Mathematics and numerics for data assimilation and state estimation – Lecture 8

Summer semester 2020

Summary of lecture 7

- Random variables can be discrete, mixed or continuous.
- It is uniquely described by its distribution $\mathbb{P}_X$, and also by its cdf

$$F_X(x) = \mathbb{P}(X \leq x)$$

  and, if it exists, also by its pdf

$$\pi_X(x) = \frac{\mathbb{P}(X \in dx)}{dx}.$$

- expectation of X given Y can be expressed through use of the
  1. conditional probability $\mathbb{P}(X \in dx \mid Y)$ when Y is a discrete rv
  2. and the conditional density $\pi_{X|Y}$ when X and Y are continuous rv.

Overview

1. $L^2(\Omega)$, sub-σ-algebras and projections
2. Forward and inverse problems
   - Bayesian vs frequentist
3. Bayesian inversion
   - Bayesian methodology

Overview

1. $L^2(\Omega)$, sub-σ-algebras and projections

2. Forward and inverse problems
   - Bayesian vs frequentist

3. Bayesian inversion
   - Bayesian methodology

General definition for conditional expectation

- Mixed rv are neither discrete nor continuous.

Example 1

$X = YZ$ where $Y \sim \mathit{Bernoulli}(1/2)$ and $Z \sim U[0,1]$ with $Y \perp Z$.
Then formally,

$$\pi_X(x) = \frac{\delta_0(x) + \mathbb{1}_{[0,1]}(x)}{2}$$

- Conditional expectations cannot always be computed using conditional densities for mixed rv.

- **Objective:** obtain a unifying definition for conditional expectations valid for all types of rv.

Sigma algebra generated by Y

- For a discrete rv

$$Y(\omega) = \sum_{k=1}^{k} b_k \mathbb{1}_{B_k}(\omega)$$

on $(\Omega, \mathcal{F}, \mathbb{P})$, with $B_k = \{Y = b_k\}$ we define

$$\sigma(Y) := \sigma(\{B_k\}) = \text{smallest } \sigma\text{-algebra containing all events } B_1, B_2, \ldots$$

- By construction Y is $\sigma(Y)$-measurable and $\sigma(Y) \subset \mathcal{F}$.

- Then, for an integrable rv X, it holds that

$$\mathbb{E}[X|Y](\omega) = \begin{cases} \frac{1}{\mathbb{P}(B_1)} \int_{B_1} X \, d\mathbb{P} & \text{if } \omega \in B_1 \\ \frac{1}{\mathbb{P}(B_2)} \int_{B_2} X \, d\mathbb{P} & \text{if } \omega \in B_2 \\ \vdots \end{cases}$$

Observations: $\mathbb{E}[X|Y]$ is a $\sigma(Y)$-measurable discrete rv for which

$$\int_B X \, d\mathbb{P} = \int_B \mathbb{E}[X \mid Y] \, d\mathbb{P} \quad \forall B \in \sigma(Y).$$

(hint: verify first for sets B_k, and extend to general set $B \in \sigma(Y)$).

Seeking to preserve these properties, observe first that for $Y : (\Omega, \mathcal{F}) \to (\mathbb{R}^k, \mathcal{B}^k)$,

$$\sigma(Y) := \text{smallest } \sigma\text{-algebra containing } Y^{-1}(C) \quad \forall C \in \mathcal{B}^k$$

similarly satisfies $\sigma(Y) \subset \mathcal{F}$ and that Y is $\sigma(Y)$-measurable.

Definition 2 (Conditional expectation for general rv)

For rv $X : \Omega \to \mathbb{R}^d$ and $Y : \Omega \to \mathbb{R}^k$ defined on the same probability space, we define $\mathbb{E}[X \mid Y]$ as any $\sigma(Y)$-measurable rv Z satisfying

$$\int_B X \, d\mathbb{P} = \int_B Z \, d\mathbb{P} \quad \forall B \in \sigma(Y).$$

Conditioning on a $\sigma-$algebra

One may relate $\mathbb{E}[X \mid Y]$ to another kind of conditional expectation:

Definition 3 (Expectation of X given $\mathcal{V} \subset \mathcal{F}$.)

Let $X : \Omega \to \mathbb{R}^d$ be an integrable rv on a probability space $(\Omega, \mathcal{F}, \mathbb{P})$ and assume $\mathcal{V}$ is a σ-algebra $\mathcal{V} \subset \mathcal{F}$. Then we define $\mathbb{E}[X \mid \mathcal{V}]$ as any $\mathcal{V}$-measurable rv Z satisfying

$$\int_B X dP = \int_B Z \, dP \quad \forall B \in \mathcal{V}.$$

Observation: Setting $\mathcal{V} = \sigma(Y)$ implies that $\mathbb{E}[X \mid Y]$ satisfies the constraints of $\mathbb{E}[X \mid \sigma(Y)]$ and vice versa.

