Predictive science

Experimental observation → model calibration → validation

Models → unification → Numerical simulation

Quantity of interest $Q_0$

Weather prediction

\[
\frac{2\xi}{\partial t} + \nabla \cdot (\xi \nabla V) = 0
\]

\[
\frac{\partial V}{\partial t} = -V \cdot \nabla V - \frac{1}{\xi} \nabla \rho - g \hat{z} - 2 \omega \times V
\]

\[
q CV \frac{\partial T}{\partial t} + p \nabla \cdot V = -\nabla \cdot F + \nabla \cdot (k \nabla T) + \varepsilon g(T, p, \varphi)
\]

Concentration of water in solid, liquid, gas

\[
j = 1, 2, 3: \quad \frac{\partial m_j}{\partial t} = -V \cdot \nabla m_j + S_{m_j}(T, m_j, x, \varphi)
\]

Atmospheric concentration

\[
j = 1, \ldots, J: \quad \frac{\partial X_j}{\partial t} = -V \cdot \nabla X_j + S_{X_j}(T, x, \varphi)
\]

Phenomenological models for $S_{m_j}, S_{X_j}$

\[
S_{m_2} = S_1 + S_2 + S_3 - S_4
\]

\[
S_1 = \frac{\varepsilon (m_2 - m_2^*)^2}{(1.2 \times 10^4 + (1.569 \times 10^{-2} \frac{\varphi}{a(h_2 - h_3^*)})} - 1
\]

Subgrid modeling
Input uncertainties

parameters $\beta, m, n, \gamma, \lambda, \omega$ in $S_1$
initial conditions, boundary conditions, model runs

Numerical errors and uncertainties

grid size: horizontal 5 km, vertical 200 m
gas concentration 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
for initial values, phenomenological parameters (fire assimilation, model calibration)

Then run model for forecasts with quantified uncertainties

(multiple)

Ensemble 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 the model

Validation: accuracy of model to quantify physical process of interest
Types of uncertainty

Aleatory 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.

\[ Y = f(X, \theta), \quad \theta = (\theta_1, \theta_2, \ldots, \theta_p) \]

\[ Y_i = f(X_i, \theta) + \delta(X_i) + \xi_i \]
Outline of course

Typical models
Probability, random processes, statistics
Random inputs
Parameter selection

Frequentist techniques for parameter est.
Bayesian techniques for parameter est.
Uncertainty propagation
Prediction with model discrepancy
Surrogate models
Local sensitivity analysis

ch.
3
4
5
6
7
8
9
12
13
14
Simple models

1.
\[
\frac{d^2 z}{dt^2} = az + b(t), \quad z(0) = z_0
\]

parameters \( q = [a, z_0] \)
\[
z(t, q) = e^{at}(z_0 + \int_0^t e^{-as} b(s) ds)
\]

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

in a 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

\[
m \frac{d^2 z}{dt^2} + c \frac{dz}{dt} + k z = \int_0^t \cos \omega t
\]
\[
z(0) = z_0, \quad \frac{dz}{dt}(0) = \dot{z}_0
\]
\[
z(t) = c_1 e^{r_1 t} + c_2 e^{r_2 t} + \frac{\int_0^t \cos \omega t}{\sqrt{m^2(\omega_0^2 - \omega^2) + c^2 \omega^2}} \cos (\omega t + \delta)
\]
\[
\cos \delta = \frac{m(\omega_0^2 - \omega^2)}{\sqrt{m^2(\omega_0^2 - \omega^2) + c^2 \omega^2}} \quad \omega_0 = \sqrt{\frac{k}{m}}
\]
\[
r_{1/2} = \frac{c \pm \sqrt{c^2 - 4km}}{2m} \quad z_{1/2} = c_1, c_2
\]

\[\text{Re } r_{1/2} < 0 \Rightarrow z(t) \sim \cos \omega t - \delta \text{ for large } t\]
\[\text{If } c^2 - 4km < 0 \text{ then oscillating first part}\]

\[
q = [m, c, k, \omega_0] \text{ has the same solution as } \]
\[
q = [m, \frac{c}{m}, \frac{k}{m}, \omega_0] \text{ only 3 parameters can be uniquely determined by data}\]
measure \( z \) or \( \bar{z} \) state \( u = [z, \bar{z}] \)

observation: \( y = C^T u \)

either position \( z \) \( C^T = (1, 0) \)
or velocity \( \bar{z} \) \( C^T = (0, 1) \)

generated from \( y = R(u, q) \)

all states may not be observable

\( x \in D, \text{ to } \mathbb{R} \times J \)

