S_4$$

$$S_1 = \bar{\rho}(m_2 - m_2^*)^2 \left(1.2 \cdot 10^4 + \left(1.569 \cdot 10^{-12} \frac{w_0}{d_0(m_2 - m_2^*)}\right)\right)^{-1}$$

subgrid modeling

Model and

Input uncertainties

parameters $\bar{\xi}, m_2^*, n_r, d_0$ in S_1
initial conditions, boundary conditions, model errors

Numerical errors and uncertainties

grid size: horizontal 5 km, vertical 200 m
discretization errors
subgrid models: cloud formation, turbulence

Measurement errors and uncertainties

limited accuracy of sensors, uncertainty in exact position and time

Weather forecasts

Determine values, quantify uncertainties (parameter estimation) for initial values, phenomenological parameters (data assimilation, model calibration)
Then run model for forecasts with quantified uncertainties

Ensemble (multiple) simulations, use initial data and parameters drawn from probability densities

Definitions

Inputs: parameters, initial conditions, boundary conditions

Quantity of interest (QoI): Output of simulation model (average temperature, precipitation ...)

Verification: Quantifying the accuracy of numerical method used to implement math. model

Validation: accuracy of math model to quantify physical process of interest

Types of uncertainty

Aleatoric uncertainty: statistical, stochastic uncertainty, inherent to a problem, cannot be reduced by more physical or experimental knowledge. Example: initial conditions for weather models

Epistemic or systematic uncertainty: caused by simplified models, missing physics, basic lack of knowledge. Example: closure relations, numerical errors

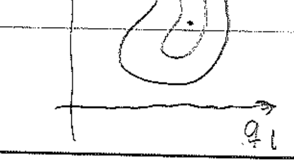
Predictive estimation

Probabilistic quantification of predicted computational outcome with identified and quantified uncertainties

Model calibration: data to quantify and update uncertainties in parameters, initial and boundary conditions

Model prediction: QoI with statistics, probability density function (pdf) for QoI

Estimation of Validation regime: Contours of constant probability for QoI



$y = f(\chi, q)$, $q = (q_1, q_2 \cdots q_p)$ parameters

random var. (epistemic) $\chi = (x, t)$, indep. var.

model err

$Y_i = f(\chi_i, q) + \delta(\chi_i) + \varepsilon_i$ ← random variable (aleatoric)

Outline of course

	Ch.
Typical models	3
Probability, Random processes, statistics	4
Random inputs	5
Parameter selection	6
Frequentist techniques for parameter est.	7
Bayesian techniques for parameter est.	8
Uncertainty propagation	9
Prediction with model Discrepancy	12
Surrogate models	13
Local sensitivity analysis	14

UQ2

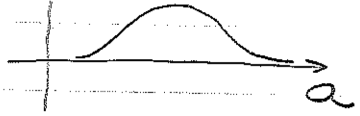
Simple models

1. $$\frac{dz}{dt} = az + b(t), \quad z(0) = z_0$$

parameters $q = [a, z_0]$

$$z(t,q) = e^{at}\left(z_0 + \int_0^t e^{-as} b(s)\,ds\right)$$

random variables $a(\omega), z_0(\omega)$



ω event in probability space

$$\frac{dz}{dt} = a(\omega)z + b(t,\omega), \quad z(0) = z_0(\omega)$$

$$z(t,q) = e^{a(\omega)t}\left[z_0(\omega) + \int_0^t e^{-a(\omega)s} b(s,\omega)\,ds\right]$$

2. Harmonic oscillator

$$m\frac{d^2z}{dt^2} + c\frac{dz}{dt} + kz = f_0 \cos \omega_F t$$

$$z(0) = z_0, \quad \frac{dz}{dt}(0) = z_1$$

$$z(t) = c_1 e^{r_1 t} + c_2 e^{r_2 t} + \frac{f_0}{\sqrt{m^2(\omega_0^2-\omega_F^2)^2 + c^2\omega_F^2}} \cos(\omega_F t - \delta)$$

$$\cos\delta = \frac{m(\omega_0^2-\omega_F^2)}{\sqrt{m^2(\omega_0^2-\omega_F^2) + c^2\omega_F^2}} \qquad \omega_0 = \sqrt{\frac{k}{m}}$$

$$r_{1,2} = \frac{-c \pm \sqrt{c^2-4km}}{2m}, \quad z_0, z_1 \Rightarrow c_1, c_2$$