Question: Does $\mathbb{E}[X \mid \mathcal{V}]$ exist and is it unique?

Yes, $\mathbb{E}[X \mid \mathcal{V}] = \text{Proj}_{L^2(\Omega,\mathcal{V})} X$ is a.s. unique.

Function space $L^2(\Omega, \mathcal{F})$

As an extension of $L^2(\Omega)$ for discrete rv, we introduce the Hilbert space

$$L^2(\Omega, \mathcal{F}) = \left\{ X : (\Omega, \mathcal{F}) \to (\mathbb{R}^d, \mathcal{B}^d) \,\middle|\, \int_\Omega |X(\omega)|^2 \, dP < \infty \right\}$$

with the scalar product

$$\langle X, Y \rangle = \int_\Omega X \cdot Y dP = \int_{\mathbb{R}^d \times \mathbb{R}^d} X \cdot Y dF(x, y)$$

and norm

$$\|X\|_{L^2(\Omega, \mathcal{F})} = \sqrt{\langle X, X \rangle}.$$

This is a Hilbert space: it is complete and for any sub-σ-algebra $\mathcal{V} \subset \mathcal{F}$, $L^2(\Omega, \mathcal{V})$ is a closed subspace of $L^2(\Omega, \mathcal{F})$.

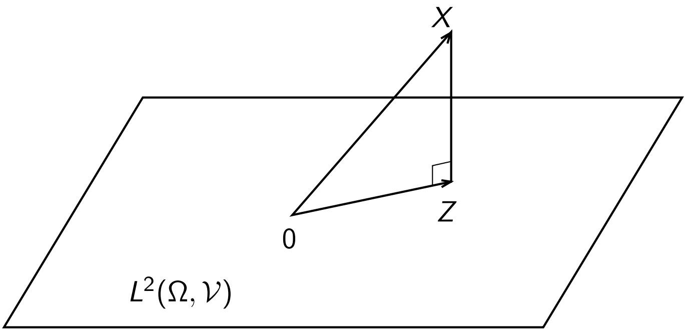
Orthogonal projections onto subspaces

Definition 4

The orthogonal projection of $X \in L^2(\Omega, \mathcal{F})$ onto the closed subspace $L^2(\Omega, \mathcal{V})$ is defined as any rv $Z \in L^2(\Omega, \mathcal{V})$ satisfying

$$\langle X - Z, W \rangle = 0 \quad \forall W \in L^2(\Omega, \mathcal{V}). \tag{1}$$

$$\text{We write} \quad Z = \text{Proj}_{L^2(\Omega,\mathcal{V})} X.$$



Orthogonal projections onto subspaces

Definition 4

The orthogonal projection of $X \in L^2(\Omega, \mathcal{F})$ onto the closed subspace $L^2(\Omega, \mathcal{V})$ is defined as any rv $Z \in L^2(\Omega, \mathcal{V})$ satisfying

$$\langle X - Z, W \rangle = 0 \quad \forall W \in L^2(\Omega, \mathcal{V}). \tag{1}$$

$$\text{We write} \quad Z = \text{Proj}_{L^2(\Omega,\mathcal{V})} X.$$

Exercise: verify that $\text{Proj}_{L^2(\Omega,\mathcal{V})} X$ satisfies the constraints of $\mathbb{E}[X \mid \mathcal{V}]$.
Hint: consider $W = \mathbb{1}_B$ for $B \in \mathcal{V}$

Exercise: verify that $Z = \text{Proj}_{L^2(\Omega,\mathcal{V})} X$ is unique in $L^2(\Omega, \mathcal{V})$ (and thus a.s. unique).

Last step: take as a fact that $\mathbb{E}[X \mid \mathcal{V}]$ satisfies the constraint (1) of $\text{Proj}_{L^2(\Omega,\mathcal{V})} X$, and conclude that $\mathbb{E}[X \mid \mathcal{V}]$ is a.s. unique.