\( L(q) u = F(q(X)) \)

\( q(X) = [q_1(X), \ldots, q_p(X)]^T \) parameters
\( X = [X(t)] \) independent variables
\( L(q) = [L_1(q), \ldots, L_N(q)]^T \)
\( \mathbf{u} = [u_1(X), \ldots, u_N(X)]^T \) state vector
\( F = [F_1, \ldots, F_N]^T \) source terms

\( B(q) u = G(q) \)

boundary and initial conditions

end lecture

Background in probability

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

\( \Omega \): sample space, set of all possible outcomes of an experiment
\( \mathcal{F} \): \( \sigma \)-field of subsets of \( \Omega \) containing all events of interest
\( P : \mathcal{F} \rightarrow [0, 1] \): probability (measure) satisfying
i: \( P(\emptyset) = 0 \)

ii: \( P(\Omega) = 1 \)

iii: if \( A_i, i \in I \) and \( A_i \cap A_j = \emptyset \) then
\[
P(U_{i=1}^\infty A_i) = \sum_{i=1}^\infty P(A_i)
\]

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

Cumulative Distribution Function (CDF)
\[
F_X(x) = P(\omega \in \Omega | X(\omega) \leq x)
= P(X \leq x)
\]
\[
\lim_{x \to -\infty} F_X(x) = 0, \quad \lim_{x \to \infty} F_X(x) = 1, \quad x_1 < x_2 \Rightarrow F_X(x_1) \leq F_X(x_2)
\]

Probability Density Function (PDF)
\[
f_X(x), \quad x \in \mathbb{R}
\]
\[
F_X(x) = \int_{-\infty}^{x} f_X(s) \, ds, \quad f_X = \frac{dF_X}{dx}
\]

\[
f_X(x) \geq 0, \quad \int_{-\infty}^{x} f_X(x) \, dx = 1
\]

\[
P(x_1 \leq X \leq x_2) = F_X(x_2) - F_X(x_1) = \int_{x_1}^{x_2} f_X(x) \, dx
\]

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

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

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

Normal distribution
\[
f_X(x) = \frac{1}{\sqrt{2\pi} \sigma} e^{-\frac{(x - \mu)^2}{2\sigma^2}}
\]
\[-\infty < x < \infty
\]

Multivariate normal distribution, determined by \( V \)
\[
\mathbf{V}_{ij} = \text{Cov}(X_i, X_j) \quad f_X(x) = \frac{1}{(2\pi)^n |V|^{1/2}} \exp \left(-\frac{1}{2}(x - \mu)^T V^{-1}(x - \mu) \right)
\]
**Inverse Transform Sampling**

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

- Realizations of $X$: $X = F_x^{-1}(u)$

**Multiple random variables**

- $X = [X_1, X_2, \ldots X_n]$

- $F_X(x_1, x_2, \ldots x_n) = P(\omega \in \Omega \mid X_j(\omega) \leq x_j)$

- $\text{cov}(X_i, Y) = E[(X - E(X))(Y - E(Y))]$

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

- $X_i \sim N(\mu_i, \sigma_i^2)$, 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
\[ X_1, X_2 \text{ have joint pdf } f_{X_1, X_2}(x_1, x_2) \]
\[ f_{X_1}(x_1) = \int f_{X_1, X_2}(x_1, x_2) \, dx_2 \]

Conditional pdf
Conditional density of \( X_1 \) given \( X_2 = x_2 \) is
\[ f_{X_1|X_2}(x_1|x_2) = \int \frac{f_{X_1, X_2}(x_1, x_2)}{f_{X_2}(x_2)} \, dx_2 \quad f_{X_2}(x_2) > 0 \]
\[ = 0 \quad \text{otherwise} \]

