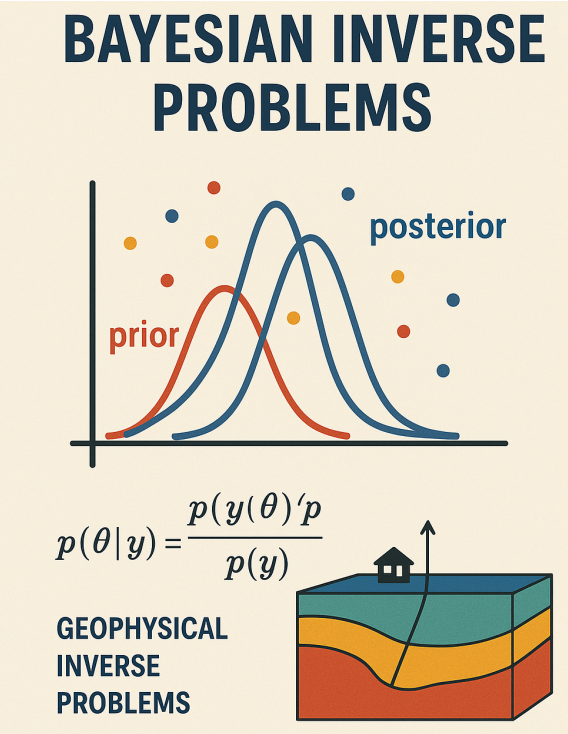
Bayesian Inverse Problems

Mark Asch - MAKUTU/2025



Outline of the course

1. Introduction to inverse problems and data assimilation: overview, setting, history, definitions, examples.

2. Bayesian inverse problems.

   (a) Bayesian inference.
   (b) Bayesian/Statistical inversion theory.
   (c) Full wave inversion example.
   (d) Point and interval estimates.

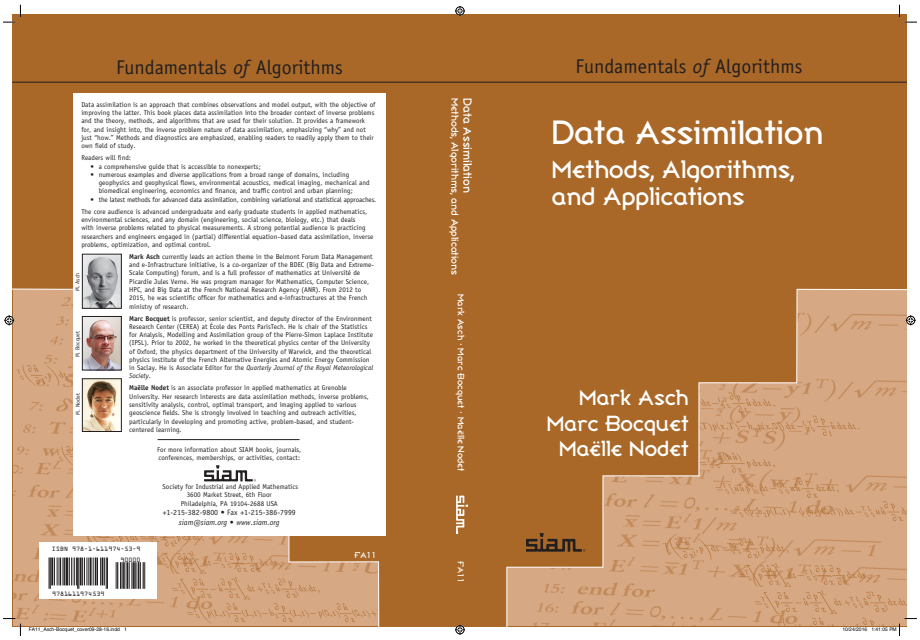
3. Posterior Estimation methods.

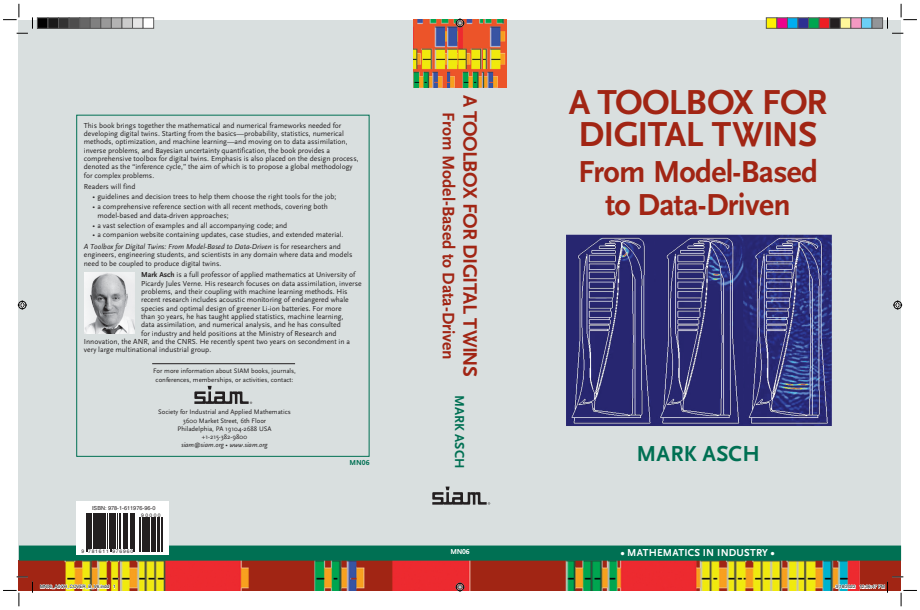
   (a) Monte Carlo methods.
   (b) Rejection Sampling. Importance Sampling.
   (c) McMC and variants for posterior estimation.
   (d) Metropolis Hastings, Gibbs and Hamiltonian McMC.
   (e) Introduction to Variational Inference (VI) for posterior estimation.

4. Statistical estimation, Kalman filters and sequential data assimilation.

   (a) Introduction to statistical DA.
   (b) The Kalman filter and Ensemble KF.
   (c) Ensemble Kalman Inversion (EKI).

2 Reference Textbooks





3 Reference Textbooks

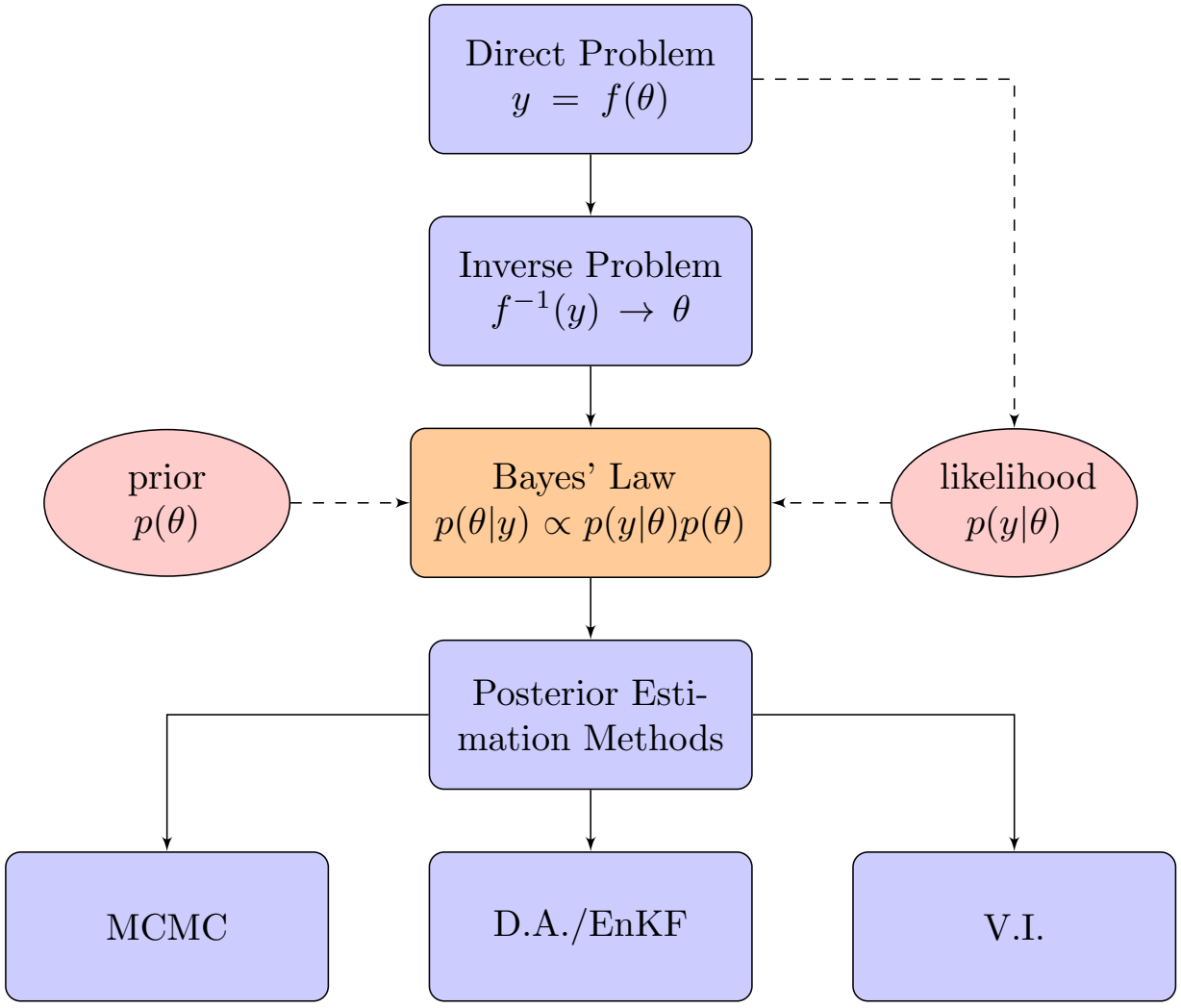
1. M. DeGroot, M. Schervisch. *Probability and Statistics*. Addison-Wesley, 2012.

2. P. Billingsley. *Probability and Measure*. John Wiley & Sons, 1995.

3. A. Gelman, et al. *Bayesian Data Analysis*. CRC Press. 2014.

INTRODUCTION

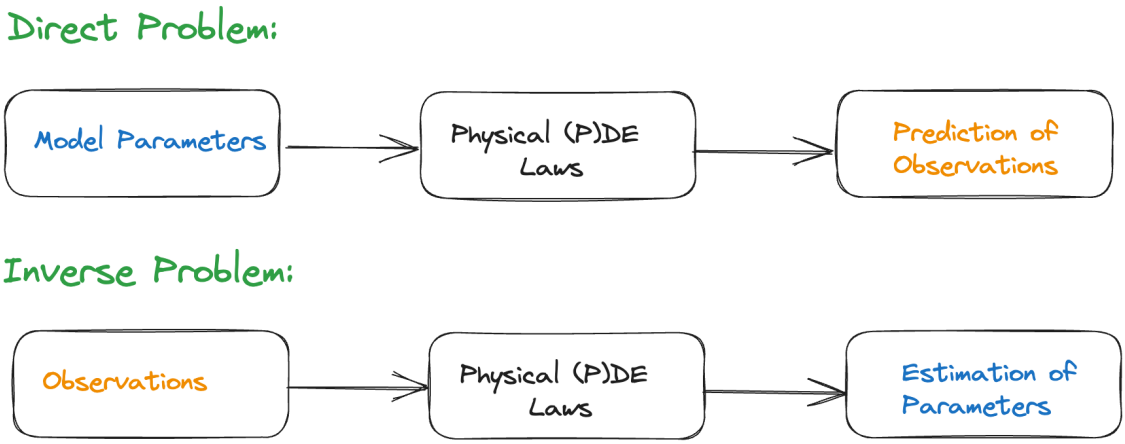
Recall: OVERVIEW



Notation: in the sequel we will replace the parameter θ by the solution of a parameter-dependent PDE, $u = u(\theta)$.

