Bayesian Inverse Problems

Jonas Latz

`Input/Output: www.latz.io`

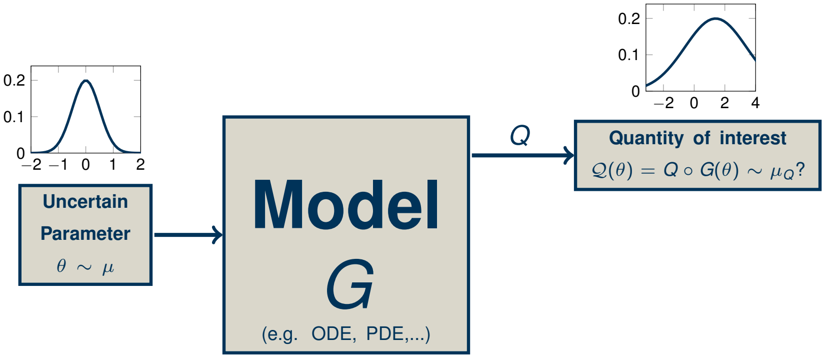
Technical University of Munich
Department of Mathematics, Chair for Numerical Analysis
Email: jonas.latz@tum.de

Garching, July 10 2018
Guest lecture in *Algorithms for Uncertainty Quantification* with Dr. Tobias Neckel and Friedrich Menhorn

Outline

Outline

Forward Problem: A Picture



Forward Problem: A few examples

Groundwater pollution.

G: Transport equation (PDE)

θ: Permeability of the groundwater reservoir

$\mathcal{Q}$: Travel time of a particle in the groundwater reservoir

Figure: Final disposal site for nuclear waste (Image: Spiegel Online)

Forward Problem: A few examples

Diabetes patient.

G: Glucose-Insulin ODE for a Diabetes-type 2 patient

θ: Model parameters such as exchange rate plasma insulin to interstitial insulin

$\mathcal{Q}$: Time to inject insulin

Figure: Glucometer (Image: Bayer AG)

Forward Problem: A few examples

Geotechnical Engineering

G: Deformation model

θ: Soil

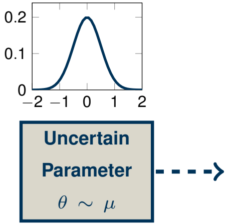
$\mathcal{Q}$: Deformation/Stability/probability of failure

Figure: Construction on Soil (Image: www.ottawaconstructionnews.com)

Distribution of the parameter θ

How do we get the distribution of θ?

Can we use data to characterise the distribution of θ?



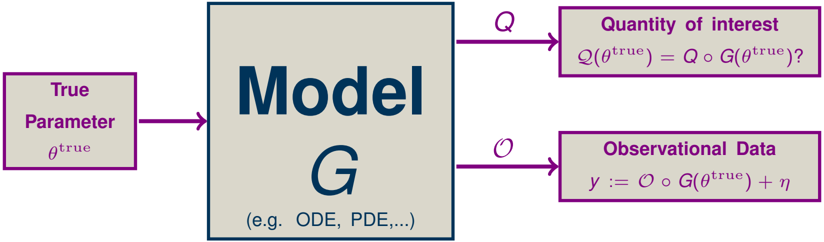
‘Data’?

Let θ^{true} be the actual parameter. We define data y by

$$y := \underbrace{\mathcal{O}}_{\text{Observation operator}} \circ \overbrace{G}^{\text{Model}} (\underbrace{\theta^{\text{true}}}_{\text{actual parameter}}) + \overbrace{\eta}^{\text{measurement noise}}$$

The measurement noise is a random variable $\eta \sim \mathrm{N}(0, \Gamma)$.

‘Data’? A Picture.



Identify θ^{true}?!

Can we use the data to identify θ^{true}?

$$\Longleftrightarrow$$

Can we solve the equation $y = \mathcal{O} \circ G(\theta^{\mathrm{true}}) + \eta$?

Identify θ^{true}?!

Can we solve equation $y = \mathcal{O} \circ G(\theta^{\text{true}}) + \eta$?

No. The problem is ill-posed.[1]

- The operator $\mathcal{O} \circ G$ is very complex
- $\dim(X \times Y) \gg \dim Y$, where $(X \times Y) \ni (\theta, \eta)$ and $Y \ni y$.

[1] Hadamard (1902) - *Sur les problèmes aux dérivés partielles et leur signification physique*, Princeton University Bulletin 13, pp. 49-52

Summary

We want to use noisy observational data y to find θ^{true}, but we cannot.

The uncertain parameter θ is still uncertain, even if we observe data y.

2 Questions:

How can we quantify the uncertainty in θ considering the data y?