Summary conditional expectations

Theorem 5

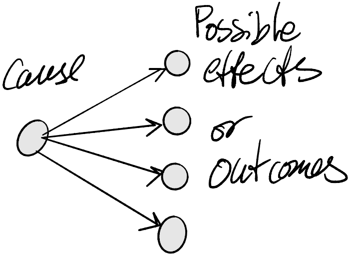
For rv $X : \Omega \to \mathbb{R}^d$ and $Y : \Omega \to \mathbb{R}^k$ defined on the same probability space and with $X \in L^2(\Omega, \mathcal{F})$, it holds that

$$\mathbb{E}[X \mid Y] = \mathbb{E}[X \mid \sigma(Y)] = Proj_{L^2(\Omega, \sigma(Y))} X.$$

Overview

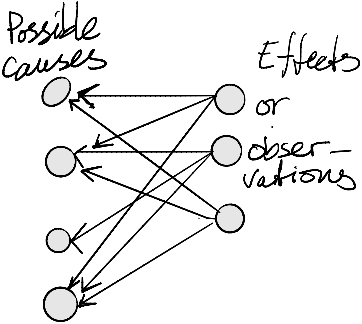
What are forward and inverse problems

Forward problem



Possible outcomes are deduced from cause

Inverse problem



Possible causes are induced from observations

A simple linear example

Forward problem: Given a matrix A and vector u compute the outcome

$$y = Au$$

Inverse problem: Given a matrix A and observation/effect y with either

(i) $y \notin \text{columnSpan}(A)$, or

(ii) $y \in \text{columnSpan}(A)$ but $Kernel(A) \neq \emptyset$,

then for (i), find the best approximate cause u to

$$Au = y$$

and for (ii), find the most suitable cause u to the above problem.

Well-posedness

Definition 6 (J. Hadamard 1902)

A problem is called **well-posed** if

1. a solution exists,
2. the solution is unique, and
3. the solution is stable with respect to small perturbations in the input.

On the other hand, if any of the above conditions are not satisfied, then the problem is **ill-posed**.

Example: The linear forward problem

$$y = Au$$

with fixed A and u is well-posed.

Well-posedness

Definition 6 (J. Hadamard 1902)

A problem is called **well-posed** if

1. a solution exists,
2. the solution is unique, and
3. the solution is stable with respect to small perturbations in the input.

On the other hand, if any of the above conditions are not satisfied, then the problem is **ill-posed**.

Example: The inverse problem:

$$Au = y$$

with fixed A and input y is well-posed if A is invertible and $|A^{-1}|$ is not too large. Since then for perturbed observations $y_\delta = y + \mathcal{O}(\delta)$,

$$|u - u_\delta| = |A^{-1}(y - y_\delta)| \leq C|A^{-1}|\delta.$$

Otherwise, it is not a well-posed problem.

Deterministic methods versus Bayesian inversion

We will consider extensions of the linear problem of the following form

$$Y = G(U) + \eta \tag{2}$$

where

- Y is the observation
- G is the possibly nonlinear forward model
- η is a perturbation/observation noise,
- U is the unknown parameter we seek to recover

Typical deterministic approach: View all variables as deterministic – also the perturbation. Find unique solution to the initially ill-posed problem (2) by introducing pseudo-inverse G^+ and solve

$$U = G^+(Y - \eta).$$

Bayesian approach: View all variables as random. Model your uncertainty through input prior $U \sim \pi_U$ and $\eta \sim \pi_\eta$. Solution is not a point in $\mathbb{R}^d$, but a posterior distribution: $\pi_{U|Y}(\cdot|y)$ (given observation $Y = y$).

Plan for this lecture

- Bayesian methodology for solving inverse problems.
- Introduce norms to study convergence of the posterior $\pi_{U|Y}$.
- Well-posedness for Bayesian inversion with perturbed input model $G_\delta \approx G$.

Bayesian vs frequentist

Definition 7 (Frequentist randomness)

The probability of an event is the long-term frequency of said event occurring when repeating an experiment multiple times:

$$\mathbb{P}(X \in A) = \lim_{M \to \infty} \sum_{i=1}^{M} \frac{\mathbb{1}_A(X_i)}{M} \quad \text{with } X_i \sim \mathbb{P}, \quad \text{and ideally independent.}$$

- Is applicable to repeatable experiments (coin flips, card games, ...), and to some degree to lagre data experiments (elections, survey polls, etc.)
- and to some degree in settings where imaginary sampling is deduced from from some form of prior information (e.g., physics argument for a coin flip being $Bernoulli(1/2)$).
- Dogmatically interpreted, not applicable to non-repeating experiments (e.g., probability that Barcelona wins a particular soccer match).

Bayesian vs frequentist

Definition 8 (Bayesian uncertainty)

$\mathbb{P}(X \in A \mid I)$ represents my degree of belief/confidence that A occurs given all my prior information I.