FORWARD AND INVERSE PROBLEMS

Recall: Classical Forward and Inverse Problems



- Consider a parameter-dependent dynamical system,

$$\frac{dz}{dt} = g(t, z; \theta), \qquad z(t_0) = z_0,$$

with g known, $\theta \in \Theta$, $z(t) \in \mathbb{R}^k$.

Forward: Given θ, z_0, find $z(t)$ for $t \geq t_0$.

Inverse: Given $z(t)$ for $t \geq t_0$, find $\theta \in \Theta$.

Stochastic Forward and Inverse Problems

- All real-life systems are subject to noise—in the "physics", in the measurements, in the model, in the simulations.

- Bayes' Theorem provides a posterior probability density that captures ALL the uncertainty.

Recall: Inverse Problems - General Formulation

- All inverse problems share a common formulation.
- Let the model parameters[1] be a vector (in general, a multivariate random variable), $\mathbf{m}$, and the data be $\mathbf{d}$,

$$\mathbf{m} = (m_1, \ldots, m_p) \in \mathcal{M},$$
$$\mathbf{d} = (d_1, \ldots, d_n) \in \mathcal{D},$$

where

$\Rightarrow$ $\mathcal{M}$ and $\mathcal{D}$ are the corresponding model parameter space and data space.

[1] Applied mathematicians often call the equation $G(m) = d$ a mathematical model and m the parameters. Other scientists call G the forward operator and m the model. We will adopt the more mathematical convention, where m will be referred to as the model parameters, G the model and d the data.

- The mapping $G\colon \mathcal{M} \to \mathcal{D}$ is defined by the direct (or forward) model

$$\mathbf{d} = g(\mathbf{m}), \tag{1}$$

where

$\Rightarrow$ $g \in G$ is an operator that describes the "physical" model and can take numerous forms, such as algebraic equations, differential equations, integral equations, or linear systems.

- Then we can add the observations or predictions, $\mathbf{y} = (y_1, \ldots, y_r)$, corresponding to the mapping from data space into observation space, $H : \mathcal{D} \to \mathcal{Y}$, and described by

$$\mathbf{y} = h(\mathbf{d}) = h\left(g(\mathbf{m})\right),$$

where

$\Rightarrow$ $h \in H$ is the observation operator, usually some projection into an observable subset of $\mathcal{D}$.

- Note that, in addition, there will be some random noise in the system, usually modeled as additive noise, giving the more realistic, stochastic direct model

$$\mathbf{d} = g(\mathbf{m}) + \epsilon, \tag{2}$$

where

$\Rightarrow$ ϵ is a random vector.

- In the sequel, we will also use the alternative Stuart's notation,

$$y = \mathcal{G}(u) + \eta,$$

where y is the observation vector, $\mathcal{G}$ is the forward model, *including* the observation operator, and η represents model and observation noise/uncertainty.

Recall: Inverse Problems - Classification

- We can now classify inverse problems as:

  $\Rightarrow$ deterministic inverse problems that solve (1) for $\mathbf{m}$,

  $\Rightarrow$ statistical/stochastic inverse problems that solve (2) for $\mathbf{m}$.

- The first class will be treated by linear algebra and optimization methods.

- The latter can be treated by a Bayesian (filtering) approach, and by weighted least-squares, maximum likelihood and DA techniques

- Both classes can be further broken down into:

  $\Rightarrow$ Linear inverse problems, where (1) or (2) are linear equations. These include linear sys-

tems—that are often the result of discretizing (partial) differential equations—and integral equations.

$\Rightarrow$ Nonlinear inverse problems where the algebraic or differential operators are nonlinear.

- Finally, since most inverse problems cannot be solved explicitly, computational methods are indispensable for their solution—see [Asch2022]

- Also note that we will be inverting here between the model and data spaces, that are usually both of high dimension and thus this model-based inversion will invariably be computationally expensive.

- This will motivate us to employ

  $\Rightarrow$ reduced order methods based on suitable projections,

  $\Rightarrow$ inversion between the data and observation spaces in a purely data-driven approach, using machine learning methods.

Statistical Inversion Theory

- Statistical inversion theory reformulates inverse problems as problems of statistical inference by means of Bayesian statistics.

- In Bayesian statistics all quantities and parameters are modeled as random variables.

- The randomness, which reflects the observer's uncertainty concerning their values, is coded in the probability distributions of the quantities/parameters.

- From the perspective of statistical inversion theory, the solution to an inverse problem is the probability distribution of the quantity of interest when all information available has been incorporated in the model.

- This distribution, the posterior distribution, describes the degree of confidence about the quantity of interest, *after* the measurement has been performed.

Statistical vs. Classical Inversion

- Regularization techniques, used in classical inversion, are typically aimed at producing a reasonable estimate of the quantities of interest based on the data available.

- In statistical inversion theory, the solution to an inverse problem is not a single estimate but a probability distribution that can be used to produce estimates.

- But satistical inversion gives more than just a single estimate: it can produce very different estimates and evaluate their reliability.

Recap: Statistical Inversion Principles

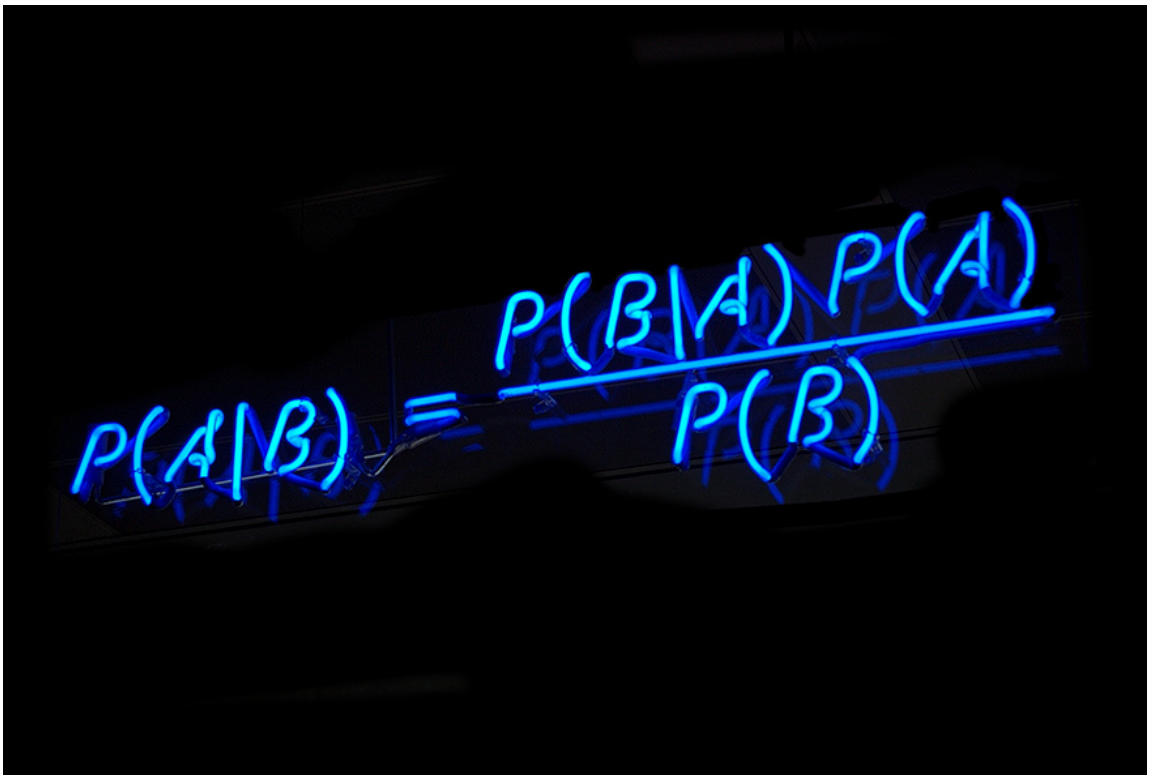
1. All variables included in the model are modelled as random variables.

2. The randomness describes our degree of information concerning their realizations.

3. The degree of information concerning these values is coded in the probability distributions.

4. The solution of the inverse problem is the posterior probability distribution.

BAYESIAN INFERENCE



Bayes' Law

- **Independence and Conditioning**:

Definition 1 (Independence). Two events, A and B, are independent if

$$\mathrm{P}(A \cap B) = \mathrm{P}(A)\mathrm{P}(B).$$

Definition 2 (Conditional Probability). The conditional probability of A on B is the probability that A occurs provided (or knowing) that B has occurred, and is given by

$$\mathrm{P}(A \mid B) = \frac{\mathrm{P}(A \cap B)}{\mathrm{P}(B)}.$$

- It follows that if A and B are mutually independent, then

$$\mathrm{P}(A \mid B) = \mathrm{P}(A), \quad \mathrm{P}(B \mid A) = \mathrm{P}(B).$$

- We can now state **Bayes' Theorem**

  $\Rightarrow$ in discrete form, for events

  $\Rightarrow$ in continuous form, for probability distributions

Theorem 1 (Bayes' Theorem for Two Events). *Let A and B be two events in Ω, then*

$$\mathrm{P}(A \mid B) = \frac{\mathrm{P}(B \mid A)\mathrm{P}(A)}{\mathrm{P}(B)}$$

or

$$\mathrm{P}(B \mid A) = \frac{\mathrm{P}(A \mid B)\mathrm{P}(B)}{\mathrm{P}(A)}.$$

- A more general situation is where Ω is subdivided into a partition of events, such as blood types in a human population.

Theorem 2 (Bayes' Theorem for a Partition). *Let $A_1, A_2, \ldots, A_k$ be a partition of Ω with $\mathrm{P}(A_i) > 0$ for each i. If $\mathrm{P}(B) > 0$, then, for each $i = 1, \ldots, k$,*

$$\mathrm{P}(A_i \mid B) = \frac{\mathrm{P}(B \mid A_i)\mathrm{P}(A_i)}{\mathrm{P}(B)},$$

where the total probability

$$\mathrm{P}(B) = \sum_{j=1}^{k} \mathrm{P}(B \mid A_j)\mathrm{P}(A_j).$$

Bayesian Inference

In the most general case, where we want to perform Bayesian inference for the estimation of parameters (an inverse problem!), we simply replace the probabilities by the corresponding density functions[2].

- Then Bayesian inference is performed in three steps:

  1. Choose a probability density $f(\theta)$, called the prior distribution, that expresses our beliefs, or prior experimental or historical knowledge, about a parameter θ before we see any data.
  2. Choose a statistical model $f(x \mid \theta)$ that reflects our beliefs about x given θ. Notice that this is expressed as a conditional probability, called the likelihood function, and not as a joint probability function.
  3. After observing data $x_1, \ldots, x_n$, update our beliefs and calculate the posterior distribution $f(\theta \mid x_1, \ldots, x_n)$.

[2] In the sequel we will use specific notation for prior, likelihood and posterior—here everything is simply denoted f.

- Let us look more closely at the three components of Bayes' Law.

Definition 3 (Prior Distribution)**.** For a given statistical model that depends on a parameter θ, considered as random, the distribution assigned to θ before observing the other random variables of interest is called the *prior distribution*. This is just the marginal distribution of the parameter.

Definition 4. [Posterior Distribution] For a statistical inference problem, with parameter θ and random sample $X_1, \ldots, X_n$, the conditional distribution of θ given $X_1 = x_1, \ldots, x_n = X_n$ is called the *posterior distribution* of θ.