Estimation
Parameter \( \theta \), samples \( X_1, \ldots, X_n \)
construct estimator of \( \theta \), based on \( X_i \)
estimator is random variable
estimator is realization of estimator
unbiased estimator if its mean equals parameter

Examples
\[ \bar{X} = \frac{1}{n} \sum_{i=1}^{n} X_i \quad \text{sample mean} \]
\[ S^2 = \frac{1}{n-1} \sum_{i=1}^{n} (X_i - \bar{X})^2 \quad \text{sample variance} \]
\[ X_i \sim N(\mu, \sigma^2) \Rightarrow \bar{X} \sim N(\mu, \frac{\sigma^2}{n}) \]
\[ S^2 \sim \frac{\sigma^2}{n-1} \chi^2(n-1) \]
\[ \chi^2 \text{ distribution: } X_i \sim N(0,1), \quad X_i \text{ and } X_j \text{ are independent} \]
\[ \Rightarrow Z = \sum_{i=1}^{k} X_i^2 \text{ is } \chi^2(k) \text{ distributed} \]
\[ E(Z) = k, \quad \text{Var}(Z) = 2k \]
Interval estimate and Confidence interval
$q_L(x), q_R(x)$ such that $q_L(x) < \theta < q_R(x)$

based on $X = (x_1, \ldots, x_n)$

$[q_L(x), q_R(x)]$ interval estimator

+ confidence coefficient $1 - \alpha$ confidence interval

$P(q_L(X) \leq \theta \leq q_R(X)) = 1 - \alpha$

Interpretation: the frequency of times that
the interval will contain parameter $\theta$

Model calibration

Statistical model

$Y_i = f(x_i; \theta) + \varepsilon_i, \; i = 1, \ldots, n$

random variable, realization $y_i$

$\varepsilon_i$ errors, independent identically distributed (i.i.d.)

$E[\varepsilon_i] = 0, \; \text{Var}(\varepsilon_i) = \sigma^2$

Ordinary Least Squares (OLS) estimator

$\hat{\theta}_{OLS} = \arg \min_{\theta \in \mathbb{R}} \sum_{i=1}^{n} (Y_i - f(x_i; \theta))^2$

$\hat{\theta}_{OLS} = \arg \min_{\theta \in \mathbb{R}} \sum_{i=1}^{n} (y_i - f(x_i; \theta))^2$

Maximum Likelihood estimator

$\nu(y; \theta)$ joint pdf for $Y = (y_1, \ldots, y_n)$

Likelihood function

$\nu(\theta) = \nu(y; \theta) = \nu(y; \theta)$

sample $y$ given, $\theta$ 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

\[ l_v(q) = L(q | y) = \log L(q | y) \]

maximum likelihood estimates (MLE)

\[ \hat{q}_{MLE} = \operatorname*{arg\,max}_{q \in Q} \prod_{i=1}^{n} f_Y(y_i | q) \]

monotonicity of \( \log \) ⇒ maximizing \( L(q | y) \)

is equivalent to maximizing \( l_v(q) \)

(computationally easier)

\[ q_v \] is the parameter value that makes the output most likely (frequentist view)

Convergence and Limit Theorems

Sequence \( X_1, X_2, \ldots \) converges to \( X \)

almost sure convergence \( X_n \overset{a.s.}{\to} X \)

if for every \( \varepsilon > 0 \)

\[ \Pr \left( \lim_{n \to \infty} (X_n - X) < \varepsilon \right) = 1 \]

Convergence in probability \( X_n \overset{P}{\to} X \)

\[ \lim_{n \to \infty} \Pr \left( |X_n - X| < \varepsilon \right) = 1 \]

Convergence in distribution \( X_n \overset{D}{\to} X \)

\[ \text{cdf's } F_{X_1}(x), F_{X_2}(x), \ldots, \lim_{n \to \infty} F_{X_n}(x) = F_X(x) \]

\( X \) 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**

\[ X_i, i=1, 2, \ldots \text{ i.i.d random variables} \]

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

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

(weak form: \( \frac{\bar{X}_n}{\sigma} \xrightarrow{d} \frac{1}{\sigma} \))

**Central Limit Theorem**

Assumptions as above

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

or \( \bar{X}_n \xrightarrow{d} X \sim N(\mu, \frac{\sigma^2}{n}) \)