How does this change the probability distribution of our Quantity of interest $\mathcal{Q}$?

Outline

An Experiment

We roll a die.

The sample space of this experiment is

$$\Omega := \{1, ..., 6\}.$$

The space of events is the power set of Ω:

$$\mathcal{A} := 2^{\Omega} := \{A : A \subseteq \Omega\}.$$

The probability measure is the Uniform measure on Ω:

$$\mathbb{P} := \mathrm{Unif}_{\Omega} := \sum_{\omega \in \Omega} \frac{1}{6} \delta_{\omega}.$$

An Experiment

We roll a die.

Consider the event $A := \{6\}$.

The probability of A is $\mathbb{P}(A) = 1/6$.

Now, an oracle tells us before rolling the die, whether the outcome would be even or odd.

$B := \{2, 4, 6\}$,
$B^c := \{1, 3, 5\}$.

How does the probability of A change, if we know whether B or B^c occurs?

$\rightarrow$ Conditional Probabilities

Conditional probabilities

Consider a probability space $(\Omega, \mathcal{A}, \mathbb{P})$ and two events $D_1, D_2 \in \mathcal{A}$, such that $\mathbb{P}(D_2) > 0$.
The conditional probability distribution of D_1 given the event D_2 is defined by:

$$\mathbb{P}(D_1|D_2) := \frac{\mathbb{P}(D_1 \text{ and } D_2)}{\mathbb{P}(D_2)} := \frac{\mathbb{P}(D_1 \cap D_2)}{\mathbb{P}(D_2)}$$

Conditional probabilities: Moving back to the experiment.

We roll a die.

$$\mathbb{P}(A|B) = \frac{\mathbb{P}(\{6\})}{\mathbb{P}(\{2,4,6\})} = \frac{1/6}{1/2} = \frac{1}{3},$$
$$\mathbb{P}(A|B^c) = \frac{\mathbb{P}(\emptyset)}{\mathbb{P}(\{1,3,5\})} = \frac{0}{1/2} = 0.$$

Probability and Knowledge

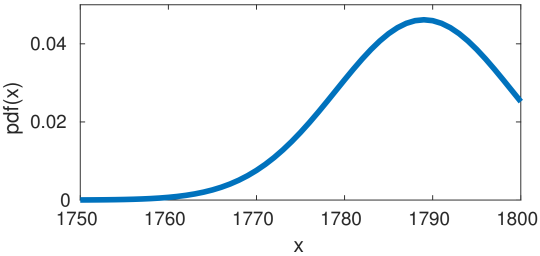
Probability distributions can be used to model knowledge.

When using a fair die, we have no knowledge whatsoever concerning the outcome:

$$\mathbb{P}(\{1\}) = \mathbb{P}(\{2\}) = \mathbb{P}(\{3\}) = \mathbb{P}(\{4\}) = \mathbb{P}(\{5\}) = \mathbb{P}(\{6\}) = 1/6$$

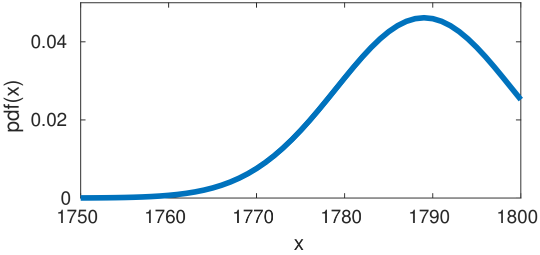
Probability distributions can be used to model knowledge.

When did the French revolution start? Rough knowledge from school: End of the 18th Century, definitely not before 1750/ after 1800.



Probability and Knowledge

Probability distributions can be used to model knowledge.



Here, the probability distribution is given by a probability density function (pdf), i.e.

$$\mathbb{P}(A) = \int_A \mathrm{pdf}(x)\mathrm{d}x$$

Conditional probability and Learning

We represent content we learn by an event $B \subseteq 2^{\Omega}$.

$$\text{Learning } B \text{ is a map } \mathbb{P}(\cdot) \mapsto \mathbb{P}(\cdot|B).$$

Conditional probability and Learning

We learn that $B = \{2, 4, 6\}$ occurs. Hence, we map

$$[\mathbb{P}(\{1\}) = \mathbb{P}(\{2\}) = \mathbb{P}(\{3\}) = \mathbb{P}(\{4\}) = \mathbb{P}(\{5\}) = \mathbb{P}(\{6\}) = 1/6]$$

$$\downarrow \qquad\qquad\qquad\qquad \downarrow$$

$$\begin{bmatrix} \mathbb{P}(\{1\}|B) = \mathbb{P}(\{3\}|B) = \mathbb{P}(\{5\}|B) = 0; \\ \mathbb{P}(\{2\}|B) = \mathbb{P}(\{4\}|B) = \mathbb{P}(\{6\}|B) = 1/3 \end{bmatrix}$$

But, how do we do this in general?