Definition 5 (Likelihood Function)**.** Suppose that $X_1, X_2, \ldots, X_n$ have a joint density function

$$f(X_1, X_2, \ldots, X_n \mid \theta).$$

Given the observations $X_1 = x_1$, $X_2 = x_2$, $\ldots$,

$X_n = x_n$, the likelihood function of θ is

$$L(\theta) = L(\theta \mid x_1, x_2, \ldots, x_n) = f(x_1, x_2, \ldots, x_n \mid \theta).$$

If the X_i are i.i.d. with density $f(X_i \mid \theta)$, , then the joint density is a product and

$$L(\theta \mid x_1, x_2, \ldots, x_n) = \prod_{i=1}^{n} f(x_i \mid \theta).$$

- We point out the following properties of the likelihood:

  $\Rightarrow$ The likelihood is not a probability density function and can take values outside the interval $[0, 1]$.

  $\Rightarrow$ Likelihood is an important concept in both frequentist and Bayesian statistics.

  $\Rightarrow$ Likelihood is a measure of the extent to which a sample provides support for particular values of a parameter in a parametric model—this will be very important when we will deal with parameter estimation, and inverse problems in general.

⇒ The likelihood measures the support (evidence) provided by the data for each possible value of the parameter. This means that if we compute the likelihood function at two points, $\theta = \theta_1$, $\theta = \theta_2$, and find that $L(\theta_1 \mid x) > L(\theta_2 \mid x)$, then the sample observed is more likely to have occurred if $\theta = \theta_1$. We say that θ_1 is a more plausible value for θ than θ_2.

⇒ For i.i.d. random variables, the log-likelihood is usually used, since it reduces the product to a sum.

General form of Bayes' Theorem

- We now formulate the general version of **Bayes' Theorem**.

Theorem 3. *Suppose that n random variables, $X_1, \ldots, X_n$, form a random sample from a distribution with density, or probability function in the case of a discrete distribution, $f(x \mid \theta)$. Suppose also that the unknown parameter, θ, has a prior pdf $f(\theta)$. Then the posterior pdf of θ is*

$$f(\theta \mid x) = \frac{f(x_1 \mid \theta) \cdots f(x_n \mid \theta) f(\theta)}{f_n(x)}, \quad (3)$$

where $f_n(x)$ is the marginal joint pdf of $X_1, \ldots, X_n$.

- In this theorem,

  $\Rightarrow$ the *prior*, $f(\theta)$, represents the credibility of, or belief in the values of the parameters

we seek, without any consideration of the data/observations;

$\Rightarrow$ the *posterior*, $f(\theta \mid x)$, is the credibility of the parameters with the data taken into account;

$\Rightarrow$ $f(x \mid \theta)$, considered as a function of θ, is the *likelihood* function, which is the probability that the data/observation could be generated by the model with a given value of the parameter;

$\Rightarrow$ the denominator, called the *evidence*, $f_n(x)$, is the total probability of the data taken over all the possible parameter values, also called the *marginal likelihood*, or the marginal, and can be considered as a normalization factor;

$\Rightarrow$ the posterior distribution is thus proportional to the product of the likelihood and the prior distribution, or, in applied terms,

$$f(\text{parameter} \mid \text{data}) \propto f(\text{data} \mid \text{parameter})\, f(\text{parameter}).$$

- What can one do with the posterior distribution thus obtained? The answer is a lot of things, in fact a complete quantification of the incertitude of the parameter's estimation is possible. We can compute:

$\Rightarrow$ Point estimates by summarizing the center of the posterior. Typically, these are the posterior mean or the posterior mode.

$\Rightarrow$ Interval estimates for a given level α—see below.

$\Rightarrow$ Estimates of the probability of an event, such as $\mathrm{P}(a < \theta < b)$ or $P(\theta > b)$.

$\Rightarrow$ Posterior quantiles.

Recap of the Bayesian Setting

We seek the posterior density of the parameters θ

$$\pi^y(\theta) := p(\theta|y) \propto \mathcal{L}(y, g(\theta))p(\theta).$$

Ingredients:

- Parameters $\theta \in \mathbb{R}^d$, data $y \in \mathbb{R}^n$.
- Prior density $p(\theta)\colon \mathbb{R}^d \to \mathbb{R}^+$.
- Forward model $g\colon \mathbb{R}^d \to \mathbb{R}^n$.
  - $\Rightarrow$ Can be a black-box function.
  - $\Rightarrow$ Each evaluation is expensive.
- Likelihood function $\mathcal{L}\colon \mathbb{R}^n \times \mathbb{R}^n \to \mathbb{R}^+$.
  - $\Rightarrow$ $\mathcal{L}(y, g(\theta)) = p(y|\theta)$, where y is the observation (data).
  - $\Rightarrow$ Each evaluation requires an evaluation of g.

Bayesian Framework for DEs

Advantages of Bayes

1. The solution to the Bayesian inverse problem, "*find* u *given* $y = G(u) + \eta$," is the posterior pdf $\pi^y(u)$. This allows for complete uncertainty quantification in the inferred parameter.

2. In the Bayesian framework, we can prove that this inverse problem is **well-posed**:

   (a) There *exists* a *unique* posterior distribution for all $y \in \mathbb{R}^{d_y}$.

   (b) If there are *multiple* minimizers of $\|y - G(u)\|_2^2$, then the posterior distribution has *multiple* modes.

   (c) The posterior distribution π^y *depends continuously* on y : if $L(y|u)$ is locally Lipschitz in y, then

$$d_{\mathrm{TV}}(\pi^y, \pi^{y'}) \leq C \|y - y'\|_2 .$$

Bayesian Framework

Challenges of Bayes

- The posterior distribution is typically not known in closed form—notable exception is the linear Gaussian case, which will be solved below and in exercises.

- Advanced sampling methods such as Markov chain Monte Carlo (McMC) methods are required for sampling from the posterior, e.g. for computing the posterior mean.

- Many variants of McMC exist, including so-called variational inference methods, where we seek an optimal posterior apprximation that minimizes a KL distance between two measures/pdf's.

- Finally, ensemble Kalman filters can be used for approximating the posterior.

EXAMPLES

Bayesian Approach

Example: linear, Gaussian, 1-D

Consider the simple, scalar (1-D) linear example,

$$y = gu + \eta$$

with $y \in \mathbb{R}^{d_y}$, $u, \eta \in \mathbb{R}^{d_u}$ and in this scalar case, $d_y = d_u = 1$.

- Forward model: $g \in \mathbb{R}$,
- Prior: $u \sim \mathcal{N}(0, \sigma_0^2)$, with Gaussian pdf

$$\pi_0(u) = \frac{1}{\sqrt{2\pi\sigma_0^2}} \exp\left(-\frac{\|u\|^2}{2\sigma_0^2}\right),$$

- Noise: $\eta \sim \mathcal{N}(0, \gamma^2)$, with likelihood function

$$L(y|u) = \frac{1}{\sqrt{2\pi\gamma^2}} \exp\left(-\frac{\|y - gu\|^2}{2\gamma^2}\right).$$

Using Bayes' Formula gives the posterior

$$\pi(u|y) := \pi^y(u) = \frac{L(y|u)\pi_0(u)}{\int_{\mathbb{R}} L(y|u)\pi_0(u)\,\mathrm{d}u},$$

which implies that the posterior law is

$$u|y \sim \mathcal{N}\left(\frac{\sigma_0^2 g}{\gamma^2 + \sigma_0^2 g^2} y,\ \sigma_0^2 \frac{\gamma^2}{\gamma^2 + \sigma_0^2 g^2}\right).$$

- The posterior has

  $\Rightarrow$ a shifted mean, and
  $\Rightarrow$ a smaller variance than the prior.

Proof

The result for the 1-D posterior is readily obtained by completing the square in the expression for the posterior pdf. Leaving out constants, we have prior and likelihood

$$\pi_0(u) \propto \exp\left(-\frac{1}{2\sigma_0^2}u^2\right)$$

$$L(y|u) \propto \exp\left(-\frac{1}{2\gamma^2}(y-gu)^2\right)$$

and by Bayes, the posterior is (regrouping powers of u)

$$\begin{aligned}\pi^y(u) &\propto \exp\left(-\frac{1}{2\sigma_0^2}u^2 - \frac{1}{2\gamma^2}(y-gu)^2\right)\\ &= \exp\left(-\frac{1}{2}\left[\frac{1}{\sigma_0^2}+\frac{g^2}{\gamma^2}\right]u^2 - 2\frac{gy}{\gamma^2}u + \frac{y^2}{\gamma^2}\right)\\ &\doteq \exp\left(-\frac{1}{2}\left[au^2-2bu+c\right]\right),\end{aligned}$$

where

$$a = \frac{1}{\sigma_0^2}+\frac{g^2}{\gamma^2}, \quad b = \frac{gy}{\gamma^2}, \quad c = \frac{y^2}{\gamma^2}.$$

We seek constants m, K and σ such that the posterior

$$\pi^y(u) \propto \exp\left(-\frac{1}{2\sigma^2}(u-m)^2 + K\right)$$

that are found by completing the square, as follows

$$\begin{aligned} au^2 - 2bu + c &= a\left(u^2 - 2\frac{b}{a}u + \frac{c}{a}\right) \\ &= a\left(u^2 - 2\frac{b}{a}u + \left(\frac{b}{a}\right)^2 - \left(\frac{b}{a}\right)^2 + \frac{c}{a}\right) \\ &= a\left(u - \frac{b}{a}\right)^2 + \left(c - \frac{b^2}{a}\right) \end{aligned}$$

and thus, by identification, the posterior variance and mean are

$$\sigma^2 = \frac{1}{a} = \frac{\sigma_0^2\gamma^2}{\gamma^2 + g^2\sigma_0^2}, \quad m = \frac{b}{a} = \frac{\sigma_0^2 g}{\gamma^2 + \sigma_0^2 g^2} y$$

and the constant K (whose exponential will be absorbed into the proportionality)

$$K = c - \frac{b^2}{a} = \frac{y^2}{\gamma^2} - \frac{\sigma_0^2 g^2 y^2}{y^4 + y^2\sigma_0^2 g^2} \qquad \blacksquare$$

Bayesian Approach

Example: linear, Gaussian, multi-D

Generalizing the 1-D example, we get the following.

- Forward model: $G \in \mathbb{R}^{d_u \times d_y}$,
- Prior: $u \sim \mathcal{N}(m_0, C_0)$
- Noise: $\eta \sim \mathcal{N}(0, \Gamma)$, with likelihood function

$$L(y|u) = \frac{1}{\sqrt{2\pi\gamma^2}} \exp\left(-\frac{1}{2}(y - Gu)^\top \Gamma^{-1}(y - Gu)\right).$$

Using Bayes' Formula gives the posterior law

$$u|y \sim \mathcal{N}(m, C),$$

with

$$m = m_0 + C_0 G^\top \left(G C_0 G^\top + \Gamma \right)^{-1} (y - G m_0) \,,$$
$$C = C_0 - C_0 G^\top \left(G C_0 G^\top + \Gamma \right)^{-1} G C_0 .$$

BIP
THEORY

(Re)Recap of Bayesian Setting

We seek the posterior density of the parameters θ

$$\pi^y(\theta) := p(\theta|y) \propto \mathcal{L}(y, g(\theta))p(\theta).$$

Ingredients:

- Parameters $\theta \in \mathbb{R}^d$, data $y \in \mathbb{R}^n$.
- Prior density $p(\theta)\colon \mathbb{R}^d \to \mathbb{R}^+$.
- Forward model $g\colon \mathbb{R}^d \to \mathbb{R}^n$.
  - $\Rightarrow$ Can be a black-box function.
  - $\Rightarrow$ Each evaluation is expensive.
- Likelihood function $\mathcal{L}\colon \mathbb{R}^n \times \mathbb{R}^n \to \mathbb{R}^+$.
  - $\Rightarrow$ $\mathcal{L}(y, g(\theta)) = p(y|\theta)$
  - $\Rightarrow$ Each evaluation requires an evaluation of g.