- **Constraint in defnition:** If you and I have the same prior information I, then your $\mathbb{P}(\cdot \mid I)$ should be the same as mine!
- The Bayesian approach leads to the same probability calculus as in the “usual” frequentist probability theory.
- More general than frequentist approach, as imaginary sampling really stems from a Bayesian viewpoint.
- Can assign probabilities/plausibility/belief to events which are either true or not (i.e., not at all random in the frequentist viewpoint):

$$\mathbb{P}(\text{John Doe committed the crime} \mid \text{evidence x, y, z})$$

Overview

Bayesian inverse problem

We consider the problem

$$Y = G(U) + \eta \tag{3}$$

where

- $U \in \mathbb{R}^d$ is the unknown parameter we seek to recover
- $Y \in \mathbb{R}^k$ is the observation
- $G : \mathbb{R}^d \to \mathbb{R}^k$ is the possibly nonlinear forward model
- η is the observation noise

Assumption 1

All parameters with the exception of G are random, described through

- $U \sim \pi_U$, $\eta \sim \pi_\eta$ and $Y \sim \pi_Y$,
- and $\eta \perp U$.

Objective: Given $Y = y$ use this to improve the estimate of the first component in the joint rv (U, Y) through determining $\pi_{U|Y}(\cdot|y)$.

Theorem 9 (Bayes theorem [Thm 1.2 Sanz-A.,Stuart, Taeb (SST)])

Let Assumption 1 hold and assume that $\pi_Y(y) > 0$. Then

$$U|Y = y \sim \pi_{U|Y}(\cdot|y)$$

with

$$\pi_{U|Y}(u|y) = \frac{\pi_\eta(y - G(u))\pi_U(u)}{\pi_Y(y)}. \tag{4}$$

Verfication: We may assume that also $\pi_U(u) > 0$, since otherwise (4) trivially holds. By the disintegration formula,

$$\pi_{UY}(u,y) = \pi_{U|Y}(u|y)\pi_Y(y) \quad \text{and} \quad \pi_{UY}(u,y) = \pi_{Y|U}(y|u)\pi_U(u).$$

And since $\pi_Y(y) > 0$, combining the above yields Bayes' rule for densities:

$$\pi_{U|Y}(u|y) = \qquad\qquad = \frac{\pi_{Y|U}(y|u)\pi_U(u)}{\pi_Y(y)}.$$

Since $Y|(U = u) = G(U) + \eta|(U = u) = G(u) + \eta$

$$\pi_{Y|U}(y|u) = \pi_{\eta+G(u)}(y) = \pi_\eta(y - G(u)).$$

Remarks

$$\pi_{U|Y}(u|y) = \frac{\overbrace{\pi_\eta(y - G(u))}^{\text{Likelihood function}}\ \pi_U(u)}{\pi_Y(y)}. \tag{5}$$

- The denominator $\pi_Y(y)$ in (5) acts as normalizing constant is called **the evidence** or the **marginal likelihood**:

$$Z := \pi_Y(y) = \int_{\mathbb{R}^d} \pi_\eta(y - G(u))\pi_U(u)\, du.$$

- $\pi_{U|Y}(u|y)$ is the **posterior density**
- To avoid clutter, we will drop density subscripts when reference is clear ($\pi(u) = \pi_U(u)$, $\pi(u|y) = \pi_{U|Y}(u|y)$ etc.)

Question: Given the posterior density, how can we produce a one-value estimate of the most plausible value of $U|Y = y$?

Posterior mean and MAP estimators

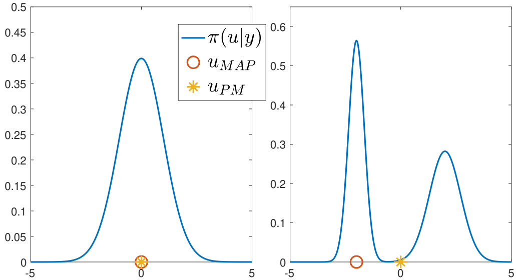
Definition 10

The posterior mean of U given $Y = y$ is defined by

$$u_{PM} := \mathbb{E}[U|Y = y] = \int_{\mathbb{R}^d} u\pi(u|y)du$$

and the maximum a posteriori (MAP) estimator is defined by

$$u_{MAP} := \arg\max_{u \in \mathbb{R}^d} \pi(u|y).$$



Example 11

Let $\eta \sim N(0, \gamma^2)$ and model $G(u) = u$ and prior

$$\pi(u) = \frac{\mathbb{1}_{(-1,1)}(u)}{2}$$

Given an observation $Y = y$, Bayes' theorem yields

$$\pi(u|y) = \frac{\pi_\eta(y-u)\pi_U(u)}{\pi_Y(y)} = \frac{\mathbb{1}_{(-1,1)}(u)\exp(-(y-u)^2/2\gamma^2)}{2Z}$$

with normalizing constant Z.
This yields:

$$u_{MAP} = \arg\max_{u\in\mathbb{R}} \pi(u|y) = \begin{cases} y & \text{if } y \in (-1,1) \\ -1 & \text{if } y \leq -1 \\ 1 & \text{if } y \geq 1 \end{cases}, \quad \text{and} \quad u_{PM} = ?$$