Elementary Bayes' Theorem

Consider a probability space $(\Omega, \mathcal{A}, \mathbb{P})$ and two events $D_1, D_2 \in \mathcal{A}$, such that $\mathbb{P}(D_2) > 0$. Then,

$$\mathbb{P}(D_1|D_2) = \frac{\mathbb{P}(D_2|D_1)\mathbb{P}(D_1)}{\mathbb{P}(D_2)}$$

Proof: We have

$$\mathbb{P}(D_1|D_2) = \frac{\mathbb{P}(D_1 \cap D_2)}{\mathbb{P}(D_2)} \text{ (1) and } \mathbb{P}(D_2|D_1) = \frac{\mathbb{P}(D_2 \cap D_1)}{\mathbb{P}(D_1)} \text{ (2) }.$$

(2) is equivalent to $\mathbb{P}(D_2 \cap D_1) = \mathbb{P}(D_2|D_1)\mathbb{P}(D_1)$, which can be substituted into (1) to get the final result.

Who is Bayes?

Figure: Bayes (Image: Terence O'Donnell, History of Life Insurance in Its Formative Years (Chicago: American Conservation Co:, 1936))

Thomas Bayes, 1701-1761

- English, Presbyterian Minister, Mathematician, Philosopher
- Proposed a (very) special case of Bayes' Theorem
- Not much known about him (the image above might be not him)

Who do we know Bayes' Theorem from?

Figure: Laplace (Image: Wikipedia)

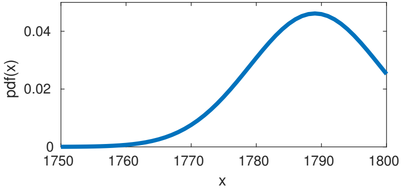
Pierre-Simon Laplace, 1749–1827

- French, Mathematician and Astronomer
- Published Bayes' Theorem in 'Théorie analytique des probabilités' in 1812

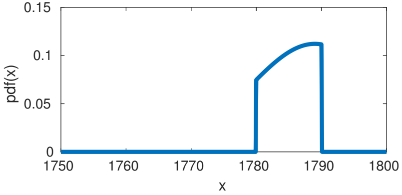
Conditional probability and Learning

When did the French revolution start?

(1) Rough knowledge from school: End of the 18th Century, definitely not before 1750/ after 1800.

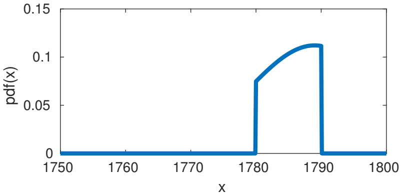


(2) Today in the radio : It was in the 1780s, so in the interval $[1780, 1790)$.



When did the French revolution start?

(2) Today in the radio : It was in the 1780s, so in the interval $[1780, 1790)$.



(Image: Wikipedia)

(3) Reading in a textbook: It was in the middle of year 1789.
Problem. The point in time x, we are looking for, is now set to a particular value $x = 1789.5 + \eta$, where $\eta \sim \mathrm{N}(0, 0.0625)$. Hence, the event we learn is $B = \{x + \eta = 1789.5\}$. But, $\mathbb{P}(B) = 0$. Hence $\mathbb{P}(\cdot|B)$ is not defined and Bayes' Theorem does not hold.

Non-elementary Conditional Probability

It is possible to define conditional probabilities for (non-empty) events B, with $\mathbb{P}(B) = 0$.
(rather complicated)

Easier: Consider the learning in terms of continuous random variables. (rather simple)

Conditional Densities

We learn a random variable x_1 and observe another random variable x_2

The joint distribution of x_1 and x_2 is given by a 2-dimensional probability density function $\mathrm{pdf}(x_1, x_2)$.