$\mathrm{Re}\, r_{1,2} < 0 \Rightarrow z(t) \sim \cos \omega_F t - \delta$ for large t

If $c^2 - 4km < 0$ then oscillatory first part

$q = [m, c, k, f_0]$ has the same solution as

$q = [1, \underbrace{\frac{c}{m}, \frac{k}{m}, \frac{f_0}{m}}]$, only 3 parameters can be uniquely determined by data

measure z or $\dot{z}$, state $u = [z, \dot{z}]$
observations $y = C^T u$
either position z $C^T = (1, 0)$
or velocity $\dot{z}$ $C^T = (0, 1)$

General form $\overset{\text{observation}}{y} = R(\overset{\text{state}}{u}, q)$ ← parameters

all states may not be observable

$$L(q)u = F(q(\chi)) \quad , \chi \in \Omega = \underset{\subset \mathbb{R}^q}{\overset{x \in D}{D}} \times \underset{\subset \mathbb{R}_+}{\overset{t \in J}{J}}$$

$q(\chi) = [q_1(\chi), \cdots q_p(\chi)]^T$ parameters
$\chi = [x, t]$ independent variables
$L(q) = [L_1(q), \cdots L_N(q)]^T$
$u = [u_1(\chi), \cdots u_N(\chi)]^T$, state vector
$F = [F_1 \cdots F_N]^T$ source terms

$$B(q)u = G(q) \quad \chi \in \partial\Omega$$

boundary and initial conditions

end lecture 1

Background in probability

Probability space $(\Omega, \mathcal{F}, P)$

Ω : sample space, set of all possible outcomes of an experiment
$\mathcal{F}$: σ-field of subsets of Ω containing all events of interest
$P : \mathcal{F} \to [0,1]$: probability (measure) satisfying

i : $P(\phi) = 0$

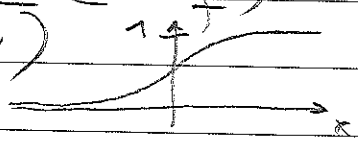
ii : $P(\Omega) = 1$

iii : if $A_i \in \mathcal{F}$ and $A_i \cap A_j = \phi$ then
$$P(\bigcup_{i=1}^{\infty} A_i) = \sum_{i=1}^{\infty} P(A_i)$$

Realization : x in $x = X(\omega)$ for $\omega \in \Omega$

Cumulative Distribution Function (cdf)

$$F_X(x) = P(\omega \in \Omega \mid X(\omega) \le x)$$
$$= P(X \le x)$$



$\lim_{x \to -\infty} F_X(x) = 0$, $\lim_{x \to \infty} F_X(x) = 1$, $x_1 \le x_2 \Rightarrow F_X(x_1) \le F_X(x_2)$

Probability Density Function (pdf)

$f_X(x)$, $x \in \mathbb{R}$

$$F_X(x) = \int_{-\infty}^{x} f_X(s)\,ds, \quad f_X = \frac{dF_X}{dx}$$

$$f_X(x) \ge 0, \quad \int_{\mathbb{R}} f_X(x)\,dx = 1$$

$$P(x_1 \le X \le x_2) = F_X(x_2) - F_X(x_1) = \int_{x_1}^{x_2} f_X(x)\,dx$$

Moments $\mathbb{E}(X^n) = \int_{\mathbb{R}} x^n f_X(x)\,dx$

expected value: $\mathbb{E}(X) = \int x f_X(x)\,dx = \mu$

second central moment, variance
$$\sigma^2 = Var(X) = \mathbb{E}((X-\mu)^2) = \int_{\mathbb{R}} (x-\mu)^2 f_X(x)\,dx$$

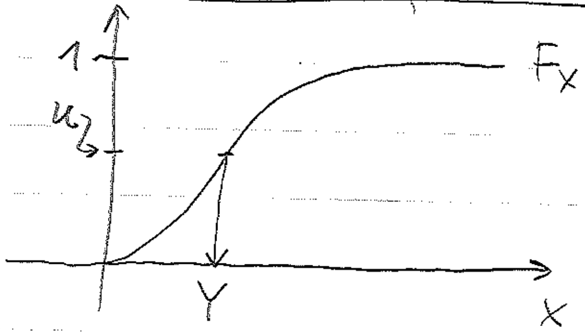
Normal Distribution $f_X(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-(x-\mu)^2/2\sigma^2}$