BAYESIAN INVERSE PROBLEMS

Bayesian Inverse Problem

Infinite Dimensional Case

"By formulating the theory and algorithms for Bayesian inversion on the underlying infinite-dimensional spaces, we obtain a framework suitable for rigorous analysis of the accuracy of reconstructions, of computational complexity, as well as naturally constructing algorithms which perform well under mesh refinement, since they are inherently well defined in infinite dimensions." [Stuart2010, Stuart2016]

- Bayesian approach to inverse problems provides a rigorous framework for the development of uncertainty quantification in the presence of data.
- Formulate Bayes' formula on a separable Banach space and study its properties in this infinite dimensional setting.
- The formulation of the Bayesian approach on a

separable Banach space has numerous benefits:

⇒ (i) it reveals an attractive well-posedness framework for the inverse problem, allowing for the study of robustness to changes in the observed data, or to numerical approximation of the forward model;

⇒ (ii) it allows for direct links to be established with the classical theory of regularization, which has been developed in a separable Banach space setting;

⇒ (iii) and it leads to new algorithmic approaches which build on the full power of analysis and numerical analysis to leverage the structure of the infinite-dimensional inference problem.

BIP Theory

Finite-dimensional Case

- For motivation (and simplicity), we start with a formulation in finite dimensions, $\mathbb{R}^n$.

- **Inverse Problem Formulation**: find $u \in \mathbb{R}^n$ from $y \in \mathbb{R}^m$, where u and y are related by the equation

$$y = G(u) + \eta \tag{4}$$

  $\Rightarrow$ y is the observed/measured data
  $\Rightarrow$ u is the unknown (which might be the parameter vector θ)
  $\Rightarrow$ $\eta \in \mathbb{R}^m$ represents noise/uncertainty in the observation
  $\Rightarrow$ $G\colon \mathbb{R}^n \to \mathbb{R}^m$ is the forward operator, and can be considered more precisely as the composition, $\mathcal{G} = O \circ G$, where O is the observation operator that defines exactly what can be measured.
  $\Rightarrow$ In the discrete (or discretized) BIP, we have a matrix operator $G\colon \mathbb{R}^{n \times m}$.

⇒ **Difficulty 1**: even if $m = n$ and G is invertible, the noise can cause $y \notin \operatorname{im} G$. Furthermore, the specific instance of η which enters the data may not be known to us; typically, at best, only the statistical properties of a typical noise are known. Thus we cannot subtract η from the observed data y to obtain something in $\operatorname{im} G$. Even if $y \in \operatorname{im} G$, the uncertainty caused by the presence of noise causes problems for the inversion.

⇒ **Difficulty 2**: when $m < n$, which is usually the case (fewer measurements than unknowns—we cannot measure everywhere), the system is underdetemined, and we have to attach a sensible meaning to the concept of solution in this case where, generically, there will be many solutions.

⇒ **Statistical Inverse Problem:** Thinking probabilistically enables us to overcome both of these difficulties. We treat u, y, and η as random variables and determine the joint probability distribution (u, y). Then define the solution of

the inverse problem to be the *conditional* probability distribution of u given y, denoted $u|y$ or $\pi^y(u)$. This enables us to

$\rightarrow$ model the noise via its statistical properties, where we do not need to know the exact instance of the noise entering the given data,

$\rightarrow$ specify a priori the form of solutions that we believe to be more likely, thereby enabling us to attach weights to multiple solutions which explain the data.

- This is the **Bayesian approach to inverse problems,** where we have reformulated the inverse problem as a *statistical inference* problem, and measurement, unknown and noise are all modeled as random variables. We automatically obtain a quantification of uncertainty by evaluating the spread of a posterior distribution.

BIP Theory

Recall: Intro to Probability Theory

Before describing the Bayesian theory, we recall some important definitions and results from measure theory applied to probability spaces.

Definition 6 (Probability Space). A probability space is a triplet $(\Omega, \mathcal{F}, \mathbb{P})$, where Ω is the sample space (universe), $\mathcal{F}$ is the σ-algebra of events, and $\mathbb{P}$ is the probability measure such that $\mathbb{P}(\Omega) = 1$.

- A measure is called σ-finite if Ω is a countable union of measurable sets with finite measure.

  $\Rightarrow$ Lebesgue measure on $\mathbb{R}^m$ is an example of a σ-finite measure.

  $\Rightarrow$ One intuitive way of thinking σ-algebras in probability theory is that they describe information.

  $\Rightarrow$ The σ-algebra contains the subsets representing the events for which we can decide, after the observation, whether they happened or not.

  $\Rightarrow$ Hence $\mathcal{F}$ represents all the information we can get from an experiment in $(\Omega, \mathcal{F}, \mathbb{P})$, while a

sub-σ-algebra $\mathcal{G} \subset \mathcal{F}$ represents partial information.

Definition 7 (Random Variable). Let $(X, \mathcal{B}(X))$ be a measurable space, with $\mathcal{B}(X)$ denoting the Borel σ-algebra generated by the open sets. We call a measurable mapping $x \colon \Omega \to X$ a random variable. Then x induces the following probability measure on X,

$$\mu(A) = \mathbb{P}\left(x^{-1}(A)\right) = \mathbb{P}\left(\omega \in \Omega \colon x(\omega) \in A\right),$$

where $A \in \mathcal{B}(X)$. The measure μ is called the probability distribution of x and we will denote $x \sim \mu$.

BIP Theory

Bayes' Theorem in Finite-dimensional Space

Define a random variable $(u, y) \in \mathbb{R}^n \times \mathbb{R}^m$ as follows:

- Let $u \in \mathbb{R}^n$ be a random variable with prior Π_0, and Lebesgue density $\pi_0(u)$.
- Let $\eta \perp u$ be a noise term, distributed according to P_0 with Lebesgue density $\rho(\eta)$.
- Let $y|u$ be defined by (4) where $G\colon \mathbb{R}^n \to \mathbb{R}^m$ is measurable.
- Then the conditional $y|u$ is obtained by simply shifting P_0 by $G(u)$ to the measure P_u with Lebesgue density $\rho^u(y) = \rho(y - G(u))$.
- Consequently, the joint random variable $(u, y) \in \mathbb{R}^n \times \mathbb{R}^m$ has Lebesgue density

$$\nu(u, y) = \rho(y - G(u))\pi_0(u)$$

thanks to the independence $\eta \perp u$ and the fact that we can rewrite (4) as $\eta = y - G(u)$.

This is encapsulated by Bayes' Theorem that allows us to calculate the distribution of the random variable $u|y$.

Theorem 4 (Bayes' Theorem Finite-Dim.). *Define the normalizing constant (evidence, marginal) as*

$$Z = \int_{\mathbb{R}^n} \rho\left(y - G(u)\right) \pi_0(u) \, \mathrm{d}u > 0.$$

Then $u|y$ is a random variable with Lebesgue density $\pi^y(u)$ given by

$$\pi^y(u) \doteq \pi(u|y) = \frac{\rho\left(y - G(u)\right) \pi_0(u)}{Z},$$

where $\pi_0(u)$ is the prior density, $\rho\left(y - G(u)\right)$ is the likelihood that measures the data misfit, and $\pi^y(u)$ is the posterior density that is a solution of the inverse problem (4) updating the prior with the given measurements.

- Define the potential Φ as the negative log likelihood,

$$\Phi(u; y) = -\log \rho\left(y - G(u)\right).$$

- Then, the conclusion of the theorem can be written as

$$\frac{\mathrm{d}\mu^y}{\mathrm{d}\mu_0}(u) = \frac{1}{Z} \exp\left(-\Phi(u; y)\right),$$

with

$$Z = \int_{\mathbb{R}^n} \exp\left(-\Phi(u; y)\right) \pi_0(u) \, \mathrm{d}u.$$

- Conclusion: the posterior is absolutely continuous with respect to the prior, and the Radon-Nikodym derivative is proportional to the likelihood—see below, in infinite-dimensional case, for all the definitions.

BIP Theory

Likelihood

- How do we derive the likelihood for the BIP? Let

$$y = f(u) + \eta$$

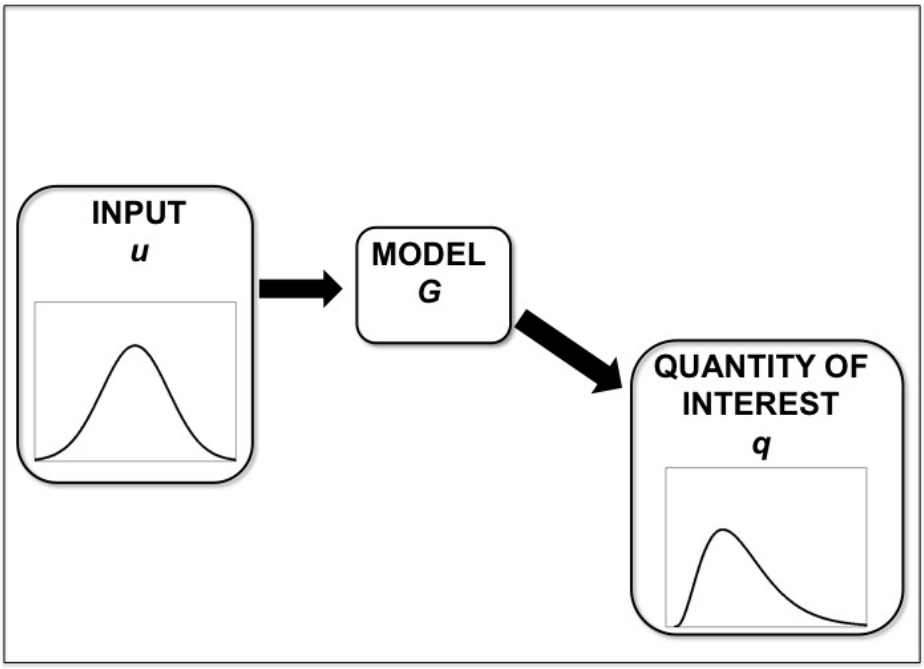
$\Rightarrow$ Define the probability distribution of the noise as π_η or $\mathrm{P}(\eta = n) = \mathrm{P}_\eta$

$\Rightarrow$ Fix a realization of u and compute the likelihood of the observation y, given u,

$$\begin{aligned}\mathrm{P}\left(f(u) + \eta = y | u\right) &= \mathrm{P}\left(f(u) + \eta = y\right) \\ &= \mathrm{P}\left(\eta = y - f(u)\right) \\ &= \mathrm{P}_\eta\left(y - f(u)\right)\end{aligned}$$

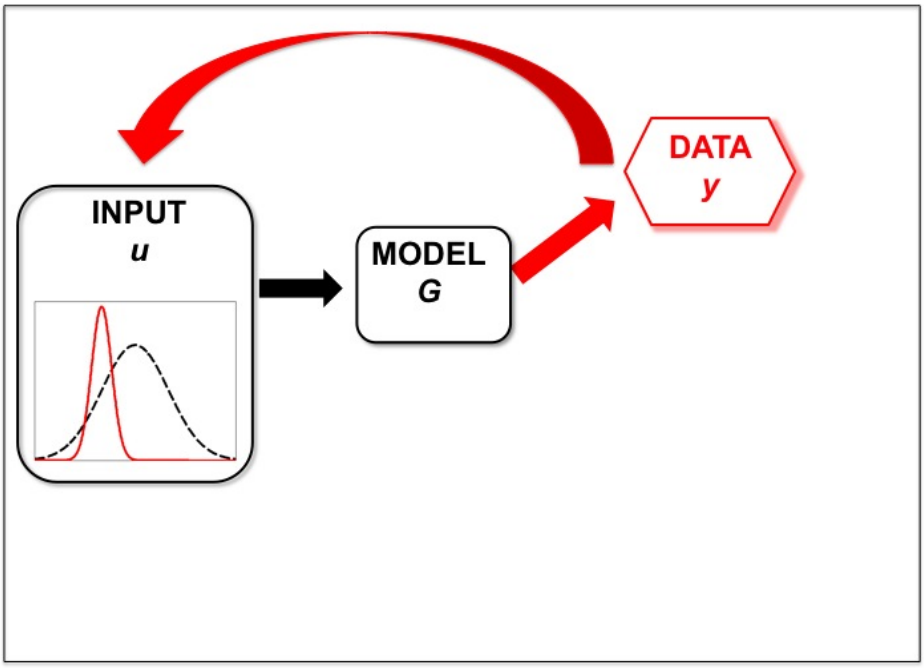
$\Rightarrow$ Hence, if we know the noise model P_η, we can immediately deduce the likelihood for the inverse problem.