Given $\mathrm{pdf}(x_1, x_2)$ the marginal distributions of x_1, x_2 are given by

$$\mathrm{mpdf}_1(x_1) = \int \mathrm{pdf}(x_1, x_2)\mathrm{d}x_2; \quad \mathrm{mpdf}_2(x_2) = \int \mathrm{pdf}(x_1, x_2)\mathrm{d}x_1$$

We learn the event $B = \{x_2 = b\}$, for some $b \in \mathbb{R}$. Here, the conditional distribution is given by

$$\mathrm{cpdf}_{1|2}(x_1|x_2 = b) = \mathrm{pdf}(x_1, b)/\mathrm{mpdf}(b)$$

Bayes' Theorem for Conditional Densities

Similarly to the Elementary Bayes' Theorem, we can give a Bayes Theorem for Densities

$$\underbrace{\mathrm{cpdf}_{1|2}(\cdot|x_2 = b)}_{\text{posterior}} = \underbrace{\mathrm{cpdf}_{2|1}(b|x_1 = \cdot)}_{\text{(data) likelihood}}\,\underbrace{\mathrm{mpdf}_1(\cdot)}_{\text{prior}}\,/\,\underbrace{\mathrm{mpdf}_2(b)}_{\text{evidence}}$$

prior: Knowledge we have a priori concerning x_1
likelihood: The probability distribution of the data given x_1
posterior: Knowledge we have concerning x_2 knowing that $x_2 = b$
evidence: Assesses the model assumptions

Laplace’s formulation

ce qui est le principe énoncé ci-dessus, lorsque toutes les causes sont *à priori* également possibles. Si cela n’est pas, en nommant p la probabilité *à priori* de la cause que nous venons de considérer; on aura $E = Hp$; et en suivant le raisonnement précédent, on trouvera

$$P = \frac{Hp}{S.Hp};$$

ce qui donne les probabilités des diverses causes, lorsqu’elles ne sont pas toutes, également possibles *à priori*.

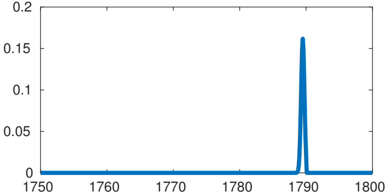
Pour appliquer le principe précédent à un exemple, supposons qu’une urne renferme trois boules dont chacune ne puisse être que

Figure: Bayes’ Theorem in ‘Théorie analytique des probabilités’ by Pierre-Simon Laplace (1812, pp. 182)

prior p, likelihood H, posterior P, integral/sum S.

When did the French revolution start?

(3) Reading in a textbook: It was in the middle of year 1789.



(4) Looking it up on wikipedia.org: The actual date is 14. Juli 1789

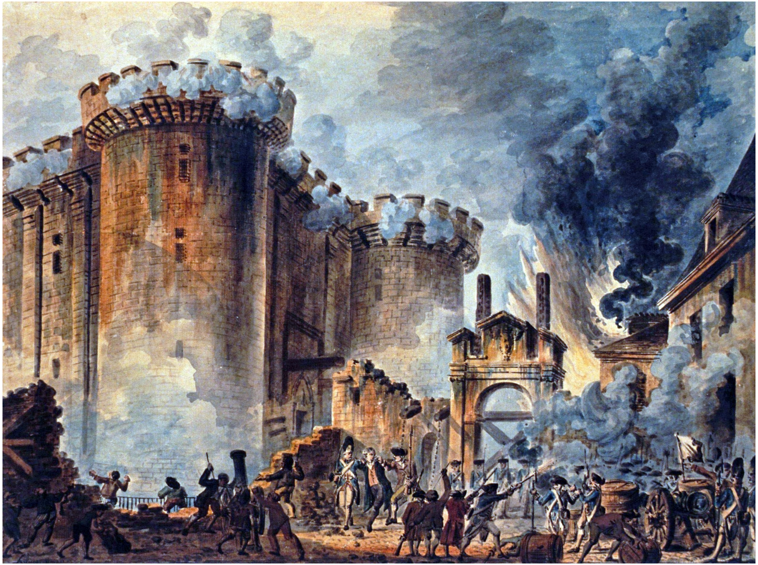


Figure: Prise de la Bastille by Jean-Pierre Louis Laurent Houel, 1789 (Image: Bibliothèque nationale de France)

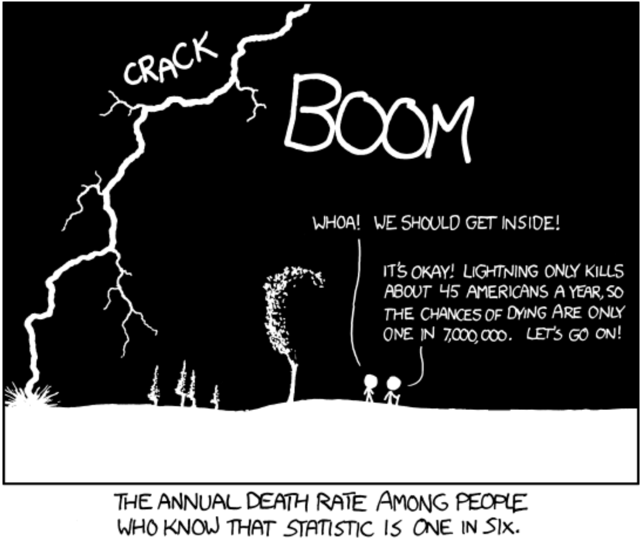


Figure: One more example concerning conditional probabilities (Image: xkcd)