$-\infty < x < \infty$

⊛ multivariate normal distribution — determinant of V

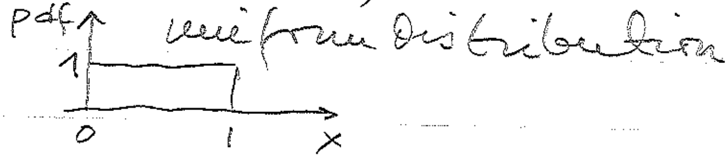
$V_{ij} = Cov(X_i, X_j)$ $\quad f_X(x) = \frac{1}{\sqrt{(2\pi)^n |V|}} \exp\left(-\frac{1}{2}(x-\mu) V^{-1} (x-\mu)\right)$

Inverse Transform Sampling



X has cdf $F_X(x)$

$U \in \mathcal{U}(0,1)$



$Y = F_X^{-1}(U)$ has the same distribution as X

$$F_Y(y) = P(Y \leq y) = P(F_X^{-1}(U) \leq y)$$
$$= P(U \leq F_X(y)) = F_X(y)$$

realizations of X: $x = F_X^{-1}(u)$

Multiple random variables

$X = [X_1, X_2, \cdots X_n]$

$$F_X(x_1, x_2 \cdots x_n) = P(\omega \in \Omega \mid X_j(\omega) \leq x_j)$$
$$= P(X_1 \leq x_1, \cdots X_n \leq x_n)$$

$$\mathrm{cov}(X,Y) = E[(X - E(X))(Y - E(Y))]$$
$$= E(XY) - E(X)E(Y)$$

X and Y are independent: $E(XY) = E(X)E(Y)$
$\Rightarrow \mathrm{cov}(X,Y) = 0$

$X_i \sim N(\mu_i, \sigma_i^2)$, independent

$$Z = \sum_{i=1}^{n} a_i X_i + b_i \sim N\left(\sum_{i=1}^{n}(a_i\mu_i + b_i), \sum_{i=1}^{n} a_i^2 \sigma_i^2\right)$$

⊛ →

Marginal pdf

X_1, X_2 have joint pdf $f_X(x_1, x_2)$

$f_{X_1}(x_1) = \int_R f_X(x_1, x_2)\, dx_2$

Conditional pdf

Conditional density of X_1 given $X_2 = x_2$ is

$$f_{X_1|X_2}(x_1|x_2) = \begin{cases} \frac{f_X(x_1, x_2)}{f_{X_2}(x_2)} & f_{X_2}(x_2) > 0 \\ 0 & \text{otherwise} \end{cases}$$

Estimator

Parameter q, samples $X_1, \dots X_n$

construct estimates of q based on X_i

estimator is random variable

estimate is realization of estimator

unbiased estimator if its mean equals paramete

Examples $\bar{X} = \frac{1}{n} \sum_{i=1}^{n} X_i$ sample mean

$S^2 = \frac{1}{n-1} \sum_{i=1}^{n} (X_i - \bar{X})^2$ sample variance

$X_i \sim N(\mu, \sigma^2) \Rightarrow \bar{X} \sim N(\mu, \frac{\sigma^2}{n})$

$S^2 \sim \frac{\sigma^2}{n-1} \chi^2(n-1)$

χ^2 distribution: $X_i \sim N(0,1)$, $\underset{i \neq j}{X_i \text{ and } X_j}$ are independent

$\Rightarrow Z = \sum_{i=1}^{k} X_i^2$ is $\chi^2(k)$ distributed

$E(Z) = k$, $Var(Z) = 2k$

Interval estimator and Confidence interval

$q_L(x), q_R(x)$ such that $q_L(x) < q < q_R(x)$
based on $x = (x_1, \cdots x_n)$
$[q_L(x), q_R(x)]$ interval estimator
+ confidence coefficient $1-\alpha$ = confidence interval

$$P(q_L(X) \le q \le q_R(X)) = 1-\alpha$$

Interpretation: the frequency of times that the interval will contain parameter q

Model calibration

Statistical model

$$\underset{\text{random variable, realization } y_i}{Y_i} = f(t_i, q_0) + \varepsilon_i, \quad i = 1, \cdots n$$