UQ vs. BIP: vanilla UQ (black=prior)



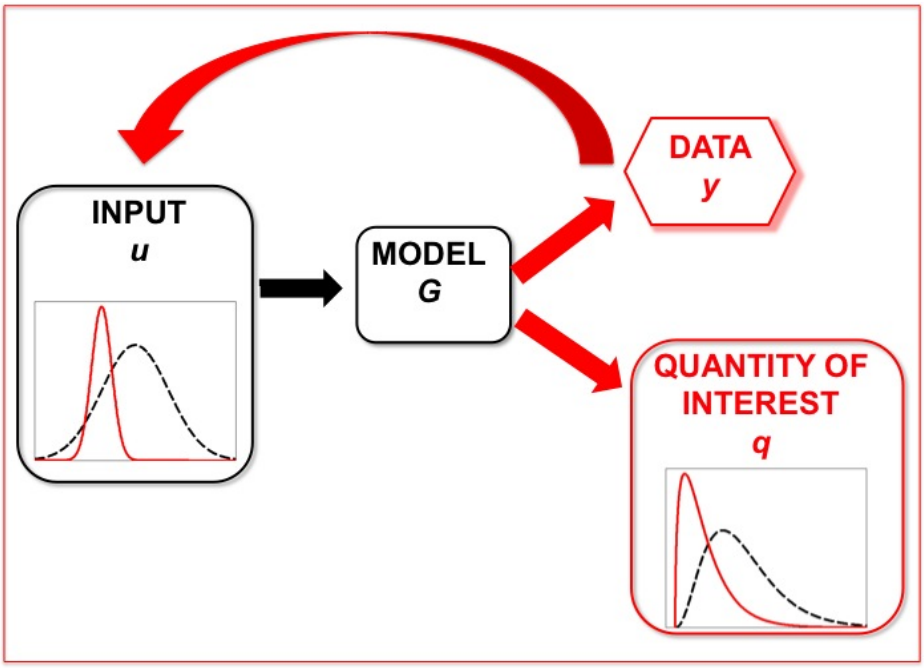
[A. Stuart]

BIP
(red=posterior)



[A. Stuart]

BIP + UQ
(red=posterior)



[A. Stuart]

BIP Examples: Finite-Dimensional Case

BIP Theory

Finite-Dimensional Examples: Scalar case

- Let $u \in \mathbb{R}$ be a scalar unknown, and suppose that we have measurements $y \in \mathbb{R}^k$, with $k \geq 1$. We are in the overdetermined case here.

- The measurement is given by the linear relation

$$y = Au + \eta,$$

  where the vector $A \in \mathbb{R}^k \setminus \{0\}$ and the noise $\eta \sim \mathcal{N}(0, \delta^2 I)$. with diagonal variance.

- We model the unknown prior of u as a standard Gaussian measure $\mathcal{N}(0, 1)$.

- Then by Bayes' Theorem, the posterior

$$\pi^y(u) \propto \exp\left(-\frac{1}{2\delta^2}\|y - Au\|^2 - \frac{1}{2}|u|^2\right).$$

- This posterior is Gaussian, and we can show that its mean and covariance are

$$\mu_\delta = \frac{A^\top y}{\delta^2 + \|A\|^2}, \quad \sigma_\delta^2 = \frac{\delta^2}{\delta^2 + \|A\|^2}.$$

- In the zero-noise limit, we get

$$\bar{\mu} = \lim_{\delta \to 0} \mu_\delta = \frac{A^\top y_0}{\|A\|^2}, \quad \bar{\sigma}^2 = \lim_{\delta \to 0} \sigma_\delta^2 = 0.$$

- The point $\bar{\mu}$ is exactly the least-squares solution (a regression) for the linear equation $y = Au$ and the prior plays no role in the limit of zero observational noise.

BIP Theory

Finite-Dimensional Examples: Vector case

- Consider now the case where the unknown is vector-valued and we have only a scalar measurement. That is, $u \in \mathbb{R}^n$ with $n \geq 2$, and $y \in \mathbb{R}$, a scalar.

- The measurement is given by the linear relation

$$y = A^\top u + \eta,$$

  where the vector $A \in \mathbb{R}^n \setminus \{0\}$.

- For the noise, we assume $\eta \sim \mathcal{N}(0, \delta^2)$ and for the prior $u \sim \mathcal{N}(0, \Sigma_0)$.

- Then by Bayes' Theorem, the posterior

$$\pi^y(u) \propto \exp\left(-\frac{1}{2\delta^2}\left|y - A^\top u\right|^2 - \frac{1}{2}u^\top \Sigma_0^{-1} u\right).$$

- Once again, this posterior is Gaussian, and we can show that its mean and covariance are

$$\mu_\delta = \frac{y\Sigma_0 A}{\delta^2 + A^\top \Sigma_0 A}, \quad \Sigma_\delta = \Sigma_0 - \frac{(\Sigma_0 A)\,(\Sigma_0 A)^*}{\delta^2 + A^\top \Sigma_0 A}.$$

- In the zero-noise limit, we get

$$\bar{\mu} = \lim_{\delta \to 0} \mu_\delta = \frac{y_0 \Sigma_0 A}{A^\top \Sigma_0 A},$$
$$\bar{\Sigma} = \lim_{\delta \to 0} \Sigma_\delta = \Sigma_0 - \frac{(\Sigma_0 A)\,(\Sigma_0 A)^*}{A^\top \Sigma_0 A}.$$

- We note that

$$\bar{\mu}^\top A = y_0, \quad \bar{\Sigma} A = 0,$$

so when the observational noise decreases:

$\Rightarrow$ Knowledge of u in the direction of A becomes certain, but the uncertainty remains in directions not aligned with A.

$\Rightarrow$ The magnitude of this uncertainty is determined by interaction between the properties of the prior and forward operator A.

$\Rightarrow$ We see that in the underdetermined case the prior plays an important role even when the observational noise disappears.

BIP Theory

Finite-Dimensional Examples: Deblurring example

- Assume we have a discrete model,

$$y = Au + \eta$$

where y and u are discretized on the same mesh, and $u, y, \eta \in \mathbb{R}^n$.

- Assume the noise is Gaussian with diagonal variance $\delta^2 I$,

$$\eta \sim \mathcal{N}(0, \delta^2 I), \quad \rho(\eta) \propto \exp\left(-\frac{1}{2\delta^2} \|\eta\|^2 .\right)$$

- The likelihood density is given as

$$\rho^u(y) = \rho(y - Au) \propto \exp\left(-\frac{1}{2\delta^2} \|y - Au\|^2 .\right)$$

- Choice of prior: assume that $u(0) = u(1) = 0$ and that u is quite smooth, that is, the value of $u(t)$ at a given point is more or less the same as its neighbour. We will then model the unknown as the noisy average,

$$u_j = \frac{1}{2}(u_{j-1} + u_{j+1}) + W_j,$$

where we take $W_j \sim \mathcal{N}(0, \gamma^2)$. The variance γ^2 determines how much the reconstructed function u deviates from the smoothness model average. We can write the discrete system in matrix form, $Lu = W$, where the tridiagonal matrix

$$L = \frac{1}{2}\begin{bmatrix} 2 & -1 & & & \\ -1 & 2 & -1 & & \\ & \ddots & \ddots & \ddots & \\ & & -1 & 2 & -1 \\ & & & -1 & 2 \end{bmatrix}.$$

- Prior can be rewritten as

$$\pi_0(u) \propto \exp\left(-\frac{1}{2\gamma^2}\|Lu\|^2 .\right)$$

- Posterior by applying Bayes' Formula is then

$$\pi^y(u) \propto \exp\left(-\frac{1}{2\delta^2}\|y - Au\|^2 - \frac{1}{2\gamma^2}\|Lu\|^2 .\right)$$

BIP Theory:
Infinite-Dimensional Case

BIP Theory

Recall: Normed, Separable Spaces

- A Banach space B is a complete normed vector space. In terms of generality, it lies somewhere in between a metric space M (that has a metric, but no norm) and a Hilbert space H (that has an inner-product, and hence a norm, that in turn induces a metric).

  $\Rightarrow$ Hilbert spaces are a subset of Banach spaces, where the additional structure of an inner product allows orthonormal bases (orthogonality), unitary operators, etc.

- More formally, if a space is endowed with an inner-product $\langle \cdot, \cdot \rangle$, then it induces a norm $\| \cdot \|$ as $\|x\| = \sqrt{\langle x, x \rangle}$, and if a space is endowed with a norm, then it induces a metric $d(x, y) = \|x - y\|$.

- By "complete" normed vector space, one usually means that every Cauchy sequence (with respect

to the norm) converges to a point that lies in the space.

- A metric space is called "separable" if it has a dense subset that is countable.

  $\Rightarrow$ A Hilbert space is separable iff it has a countable orthonormal basis.

- When the underlying space is simply $\mathbb{C}^n$ or $\mathbb{R}^n$, any choice of norm$\|\cdot\|_p$ for $1 \leq p \leq \infty$ yields a Banach space, while only the choice $\|\cdot\|_2$ leads to a Hilbert space.

- Similarly, if (X, Ω, μ) is a probability space, then the following space is a Banach space $L_p(X, \Omega, \mu) := f : X \to \mathbb{C}$ such that f is Ω-measurable and $\int |f(x)|^p \mathrm{d}\mu(x) < \infty$ with norm $\|f\|_p := \left(\int |f(x)|^p \mathrm{d}\mu(x)\right)^{1/p}$ (with $f = g$ meaning that they are equal μ-a.e.).

- When $X = \mathbb{R}$ or $X = \mathbb{C}$ and μ is the

Lebesgue measure, we sometimes just write $L_p := f : \mathbb{C} \to \mathbb{C}$ such that $\int |f(x)|^p \, \mathrm{d}x < \infty$.

BIP Theory

Recall: Radon-Nikodym Derivative and Theorem

- **Motivation**: Consider two probability measures on the same measurable space. When can we express one measure as a "density" with respect to another? This fundamental question arises naturally in:

  ⇒ *Probability theory*: Relating different probability distributions.
  ⇒ *Statistics*: Change of variables in transformations.
  ⇒ *Bayesian inference*: Computing posterior distributions from priors and likelihoods.

- The Radon-Nikodym theorem provides the mathematical framework to answer this question rigorously.