Outline

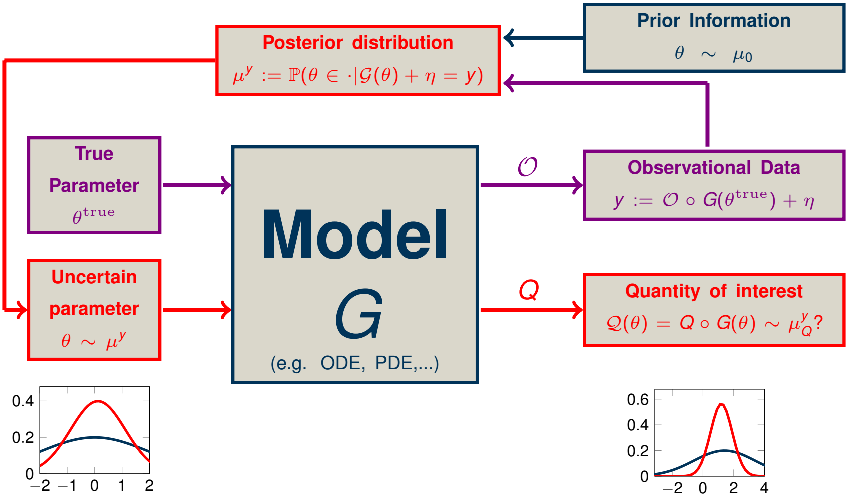
Bayesian Inverse Problem

Given data y and a prior distribution μ_0 - the parameter θ is a random variable: $\theta \sim \mu_0$.
Determine the posterior distribution μ^y, that is

$$\mu^y = \mathbb{P}(\theta \in \cdot | \mathcal{O} \circ G(\theta) + \eta = y)$$

The problem 'find μ^y' is well-posed

Bayesian Inverse Problem



Bayes' Theorem (revisited)

$$\underbrace{\mathrm{cpdf}(\theta|\mathcal{O}\circ G(\theta)+\eta=y)}_{\text{posterior}} = \underbrace{\mathrm{cpdf}(y|\theta)}_{\text{(data) likelihood}}\ \underbrace{\mathrm{mpdf}_1(\theta)}_{\text{prior}}\,/\,\underbrace{\mathrm{mpdf}_2(y)}_{\text{evidence}}$$

prior: Given by the probability measure μ_0

likelihood: $\mathcal{O}\circ G(\theta)-y=\eta\sim\mathrm{N}(0,\Gamma)\Leftrightarrow y\sim\mathrm{N}(\mathcal{O}\circ G(\theta),\Gamma)$

posterior: Given by the probability measure μ^y

evidence: Chosen as a normalising constant

How do we invert Bayesian?

Sampling based: Sample from the posterior measure μ^y

Importance Sampling
Markov Chain Monte Carlo
Sequential Monte Carlo/Particle Filters

Deterministic: Use a deterministic quadrature rule, to approximate μ^y

Sparse Grids
QMC

Idea: Generate samples from μ^y.

Use these samples in a Monte Carlo manner to approximate the distribution of $\mathcal{Q}(\theta)$, where $\theta \sim \mu^y$.

Problem: We typically can't generate iid. samples of μ^y

- weighted samples of the wrong distribution (Importance Sampling, SMC)
- dependent samples of the right distribution (MCMC)

Importance sampling applies directly Bayes' Theorem and uses the following identity:

$$\mathbb{E}_{\mu^y}[Q] = \mathbb{E}_{\mu_0}[Q \cdot \underbrace{\mathrm{cpdf}(y|\cdot)}_{\text{likelihood}}] / \underbrace{\mathbb{E}_{\mu_0}[\underbrace{\mathrm{cpdf}(y|\cdot)}_{\text{likelihood}}]}_{\text{evidence}}$$

Hence, we can integrate w.r.t. to μ^y, using only integrals w.r.t. μ_0.
In practice: Sample iid. from $(\theta_j : j = 1, ..., J) \sim \mu_0$ and approximate:

$$\mathbb{E}_{\mu^y}[Q] \approx J^{-1} \sum_{j=1}^{J} Q(\theta_j)\mathrm{cpdf}(y|\theta_j) / J^{-1} \sum_{j=1}^{J} \mathrm{cpdf}(y|\theta_j)$$

Markov Chain Monte Carlo

Construct an ergodic Markov chain $(\theta_n)_{n\geq 1}$ that is stationary with respect to μ^y.