ε_i, errors, independent identically distributed (iid)
unbiased $E[\varepsilon_i] = 0$, $Var(\varepsilon_i) = \sigma_0^2$

Ordinary Least Squares (OLS) estimator

$$\hat{q}_{OLS} = \arg\min_{q \in Q} \sum_{i=1}^{n} (Y_i - f(t_i, q))^2$$

$$q_{OLS} = \arg\min_{q \in Q} \sum_{i=1}^{n} (y_i - f(t_i, q))^2$$

Maximum Likelihood (ML) estimator

$f_Y(y; q)$ joint pdf for $\overset{\text{random vector}}{Y} = (Y_1, \cdots Y_n)$

Likelihood function

$$L_Y(q) = \overbrace{L(q|y)}^{\text{not conditional pdf, function of } q} = f_Y(y; q)$$

sample y given, q varies

n iid random variables Y_i

$$L(q|y) = f_Y(y;q) = \prod_{i=1}^{n} f_Y(y_i;q)$$

log-likelihood function

$$l_Y(q) = l(q|y) = \log L(q|y)$$

maximum likelihood estimate (MLE)

$$q_{MLE} = \arg\max_{q \in Q} \prod_{i=1}^{n} f_Y(y_i;q)$$

monotonicity of log ⇒ maximizing $L(q|y)$ is equivalent to maximizing $l(q|y)$ (computationally easier)

q_{MLE} is the parameter value that makes the output most likely (frequentist view)

Convergence and Limit theorems

Sequence $X_1, X_2, \cdots$ converges to X

Almost sure convergence $X_n \xrightarrow{a.s.} X$

if for every $\varepsilon > 0$

$\Downarrow$ $$P\left(\lim_{n\to\infty} |X_n - X| < \varepsilon\right) = 1$$

Convergence in probability $X_n \xrightarrow{P} X$

$$\lim_{n\to\infty} P(|X_n - X| < \varepsilon) = 1$$

$\Downarrow$ Convergence in Distribution $X_n \xrightarrow{D} X$

cdf's $F_{X_1}(x), F_{X_2}(x), \cdots, \lim_{n\to\infty} F_{X_n}(x) = F_X(x)$

X has cdf $F_X(x)$

Law of large numbers

estimator of mean $\bar{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i$

Strong law of large numbers

X_i, $i=1,2\cdots$ iid random variables

$E(X_i)=\mu$, $Var(X_i)=\sigma^2<\infty$

Then $P(\lim_{n\to\infty} |\bar{X}_n-\mu|<\varepsilon)=1$, $\bar{X}_n \xrightarrow{a.s.} \mu$

(weak form: $\bar{X}_n \xrightarrow{P} \mu$)

Central limit theorem

Assumptions as above

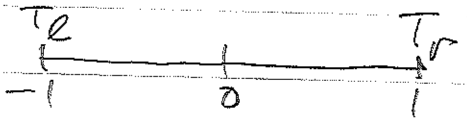
$$Z_n = \frac{\sqrt{n}(\bar{X}_n-\mu)}{\sigma} \xrightarrow{D} Z \sim N(0,1)$$

$$\text{or } \bar{X}_n \xrightarrow{D} \bar{X} \sim N\left(\mu, \frac{\sigma^2}{n}\right)$$

Random processes

temperature T, heat conductivity $\alpha(x,\omega)$

$$\frac{\partial T}{\partial x} = \frac{\partial}{\partial x}\left(\alpha(x,\omega)\frac{\partial T}{\partial x}\right),$$



, $t \geq 0$, $T(x,0)=T_0(x)$

$T(x,t,\omega)$ is $\underbrace{\text{a random variable for all}}_{=\{X_t,\, t\in\mathcal{J}\}}$ x,t

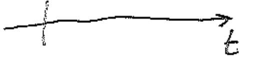
Stochastic process: $X^t = \{X(t), t\in\mathcal{J}\}$

$\mathcal{J}$ is interval (continuous process)

or consecutive integers (discrete process)

$X_t(\omega)$ realization of process, $t\in\mathcal{J}$, $\omega\in\Omega$

sample path or trajectory of ω



Second order stochastic process: $E(X_t^2) < \infty$

$$\mu(t) = E(X_t), \quad t \in J$$

$$C(t,s) = Cov(X_t, X_s) = E((X_t - \mu(t))(X_s - \mu(s))) \quad t, s \in J$$

<u>Markov chains</u>

Discrete state space S, $X = \{X_i, i \in \overset{\text{non-negative integers}\atop\text{time}}{\mathbb{Z}}\}$