- Let $(\Omega, \mathscr{F}, \mu)$ be a measure space, where

  ⇒ Ω is the sample space,

$\Rightarrow$ $\mathscr{F}$ is a σ-algebra of measurable sets,
$\Rightarrow$ μ is a measure on $(\Omega, \mathscr{F})$.

- Absolute Continuity

Definition 8 (Absolutely Continuous). Let μ and ν be two measures on the same measure space. The measure ν is **absolutely continuous** with respect to μ (written $\nu \ll \mu$) if for every $A \in \mathscr{F}$,

$$\mu(A) = 0 \Longrightarrow \nu(A) = 0.$$

Intuition: ν assigns zero measure to any set that μ considers negligible.

- σ-Finite Measures

Definition 9. A measure μ is **σ-finite** if there exists a sequence $\{A_n\}_{n=1}^{\infty}$ of measurable sets such that: $\Omega = \cup_{n=1}^{\infty} A_n$ and $\mu(A_n) < \infty$ for all n.

- **Examples**

$\Rightarrow$ Lebesgue measure on $\mathbb{R}$ is σ-finite.

$\Rightarrow$ Any finite measure (like probability measures) is σ-finite.

$\Rightarrow$ Counting measure on $\mathbb{R}$ is not σ-finite.

Theorem 5 (Radon-Nikodym Theorem). *Let $(\Omega, \mathscr{F}, \mu)$ be a σ-finite measure space, and let ν be a σ-finite measure on the same measure space $(\Omega, \mathscr{F})$. Then $\nu \ll \mu$ if and only if there exists a non-negative, $\mathscr{F}$-measurable function $f \in L^1_\mu$ (also denoted as a density) such that*

$$\nu(A) = \int_A f \, \mathrm{d}\mu \quad \textit{for all} \quad A \in \mathscr{F}.$$

Moreover, f is unique μ-almost everywhere.

- Key Points:

1. *Existence*: The function f exists when absolute continuity holds.

2. *Uniqueness*: f is determined uniquely up to sets of μ-measure zero.

3. *Notation*: this unique density f is called the **Radon-Nikodym derivative** of ν with respect to μ, formally written as

$$f \doteq \frac{\mathrm{d}\nu}{\mathrm{d}\mu}.$$

- *Geometric Interpretation*: we can think of the Radon-Nikodym derivative as a "local scaling factor" that tells us how ν "stretches" or "compresses" μ at each point, or it is the Jacobian for the change-of-variable in an integral.

- Conditional Probability theory:

  $\Rightarrow$ The probability density function of a random variable is the Radon–Nikodym derivative of the induced measure with respect to some base measure (usually the Lebesgue measure for continuous random variables). See example below.

⇒ In probability theory, conditional probability densities are essentially Radon-Nikodym derivatives. Given random variables X and Y with joint density $f_{X,Y}(x,y)$, the conditional density of Y given $X = x$ is

$$f_{Y|X=x}(y) = \frac{f_{X,Y}(x,y)}{f_X(x)}.$$

This is the Radon-Nikodym derivative of the joint measure (numerator) with respect to the marginal measure (denominator).

⇒ *Rigorous construction*: Let $(\Omega, \mathscr{F}, P)$ be a probability space and X, Y be random variables.

1. *Joint measure*: Define $\mu_{X,Y}(A \times B) = P(X \in A, Y \in B)$.
2. *Marginal measure*: Define $\mu_X(A) = P(X \in A)$.
3. *Conditional measure*: For fixed x, define $\mu_{Y|X=x}(B) = P(Y \in B | X = x)$.

The conditional density $f_{Y|X=x}$ is the Radon-

Nikodym derivative

$$f_{Y|X=x} = \frac{\mathrm{d}\mu_{Y|X=x}}{\mathrm{d}\lambda}.$$

where λ is a reference measure, usually Lebesgue measure.

- Relation with McMC and Importance Sampling: (this will be important when developing rigorous numerical methods for estimating the posterior in Bayesian Inversion—see next lecture)

  $\Rightarrow$ *McMC*: The acceptance probability in Metropolis-Hastings involves Radon-Nikodym derivatives.

  $\Rightarrow$ *Importance Sampling*: Weights are ratios of densities, i.e. Radon-Nikodym derivatives.

BIP Theory

Recall: Example of a Probability Density Function

Example 1. Let μ be a probability measure on $(X, \mathcal{B}(X))$, with $X = \mathbb{R}^d$ and suppose that $\mu \ll \nu_L$, where ν_L is the standard Lebesgue measure on $\mathbb{R}^d$. Since ν_L is σ-finite, we can use Theorem 5 and conclude that there exists a function $f \in L^1\left(\mathbb{R}^d\right)$ such that for any $A \in \mathcal{B}\left(\mathbb{R}^d\right)$,

$$\mu(A) = \int_A f(t)\,\mathrm{d}t.$$

The function f is called the probability density of $x \sim \mu$.

BIP Theory

Bayes' Theorem in Infinite-dimensional Space

- Let X and Y be separable Banach spaces, equipped with the Borel σ-algebra, and let $G\colon X \to Y$ be a measurable mapping.

- Solve the inverse problem: find u from y where

$$y = G(u) + \eta$$

  and the noise $\eta \in Y$.

- Bayesian approach: let $(u, y) \in (X, Y)$ be a random variable and compute $u|y$ where we specify the random variable (u, y) as follows:

  $\Rightarrow$ Prior: $u \sim \mu_0$ measure on X.
  $\Rightarrow$ Noise: $\eta \sim \mathbb{Q}_0$ measure on Y with $\eta \perp u$.

- The random variable $y|u$ is then distributed according to the measure $\mathbb{Q}_u$, the translate of $\mathbb{Q}_0$ by $G(u)$.

- Assume that $\mathbb{Q}_u \ll \mathbb{Q}_0$ for u μ_0-a.s. Then for some potential $\Phi\colon X \times Y \to \mathbb{R}$,

$$\frac{\mathrm{d}\mathbb{Q}_u}{\mathrm{d}\mathbb{Q}_0}(y) = \exp\left(-\Phi(u;y)\right). \tag{5}$$

  Thus for fixed u, $\Phi(u;\cdot)\colon Y \to \mathbb{R}$ is measurable and for a given instance y of the data, $-\Phi(u;y)$ is called the log-likelihood.

- Define the product measure

$$\nu_0(\mathrm{d}u, \mathrm{d}y) = \mu_0(\mathrm{d}u)\mathbb{Q}_0(\mathrm{d}y)$$

  and assume that $\Phi(\cdot;\cdot)$ is ν_0-measurable.

- Then the random variable $(u, y) \in (X, Y)$ is distributed according to measure $\nu(\mathrm{d}u, \mathrm{d}y) = \mu_0(\mathrm{d}u)\mathbb{Q}_u(\mathrm{d}y)$ and it follows that $\nu \ll \nu_0$ with Radon-Nikodym derivative

$$\frac{\mathrm{d}\nu}{\mathrm{d}\nu_0}(u, y) = \exp\left(-\Phi(u;y)\right).$$

Theorem 6 (Bayes' Theorem Infinite-Dim.). *Assume that $\Phi\colon X \times Y \to \mathbb{R}$ is ν_0-measurable and that for a given y we have $\mathbb{Q}_0$-a.s.*

$$Z = \int_X \exp\left(-\Phi(u;y)\right) \mu_0(u)\, \mu_0(\mathrm{d}u) > 0. \tag{6}$$

Then the conditional distribution of $u|y$ exists under ν and is denoted by μ^y. Furthermore $\mu^y \ll \mu_0$ and for y ν-a.s.

$$\frac{\mathrm{d}\mu^y}{\mathrm{d}\mu_0}(u,y) = \frac{1}{Z}\exp\left(-\Phi(u;y)\right). \tag{7}$$

Proof. Application of the definition of a conditional random variable in a Banach space setting. See [Dashti, Stuart]. □

BIP Theory

Bayes' Theorem Implementation in Infinite-dimensional Space

Four essential steps are required to apply Bayes Formula (7) to a given inverse problem:

1. Define a suitable prior measure μ_0 and noise measure $\mathbb{Q}_0$ whose independent product, form the reference measure ν_0.

2. Determine the potential Φ such that formula (5) holds.

3. Show that Φ is ν_0-measurable.

4. Show that the normalization constant Z given by (6) is positive almost surely with respect to $y \sim \mathbb{Q}_0$.

BIP Theory:

Applications

BIP Theory

Darcy Flow Problem

A stochastic inverse problem formulation for Darcy flow (and other pde's...) is:

- Consider Darcy flow with (log) permeability $u \in X = L^\infty(D)$ for the pressure p,

$$-\nabla \cdot (\exp(u)\nabla p) = 0, \quad x \in D \qquad (8)$$
$$u = g, \quad x \in \partial D$$

- Find $u \in X$, given noisy observations

$$y_j = p(x_j) + \eta_j, \qquad (9)$$

  where $\eta \sim \mathcal{N}(0, \Gamma)$ and the prior, μ_0 is a Gaussian measure on u

- Abstractly the inverse problem is: for $\mathcal{G} : X \mapsto Y = R^J$, find u given noisy measurements,

$$y = \mathcal{G}(u) + \eta, \ \text{noise}. \qquad (10)$$

Theorem 7 (A. Stuart). *Consider the Bayesian Inverse Problem (BIP) for* $u(x) = \ln k(x)$ *subject to observations of the form (9) where* p *solves the flow equation (8). Suppose that we have a prior measure* $\mu_0 = \mathcal{N}(0, \beta)$, *then* $\mu^y(\mathrm{d}u) = P(\mathrm{d}u \mid y)$ *is absolutely continuous with respect to* μ_0, *with Radon-Nikodym derivative*

$$\frac{\mathrm{d}\mu^y}{\mathrm{d}\mu_0}(u) \propto \exp\left(-\frac{1}{2}|y - \mathcal{G}(u)|_\Gamma^2\right)$$

with $\mathcal{G}$ *given by (10). The expression* $\mathrm{d}\mu^y/\mathrm{d}\mu_0$ *is precisely the* posterior probability density function *expressed in terms of measures.*

BIP Theory

1D Wave Equation

Consider the 1D wave equation

$$\frac{\partial v}{\partial t} + c(x)\frac{\partial v}{\partial x} = 0, \quad (x,t) \in \mathbb{R} \times (0,\infty), \quad (11)$$
$$v = f \quad (x,t) \in \mathbb{R} \times \{0\}.$$

The problem formulation is:

- Find $v \in X$ given noisy Eulerian observations

$$y_k = v(1,t_k) + \eta_k. \quad (12)$$

- Abstractly, for $\mathcal{G} : X \mapsto Y = \mathbb{R}^K$, find u given noisy measurements

$$y = \mathcal{G}(u) + \eta.$$

- All of these formulations can be captured again by the powerful Bayes' theorem.