- $\theta_n \sim \mu^y$ for n large,
- dependent samples can be used for MC type estimation
- some methods
  - Metropolis-Hastings MCMC
  - Gibbs sampling
  - Hamiltonian/Langevin MCMC
  - Slice sampling
  - ...
- often: accept-reject mechanisms

Deterministic Strategies

Several deterministic methods have been proposed

General issue: Estimating the model evidence is difficult
(this also contraindicates importance sampling)

Outline

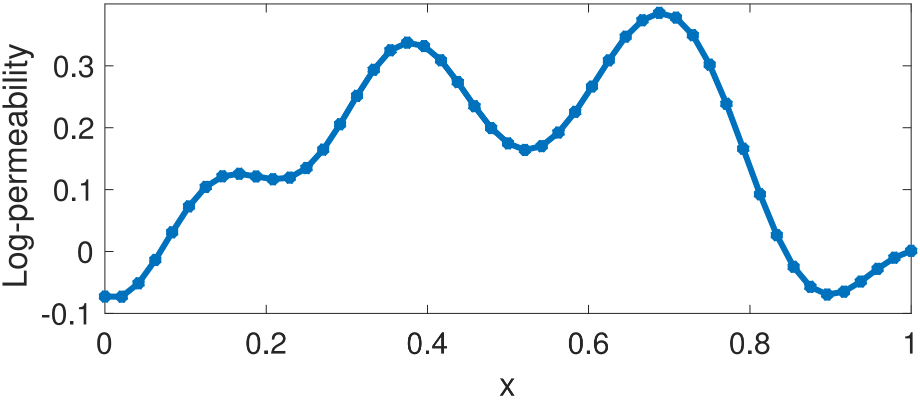
Example 1: 1D Groundwater flow with uncertain source

Consider the following partial differential equation on $D = [0, 1]$

$$
\begin{aligned}
-\nabla(k\nabla)p &= f(\theta) && \text{(on } D\text{)} \\
p &= 0 && \text{(on } \partial D\text{)},
\end{aligned}
$$

where the diffusion coefficient k is known. The source term $f(\theta)$ contains one Gaussian-type source at position $\theta \in [0.1, 0.9]$.
(We solve the PDE using 48 linear Finite Elements.)

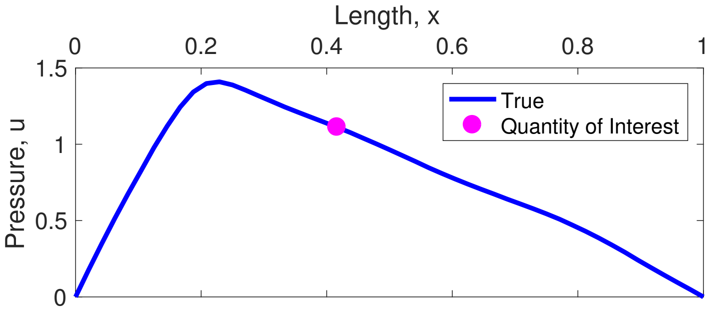
Example 1: (deterministic) log-Permeability



Example 1: Quantity of Interest

Considering the uncertainty in $f(\theta)$, determine the distribution of the Quantity of interest

$$\mathcal{Q} : [0.1, 0.9] \to \mathbb{R}, \theta \mapsto p(5/12).$$

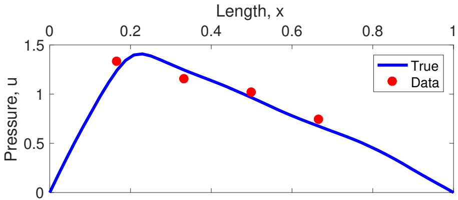


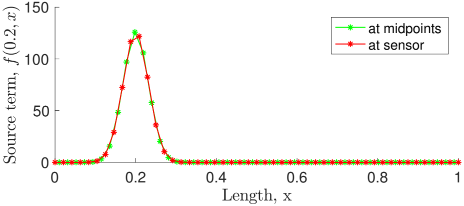
Example 1: Data

The observations are based on the observation operator $\mathcal{O}$, which maps

$$p \mapsto [p(2/12), p(4/12), p(6/12), p(8/12)],$$

given $\theta^{\text{true}} = 0.2$.