Assimilating two observations

Example 12

Consider the ordinary differential equation

$$\dot{x}(t;u) = x(t;u) \quad t > 0 \quad \text{and} \quad x(0;u) = u$$

and

$$G(u) = x(1;u) = ue^1,$$

and assume we have two different observations

$$Y_1 = G(U) + \eta_1, \quad \text{and} \quad Y_2 = G(U) + \eta_2,$$

with the prior density $U \sim U[-1, 4]$, and $\eta_1 \sim N(0, 1)$ and $\eta_2 \sim U[-0.5, 0.5]$ with $\eta_1 \perp \eta_2$.

Problem: Compute the posterior density for $U|(Y_1 = 0.2, Y_2 = -0.4)$.

Solution to Example 12

Set $Y = (Y_1, Y_2)$ with $\eta = (\eta_1, \eta_2)$ and apply Theorem 1 to the joint rv (U, Y), for the observation $y = (0.2, -0.4)$.

$$\pi(u|y) = \frac{\pi_\eta((y_1 - G(u), y_2 - G(u))\pi_U(u)}{Z}$$
$$= \frac{\pi_{\eta_2}(y_2 - G(u)))\pi_{\eta_1}((y_1 - G(u))\pi_U(u)}{Z}$$

Motivation: For $\pi_U(u) > 0$,

$$Y|(U = u) = (G(u) + \eta_1, G(u) + \eta_2)$$
$$\implies \pi(y|u) = \pi_{\eta_1\eta_2}(y_1 - G(u), y_2 - G(u)).$$

and

$$\pi(u|y) =$$

Observation: The posterior $\pi(u|y_1, y_2)$ can be obtained in two ways:

- In one go: $U|(Y_1 = y_1, Y_2 = y_2)$ mapping $\pi_U(u)$ to $\pi(u|y_1, y_2)$:

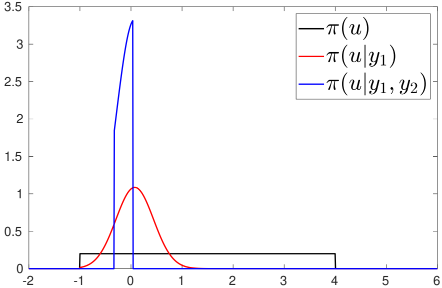
$$\pi(u|y) = \frac{\pi_\eta((y_1 - G(u), y_2 - G(u)))\pi_U(u)}{Z}$$

- Or sequentially: 1. $U|(Y_1 = y_1)$ mapping $\pi_U(u)$ to $\pi(u|y_1)$:

$$\pi(u|y_1) = \frac{\pi_{\eta_1}(y_1 - G(u))\pi_U(u)}{Z_1}$$

and 2. $U|(Y_1 = y_1, Y_2 = y_2)$ mapping $\pi(u|y_1)$ to $\pi(u|y_1, y_2)$

$$\pi(u|y_1, y_2) = \frac{\pi_{\eta_2}(y_2 - G(u))\pi(u|y_1)}{Z_2}$$



Well-posedness for Bayesian inversion

Relation between observation and underlying parameter

$$Y = G(U) + \eta$$

Inverse problem: what is the most likely/plausible U given $Y = y$

Bayesian inversion solution: the posterior density $\pi(u|y)$, or a function of the density, e.g., u_{PM} and u_{MAP}.

Hadamard's definition of well-posedness requires that a solution (i) exists, (ii) is unique and (iii) **is stable with respect to small perturbations of the input**.

By construction, $\pi(u|y)$ exists and is unique as long as $\pi_Y(y) > 0$.

Objective: Study condition (iii) under perturbations in the model. We seek a result along the lines of

$$|G_\delta - G| = \mathcal{O}(\delta) \implies d(\pi^\delta(\cdot|y), \pi(\cdot|y)) = \mathcal{O}(\delta), \quad \text{but what is } d?$$

Next time

- Assumptions on the noise η and perturbations G_δ that gives stability,
- Perturbed forward problems G_δ to which said assumptions apply,
- Bayesian inversion in the linear setting with Gaussian distributions.