Theorem 8 (A. Stuart). *Consider the Bayesian Inverse Problem (BIP) for* $u(x) = \ln c(x)$ *subject to observations of the form (12), where* v *solves the wave equation (11). Suppose that we have a prior measure* μ_0*; then* $\mu^y(\mathrm{d}u) = \mathrm{P}(\mathrm{d}u \mid y)$ *is absolutely continuous with respect to* $\mu_0(\mathrm{d}u)$*, with Radon-Nikodym derivative*

$$\frac{\mathrm{d}\mu^y}{\mathrm{d}\mu_0}(u) \propto \exp\left(-\frac{1}{2}|y - \mathcal{G}(u)|_\Gamma^2\right),$$

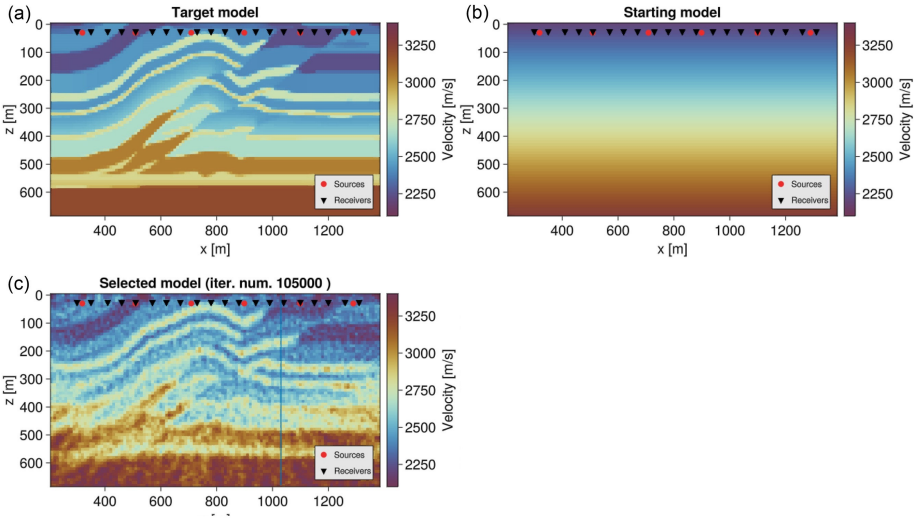
where $\eta \sim \mathcal{N}(0, \Gamma)$*. The expression* $\mathrm{d}\mu^y/\mathrm{d}\mu_0$ *is precisely the* posterior probability density function *expressed in terms of measures.*

- **Remark**: these results come directly from measure theory, which provides the exponential expression for the posterior.

APPLICATION:
Full Wave Inversion

Bayesian Full Wave Inversion:

Problem Setting



- Controlled seismic sources (explosives, vibrators, or airguns) generate acoustic/elastic waves.
- Waves propagate through the Earth's subsurface and interact with geological structures.
- Signals recorded by an array of receivers (geophones or hydrophones) at the surface or in boreholes.

- Inverse problem:

  $\Rightarrow$ Reconstruct high-resolution images of subsurface material properties from seismic wave measurements.

  $\Rightarrow$ Bayesian framework provides a principled approach to handle the inherent uncertainties and non-uniqueness in this problem.

Bayesian Full Wave Inversion:

Problem Formulation

- Elastic wave equation

$$\rho(\mathbf{x})\frac{\partial^2 u_i}{\partial t^2} = \frac{\partial \sigma_{ij}}{\partial x_j} + f_i$$

where:

$\Rightarrow$ $\rho(x)$ is the spatially varying density
$\Rightarrow$ u_i are the displacement components $(i; j = 1, 2, 3)$
$\Rightarrow$ σ_{ij} are the stress tensor components
$\Rightarrow$ f_i are the body force components
$\Rightarrow$ Einstein summation convention is implied

- Constitutive Relation

$$\sigma_{ij} = C_{ijkl}(\mathbf{x})\varepsilon_{kl}$$

- Strain-Displacement Relation

$$\varepsilon_{kl} = \frac{1}{2}\left(\frac{\partial u_k}{\partial x_l} + \frac{\partial u_l}{\partial x_k}\right)$$

- Complete Wave Equation (General Anisotropic)—substituting the constitutive and kinematic relations

$$\rho(\mathbf{x})\frac{\partial^2 u_i}{\partial t^2} = \frac{\partial}{\partial x_j}\left[C_{ijkl}(\mathbf{x})\frac{1}{2}\left(\frac{\partial u_k}{\partial x_l} + \frac{\partial u_l}{\partial x_k}\right)\right] + f_i$$

where the isotropic constitutive relation is

$$\sigma_{ij} = \lambda(\mathbf{x})\delta_{ij}\varepsilon_{kk} + 2\mu(\mathbf{x})\varepsilon_{ij}$$

with $\lambda(x)$ and $\mu(x)$ the spatially varying Lamé parameters, and δ_{ij} the Kronecker delta, or in

matrix form

$$\mathbf{C} = \begin{bmatrix} \lambda + 2\mu & \lambda & \lambda & 0 & 0 & 0 \\ \lambda & \lambda + 2\mu & \lambda & 0 & 0 & 0 \\ \lambda & \lambda & \lambda + 2\mu & 0 & 0 & 0 \\ 0 & 0 & 0 & \mu & 0 & 0 \\ 0 & 0 & 0 & 0 & \mu & 0 \\ 0 & 0 & 0 & 0 & 0 & \mu \end{bmatrix}$$

and we used Voigt notation to convert from $i, j, k, l = 1, 2, 3$ to a 6-dimensional system.

Bayesian Full Wave Inversion:

Problem Formulation—Parametrization

The subsurface model m typically represents:

- **Acoustic Case:**

$$m(x) = [\rho(x), \kappa(x)]^{\mathsf{T}}$$

or

$$m(x) = [\rho(x), c(x)]^{\mathsf{T}}$$

- **Elastic Case:**

$$m(x) = [\rho(x), V_p(x), V_s(x)]^{\mathsf{T}}$$

or

$$m(x) = [\rho(x), \lambda(x), \mu(x)]^{\mathsf{T}}$$

where

$\Rightarrow$ V_p is the P-wave velocity $\sqrt{((\lambda + 2\mu)/\rho)}$
$\Rightarrow$ V_p is the S-wave velocity $\sqrt{(\mu/\rho)}$

Bayesian Full Wave Inversion:

Problem Formulation—Forward and Inverse

- The forward modeling operator, G, expressed as a function of the parameter vector m, is then the solution $u(x_r, t; m)$ of the wave equation, measured at the receiver locations x_r,

$$y = G(m) + \eta,$$

where η represents (additive) observation noise (or just uncertainty, in general).

- The measurement vector $\mathbf{y}$ consists of

$$y = [u(x_1, t_1; m), u(x_1, t_2; m), ..., u(x_{N_r}, t_{N_t}; m)]^{\mathsf{T}}$$

- Inverse problem: given measurements, y, find the material parameters m.

Bayesian Full Wave Inversion:

Problem Formulation—Prior, Likelihood, Posterior

We briefly present some pertinent priors and likelihood options—these will be discussed in more detail in the relevant sections below.

- Priors:
  - ⇒ Gaussian with spatial correlation, based on estimated horizontal and vertical correlation lengths.
  - ⇒ TV and Gaussian smoothness priors. TV favors piecewise-constant solutions with sharper boundaries.
  - ⇒ Geologically constrained priors based on facies, or rock physics.
  - ⇒ Hierarchical priors.

- Likelihoods:
  - ⇒ Additive Gaussian Noise Model—the most common assumption.

⇒ Noise Covariance Structure to account for spatial, temporal or frequency correlations.
⇒ Robust Likelihood Functions based on Student-t (heavy-tailed) or Laplace (L_1) distributions.
⇒ Amplitude and Phase Inversions.

- Posterior Distribution and Bayesian Inference:

  ⇒ Apply Bayes' Theorem

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

  ⇒ Log-posterior form:

$$\begin{aligned}\log \pi(m|y) &= \log \pi(y|m) + \log \pi_0(m) + \text{constant} \\ &= -\frac{1}{2} \|y - G(m)\|^2_\Sigma \\ &\quad - \frac{1}{2}(m - m_0)^\top C_0^{-1}(m - m_0) + \text{ct.}\end{aligned}$$

  where m_0 and C_0 are the mean and variance of the prior.

Bayesian Full Wave Inversion:

Numerical Solution

- Toy problems and Simplified versions, first!
- Existing packages:
  - ⇒ HMC Lab (Fichtner, ETHZ) HMClab
  - ⇒ HIPPy Lib (Gattaz, UT Austin) hIPPYlib
  - ⇒ VIP (Curtis, Edinburgh) VIP
  - ⇒ TorchFWI (Darve, Stanford) TorchFWI
  - ⇒ others...
- VI and Ensemble KF—see next lecture series...

Bayesian Full Wave Inversion:

Numerical Solution - Priors

Statement of model parameterization and prior uncertainty

Parameter Name	Parameter code	Amount of parameters	Type of uncertainty	Established from	Fields of science
Local model lithology	ma	1	Scenario	Geophysics wells	geophysics, petrophysics, geostatistics
Regional and local permeability	Kh	22	Log-normal pdfs	Head data Well data	hydrogeology
pressure boundary conditions	ch	5	Uniform	Experience	hydrology engineering
River flows and conductance	riv	8	Conductance: log-normal DEM: uniform	Experience from previous studies	River science
Drain conductance	drn	8	Conductance: log-normal DEM: uniform	Experience from previous studies	hydrology
Aquifer Recharge	rch	1	Trapezoidal	Base-flow estimates	hydrology, meteorology, climate science

[J. Caers]

Bayesian Full Wave Inversion:

Numerical Solution - Priors

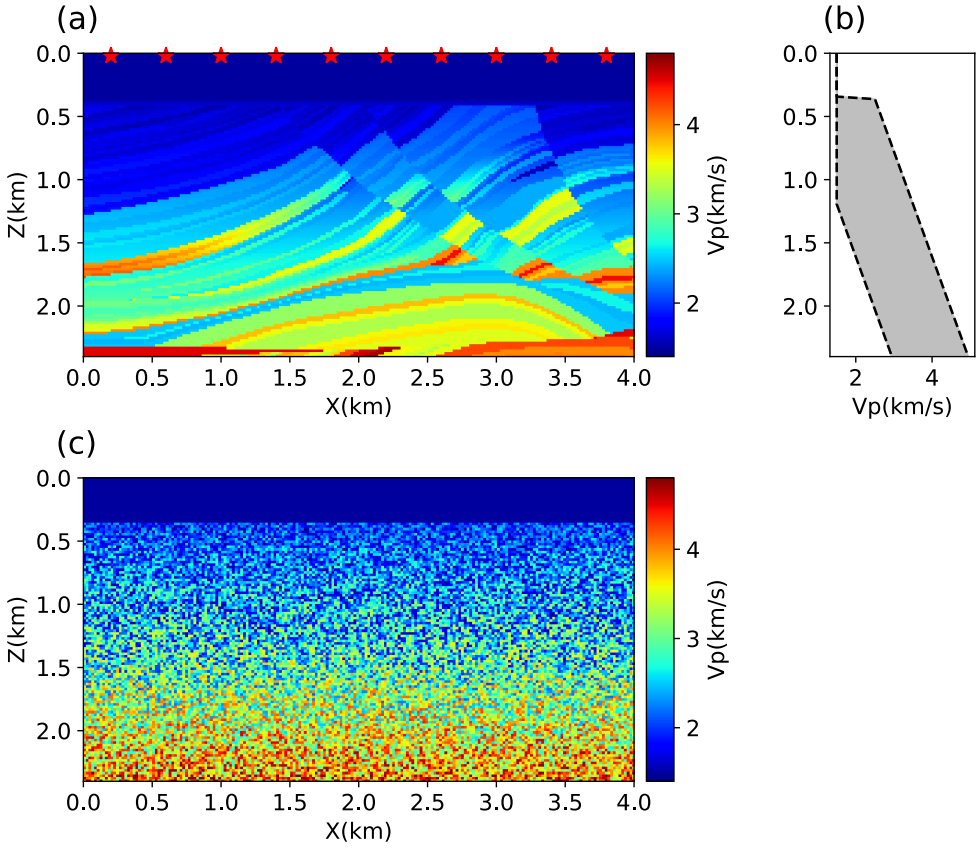


Figure 4. **(a)** The true model used in the full waveform inversion example. 10 sources are located at the depth of 20 m (red stars) and 200 receivers (not shown) are equally spaced at the depth of 360 m on the seabed. **(b)** The prior distribution of seismic velocity, which is set to be a Uniform distribution with an interval of 2 km/s at each depth. An additional lower bound of 1.5 km/s is also impose to the velocity to ensure that the rock velocity is higher than the velocity in water. **(c)** An example particle generated from the prior distribution.