Example 1: Bayesian Setting

We assume uncorrelated Gaussian noise, with different variances:

(a) $\Gamma = 0.8^2$
(b) $\Gamma = 0.4^2$
(c) $\Gamma = 0.2^2$
(d) $\Gamma = 0.1^2$

Prior distribution $\theta \sim \mu_0 = \mathrm{Unif}[0.1, 0.9]$

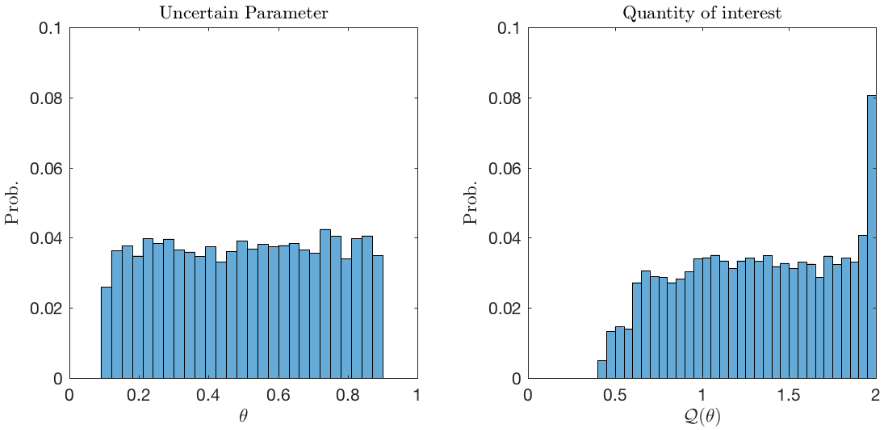
Compare

prior and different posteriors (with different noise levels)

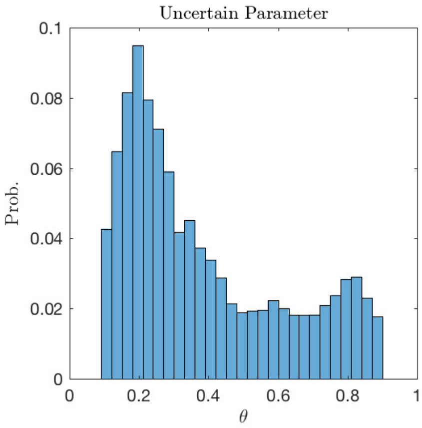
the uncertainty propagation of prior and the posteriors

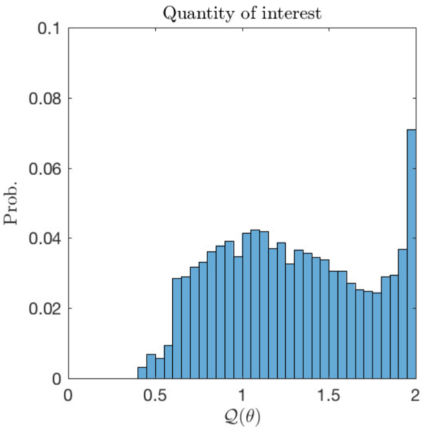
(Estimations with standard Monte Carlo/Importance Sampling using $J = 10000$.)

Example 1: No data (i.e. prior)

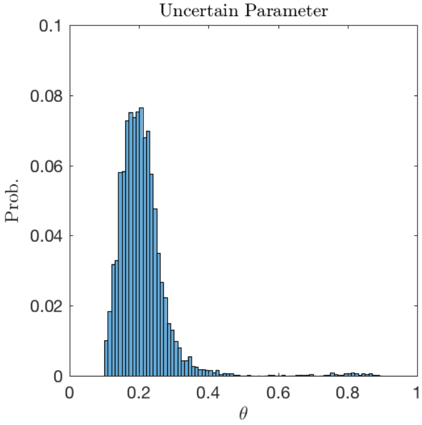


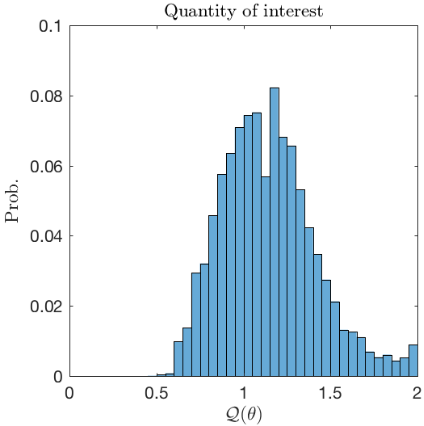
Example 1: Very high noise level $\Gamma = 0.8^2$