Markov property: X_{n+1} depends only on X_n

$$P(X_{n+1} = x_{n+1} | X_0 = x_0 \dots, X_n = x_n) = P(X_{n+1} = x_{n+1} | X_n = x_n)$$

x_i is the state at time i

finite discrete state space

$S = \{x_1, x_2, \cdots x_k\}$, initial distribution p^0

transition probability $p_{ij} = P(X_{n+1} = x_j | X_n = x_i)$

$\sum_j p_{ij} = 1$ $\quad i,j = 1 \dots k \quad p_i^0 = P(X_0 = x_i)$

$$P_{ij} = p_{ij},$$

$$p_j^1 = (p^0 P)_j = \sum_{i=1}^k p_i^0 p_{ij}$$

$$= \sum_{i=1}^k P(X_1 = x_j | X_0 = x_i) P(X_0 = x_i)$$

$$= \sum_{i=1}^k P(X_1 = x_j, X_0 = x_i) = P(X_1 = x_j)$$

$$p^n = p^{n-1} P = p^0 P^n$$

(stationary)

Limiting distribution $\quad \pi = \pi P, \quad \sum \pi_i = 1$

$$\pi = \lim_{n\to\infty} p^n = \lim_{n\to\infty} p^0 P^n = \lim_{n\to\infty} p^0 P^{n+1} = (\lim_{n\to\infty} p^0 P^n) P = \pi P$$

(✱) P.T.O.

lit. lect. 2

<u>Random differential equations</u>

Random effects in parameters, initial or boundary conditions

$$\frac{dz}{dt} = a(\omega) z + b(t, \omega), \quad z(0) = z_0(\omega)$$

<u>Stochastic differential equations</u>

Wiener process

$$dZ(t) = -aZ(t)dt + b\, dW(t) \quad \text{or}$$

$$Z(t) = Z_0 - \int_0^t aZ(s)ds + \underbrace{\int_0^t b\, dW(s)}_{\text{Ito integral}}$$

models Brownian motion

Statistical inference

Given $S = \{x_1, x_2 \cdots x_n\}$, $x_j \in \mathbb{R}^N$
n observed realizations of X
want to infer probability distribution of X
parametric inference: prob. distribution has
small number of parameters, e.g. mean, variance
problem: estimate parameters
non-parametric inference: construct distribution
based on observations only

Frequentist inference

probabilities are defined as the frequencies
with which an event occurs if experiment is
repeated a large number of times
OLS or ML estimators
fixed parameter q_0, $(1-\alpha)\cdot 100\,\%$ of confidence
intervals contain q_0

Bayesian inference

parameter estimation solution is a
posterior probability density
parameters are random variables
credible interval: has $(1-\alpha)\cdot 100\%$ chance
of containing expected parameter
update probability when more data are known

Ch 4 Exercises: 4.4, 4.5, 4.7

Bayes' formula

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)} = \frac{P(A,B)}{P(B)}$$

$$\pi(q|y) = \frac{\pi(y|q)\pi_0(q)}{\pi_Y(y)}$$

prior density

posterior density

likelihood normalization factor

$\pi_0(q)$ quantifies prior knowledge of parameter q
no such information: use noninformative prior
e.g. $\pi_0(q) = \chi_{[0,\infty)}(q)$

χ_J is indicator function $\chi_J = \begin{cases} 1, & x \in J \\ 0, & x \notin J \end{cases}$

$\pi(\overset{\text{fixed}}{y}|q)$ quantifies likelihood $L(q|y)$ of observing y given parameter realizations q

Joint density $\pi(q,y) = \pi(y|q)\pi_0(q)$ is normalized to 1 by $\pi_Y(y)$

Posterior density $\pi(q|y)$ quantifies probability to get parameter q given observations y

Model calibration with $\pi_Y(y) = \int_{\mathbb{R}^p} \pi(y|q)\pi_0(q)\,dq$

$$\Rightarrow \quad \pi(q|y) = \frac{\pi(y|q)\pi_0(q)}{\int \pi(y|q)\pi_0(q)\,dq}$$

Conjugate priors

prior and posterior distributions have the same parametric form, $\pi_0(q)$ is then conjugate prior of $\pi(y|q)$