[A. Curtis]

BIP Theory: Implementation– Estimators, Priors, Posterior

Bayesian Inverse Problem

Estimators

- The dimension of the inverse problem can be very large and consequently the posterior distribution lives in a high dimensional space—this makes its computation and visualisation very difficult.

- However, we can calculate different point estimators, and spread or region estimators.

  $\Rightarrow$ The point estimators approximate the most probable value of the unknown given the data and the prior.

  $\Rightarrow$ The spread estimators give a region that contain the unknown with some high probability.

- Estimation of the complete posterior is treated in detail below (McMC, VI, EnKF).

Bayesian Inverse Problem

Point Estimators

- One of the most used statistical estimators is the maximum a posterior estimate (MAP)

  $\Rightarrow$ The MAP is the mode of the posterior distribution. That is, given the posterior density $\pi^y(u)$ the MAP estimate u_{MAP} satisfies

  $$u_{\mathrm{MAP}} = \arg\max_{u \in \mathbb{R}^n} \pi^y(u)$$

  if it exists. The MAP may not be unique—eg. for bi-modal distributions.

- Another widely used point estimate is the conditional mean (CM) of the unknown u given the data y, which is defined by

  $$u_{\mathrm{CM}} = \mathrm{E}(u|y) = \int_{\mathbb{R}^n} u\, \pi^y(u)\, \mathrm{d}u,$$

  proivided that the integral converges.

Bayesian Inverse Problem

Spread Estimators

- The most common spread estimator for the BIP is the Bayesian Credible Interval or Set (BCI)

  $\Rightarrow$ A level $(1-\alpha)$ credible set, C_α, for given small $\alpha \in (0,1)$ is defined by

$$\Pi(C_\alpha|y) = \int_{C_\alpha} \pi^y(u)\,\mathrm{d}u = 1-\alpha.$$

  $\Rightarrow$ Thus, a credible set C_α is a region that contains a large proportion of the posterior mass.

- The random sets C_α that frequently contain the 'true' unknown $u^\dagger$ such that

$$\mathrm{P}(u^\dagger \in C_\alpha) = 1-\alpha,$$

  are known as confidence regions of level $(1-\alpha)$.

Bayesian Inverse Problem

Simple Estimator Examples

- See Practical #2.

Bayesian Inverse Problem

Estimator Examples–sum of Gaussians

- Point Estimators: let $u \sim \mathcal{N}(0,1)$, with density

$$\phi(u) = \frac{1}{\sqrt{2\pi}} \mathrm{e}^{-u^2/2}$$

and suppose that the posterior distribution is the sum of two Gaussians,

$$\pi^y(u) = \frac{c}{\sigma_1}\phi\left(\frac{u}{\sigma_1}\right) + \frac{1-c}{\sigma_2}\phi\left(\frac{u-1}{\sigma_2}\right),$$

where $0 < c < 1$ and $\sigma_1, \sigma_2 > 0$.

$\Rightarrow$ Conditional Mean: we obtain by taking expectation,

$$u_{\mathrm{CM}} = \mathrm{E}(u|y) = 1 - c.$$

$\Rightarrow$ MAP: we obtain by maximizing

$$u_{\mathrm{MAP}} = \begin{cases} 0, & \text{if } c/\sigma_1 > (1-c)/\sigma_2, \\ 1, & \text{if } c/\sigma_1 < (1-c)/\sigma_2. \end{cases}$$

$\Rightarrow$ Posterior covariance:

$$\begin{aligned} \sigma^2 &= \int_{-\infty}^{\infty} (u - u_{\mathrm{CM}})^2 \pi^y(u)\,\mathrm{d}u \\ &= \int_{-\infty}^{\infty} u^2 \pi^y(u)\,\mathrm{d}u - (u_{\mathrm{CM}})^2 \\ &= c\sigma_1^2 + (1-c)(\sigma_2^2 + 1) - (1-c)^2 \end{aligned}$$

Bayesian Inverse Problem

Estimator Examples--deblurring

Recall the finite-dimensional, deblurring problem, where we calculated the posterior distribution

$$\pi^y(u) \propto \exp\left(-\frac{1}{2\delta^2}\|y - Au\|^2 - \frac{1}{2\gamma^2}\|Lu\|^2 .\right)$$

- Since the posterior distribution is Gaussian, we know that the MAP and CM estimators coincide, and we have an estimator

$$\begin{aligned} u^\delta_{\text{MAP}} &= \arg\max_{u\in\mathbb{R}^n} \pi^y(u) \\ &= \arg\min_{u\in\mathbb{R}^n} \left\{ \frac{1}{2\delta^2}\|y - Au\|^2 + \frac{1}{2\gamma^2}\|Lu\|^2 \right\} \end{aligned}$$

- Notice that u_{MAP} is of the same form as the Tikhonov estimator...

- Completing the square we can write the posterior in the form

$$\pi^y(u) \propto \exp\left(-\frac{1}{2}\left\|u - \frac{1}{\delta^2}\Gamma^{-1}A^\top y\right\|_\Gamma^2\right),$$

where the weighted norm $\|\cdot\|_\Gamma = \left\|\Gamma^{1/2}\cdot\right\|$, with

$$\begin{aligned}\Gamma &= \frac{1}{\delta^2}\Gamma^{-1}A^\top y \\ &= \left(A^\top A + \frac{\delta^2}{\gamma^2}L^\top L\right)^{-1} A^\top y\end{aligned}$$

and the posterior covariance is

$$\Sigma = \Gamma^{-1}.$$

Bayesian Inverse Problem

Priors

Constructing a good prior density is one of the most challenging parts of solving a Bayesian inverse problem. The main problem is transforming our qualitative information into a quantitative form that can be coded as a prior density. The prior probability distribution should be concentrated on those values of u we expect to see and assign a clearly higher probability to them than to the unexpected ones.

- In inverse problems, prior information plays a key role.

  $\Rightarrow$ Broadly speaking, priors serve as regularizers.

- Intuitive idea:

  $\Rightarrow$ assign lower probability to neighborhoods of θ that you don't expect to see,

$\Rightarrow$ higher probability to neighborhoods of θ that you do expect to see.

- Examples:

  $\Rightarrow$ Uniform priors to respect known bounds.
  $\Rightarrow$ Gaussian processes with specified covariance kernel.
  $\Rightarrow$ Gaussian Markov random fields.
  $\Rightarrow$ Gaussian priors derived from differential operators.
  $\Rightarrow$ Hierarchical priors.
  $\Rightarrow$ Besov space priors.
  $\Rightarrow$ Other non-Gaussian priors.
  $\Rightarrow$ Higher-level representations (objects, marked point processes).

- Large samples: when the sample size is large, the likelihood outweighs the prior in determining the posterior, i.e., when the sample size is large, the prior is not crucial.

Bayesian Inverse Problem

Gaussian Priors

- Gaussian probability densities are the most widely-used priors in statistical inverse problems.

  $\Rightarrow$ They are easy to construct and form a versatile class of densities.
  $\Rightarrow$ They also often lead to explicit estimators.
  $\Rightarrow$ Thanks to the central limit theorem Gaussian densities are often good approximations to inherently non-Gaussian distributions when the observation is based on a large number of mutually independent random events. This is also the reason why the noise is often assumed to be Gaussian.

- The Gaussian noise model is the least restrictive one, and is based on observations of real systems. The only hypotheses are:

$\Rightarrow$ The signal has a central tendency that is zero (if it is non-zero, then it can simply be centered, and the bias is then absorbed into the deterministic part.)

$\Rightarrow$ Deviations around the central tendency are smaller and smaller as we move away from it.

Definition 10 (Multivariate Gaussian Distribution). Let $\theta \in \mathbb{R}^d$ and $\Sigma \in \mathbb{R}^{d \times d}$ be a symmetric positive definite matrix. A Gaussian d-variate random variable u with mean θ and covariance Σ is a random variable with the probability density

$$\pi(u) = \frac{1}{(2\pi\,|\Sigma|)^{d/2}} \exp\left(-\frac{1}{2}\,(u-\theta)^\top \Sigma^{-1}\,(u-\theta)\right),$$

where $|\Sigma| = \det(\Sigma)$, and we then denote

$$u \sim \mathcal{N}(\theta, \Sigma).$$

Bayesian Inverse Problem

Gaussian Prior Example

The important case where both the unknown u and the noise η are multivariate Gaussian, produces a Gaussian posterior—this is a case of so-called *conjugate priors*.

- Assume that $u : \Omega \to \mathbb{R}^n$ and $\eta : \Omega \to \mathbb{R}^m$ are mutually independent Gaussian random variables,

$$u \sim \mathcal{N}(\theta_u, \Sigma_u), \quad \eta \sim \mathcal{N}(0, \Sigma_\eta),$$

  where $\Sigma_u \in \mathbb{R}^{n \times n}$ and $\Sigma_\eta \in \mathbb{R}^{m \times m}$ are positive definite (covariance) matrices.

- Noisy measurements are given by

$$y = Au + \eta,$$

  where $y \in \mathbb{R}^m$ and $A \in \mathbb{R}^{m \times n}$ is a known matrix.

- Then the posterior probability density of u given the measurement y is

$$\pi(u|y) \propto \exp\left(-\frac{1}{2}(u-\bar{u})^\top \Sigma^{-1}(u-\bar{u})\right),$$

where the posterior mean and variance are given by

$$\bar{u} = \theta_u + \Sigma_u A^\top \left(A\Sigma_u A^\top + \Sigma_\eta\right)^{-1}(y - A\theta_u)$$

and

$$\Sigma = \Sigma_u - \Sigma_u A^\top \left(A\Sigma_u A^\top + \Sigma_\eta\right)^{-1} A\Sigma_u.$$

- The Sherman-Morrison-Woodbury formula provides a more compact form:

$$\Sigma = \left(A^\top \Sigma_\eta^{-1} A + \Sigma_u^{-1}\right)^{-1}$$

and

$$\bar{u} = \Sigma\left(A^\top \Sigma_\eta^{-1} y + \Sigma_u^{-1}\theta_u\right).$$

Posteriors

- See next Lecture

CONCLUSIONS

General

- Bayes' framework restores well-posedness to the (stochastic) inverse problem.

- This comes at a cost:

  $\Rightarrow$ theoretically more complex,
  $\Rightarrow$ computationally complex,
  $\Rightarrow$ requires new thinking—uncertainties and risks.

Post-Bayes Research

- 3 major limiting assumptions of Bayesian approach: [J. Knoblauch, UCL]

  $\Rightarrow$ (A1) Model is well-specified.
  $\Rightarrow$ (A2) Prior is well-specified.
  $\Rightarrow$ (A3) Computational feasibility.

- Generalized Variational Inference (see also V.I. in Lecture #3) can overcome all three...

References

1. A.M. Stuart. Inverse problems: a Bayesian perspective, *Acta Numerica*, 19 , pp. 451–559 , 2010.

2. M. Dashti, A.M. Stuart. The Bayesian approach to inverse problems, *Handbook of Uncertainty Quantification*, pp. 1–118, 2016

3. A. Tarantola. *Inverse problem theory and methods for model parameter estimation.* SIAM. 2005.

4. H. Igel. *Computational Seismology: A Practical Introduction.* Oxford University Press, 2017. website

5. A. Fichtner. *Full Seismic Waveform Modelling and Inversion.* Springer. 2011.