---

**Random Processes**

Temperature \( T \), heat conductivity \( \alpha(x, \omega) \)

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

\[ T(x, t, \omega) = \begin{cases} T_0(x) & t < 0, T(x, 0) = T_0(x) \\ \frac{\partial T}{\partial x}, & t > 0, T(x, t, \omega) \end{cases} \]

\( T(x, t, \omega) \) is a random variable for all \( x, t \)

**Stochastic Process**

\( X_t = \{ X(t) : t \in T \} \)

\( T \) is interval (continuous process)

or consecutive integers (discrete process)

\( X_t(\omega) \) realization of process \( t \in T \), \( \omega \in \Omega \)

Sample path or trajectory of \( \omega \)
Second order stochastic process: \( E(X_t^2) < \infty \)
\[
\begin{align*}
\mu(t) &= E(X_t), \quad t \in T \\
C(t,s) &= \text{Cov}(X_t, X_s) = E((X_t - \mu(t))(X_s - \mu(s))) \\
&= \begin{cases} \\
&\text{non-negative} \\
&\text{symmetric} \end{cases} \\
&t, s \in T
\end{align*}
\]

Markov chain
Discrete state space \( S \), \( X = \{ x_i \mid i \in \mathbb{Z} \} \)

Markov property: \( X_{n+1} \) depends only on \( X_n \)
\[
P(X_{n+1} = x_{n+1} \mid X_n = x_n, \ldots, X_0 = x_0) = P(X_{n+1} = x_{n+1} \mid X_n = x_n)
\]
\( x_i \) is the state at time \( i \)
Finite discrete state space
\( S = \{ x_1, x_2, \ldots, x_k \} \), initial distribution \( p^0 \)
Transition probability \( p_{ij} = P(X_{n+1} = x_j \mid X_n = x_i) \)
\[
P^n = p^0 P \quad (\text{stationary})
\]

Limiting distribution
\[
\pi = \lim_{n \to \infty} P^n = \lim_{n \to \infty} p^0 P^n = \lim_{n \to \infty} p^0 P^{n+1} = (\lim_{n \to \infty} p^0 P^n) P = \pi P
\]

Random differential equations
Random effects via parameters, initial, or boundary conditions
\[
\frac{dZ(t)}{dt} = a(Z(t))Z + b(t; w), \quad Z(0) = Z_0 \quad \text{w} \text{eak. lec. 2}
\]

Stochastic differential equations
\[
\text{Wiener process}
\]
\[
\begin{align*}
&dZ(t) = -aZ(t)dt + b dW(t) \quad \text{or} \\
&Z(t) = Z_0 - \int_0^t aZ(s)ds + \int_0^t b dW(s) \\
&\text{Ito integral}
\end{align*}
\]

models Brownian motion.
Statistical inference

Given \( S = \{ x_1, x_2, \ldots, x_n \in \mathbb{R}^n \} \) realized realizations of \( X \)

want to infer probability distribution of \( X \)

parametric inference: prob. distribution has a 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

O.L.S or M.L. estimators

fixed parameter \( \eta \), \((1-\alpha)\times 100\% \) of confidence intervals contain \( \eta \)

Bayesian inference

parameter estimation solution is a posterior probability density

parameters are random variables

credible interval: has \((1-\alpha)\times 100\% \) chance of containing expected parameter

update probability when more data are known
Bayes' formula

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

Posterior density

\[
\pi(q | y) = \frac{\pi(y | q) \pi_0(q)}{\pi_y(y)} \quad \text{prior density}
\]

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

\[X\text{ is indicator function } \begin{cases} 1, & x \in I \\ 0, & x \notin I \end{cases}\]

\[
\pi(y | q) \text{ quantifies likelihood } L(q | y) \text{ of observing } y \text{ given parameter realization } q
\]

Joint density \[
\pi(q, y) = \pi(y | q) \pi_0(q)
\]

is normalized to 1 by \[
\pi_y(y)
\]

Posterior density \[
\pi(q | y) \text{ quantifies probability to get parameter } q \text{ given observations } y
\]

Model calibration with \[
\pi_y(y) = \int \pi(y | q) \pi_0(q) \, dq
\]

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

Conjugate priors

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