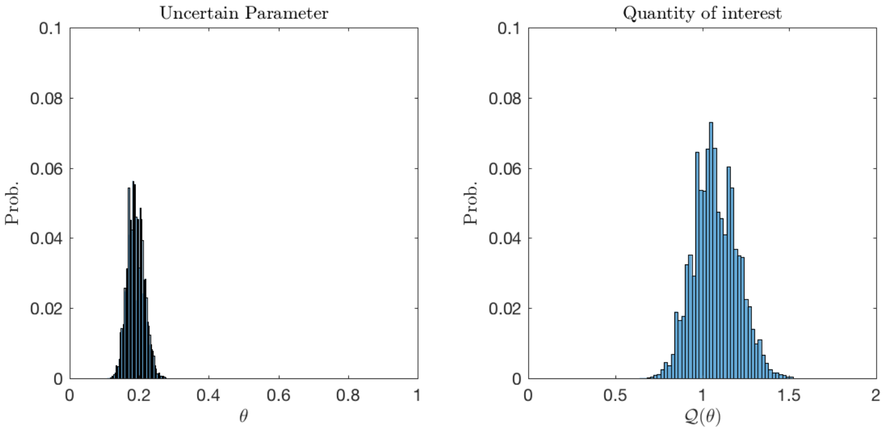


Example 1: High noise level $\Gamma = 0.4^2$

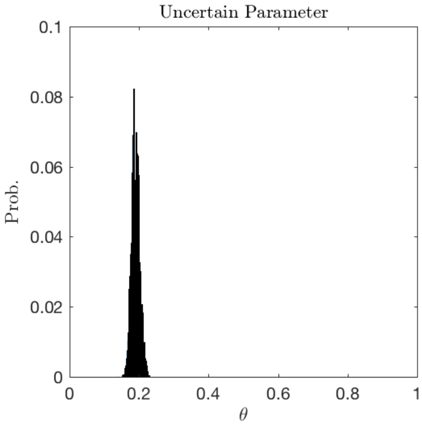


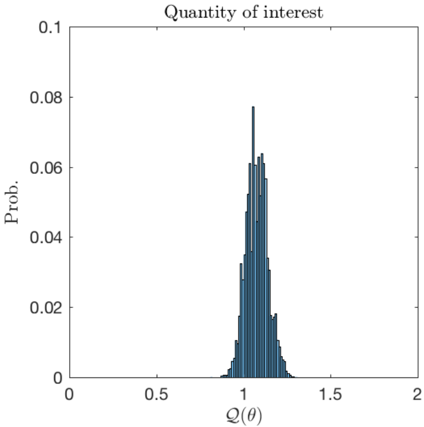


Example 1: Small noise level $\Gamma = 0.2^2$



Example 1: Very small noise level $\Gamma = 0.1^2$





Example 1: Summary

Smaller noise level $\Leftrightarrow$ less uncertainty in the parameter $\Leftrightarrow$ less uncertainty[2] in the quantity of interest

The unknown parameter can be estimated pretty well in this setting

Importance Sampling can be used in such simple settings.

[2] less uncertainty meaning ‘smaller variance’.

Example 2: 1D Groundw. flow with uncertain number and pos. of sources

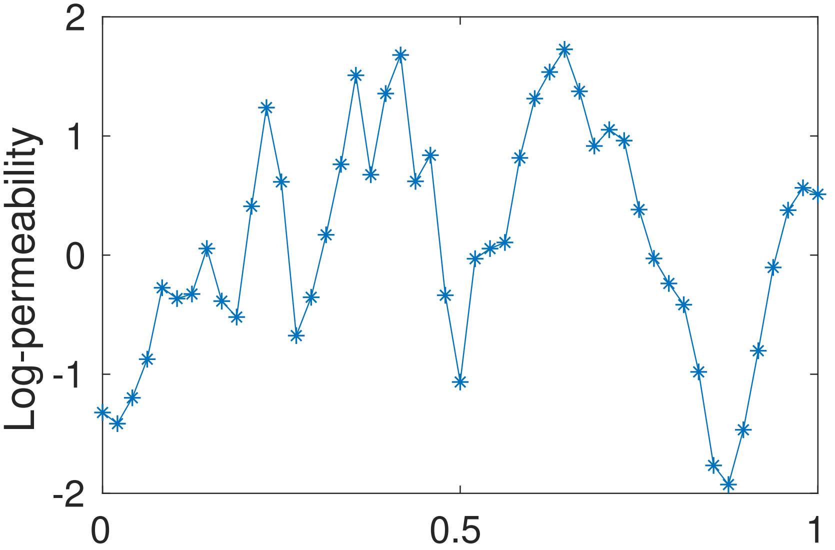
Consider again

$$
\begin{aligned}
-\nabla(k\nabla)p &= g(\theta) && \text{(on } D\text{)} \\
p &= 0 && \text{(on } \partial D\text{)},
\end{aligned}
$$

$\theta := (N, \xi_1, ..., \xi_N)$, where N is the number of Gaussian type sources and $\xi_1, ..., \xi_N$ are the positions of the sources (sorted ascendingly)

the log-permeability is known, but with a higher spatial variability

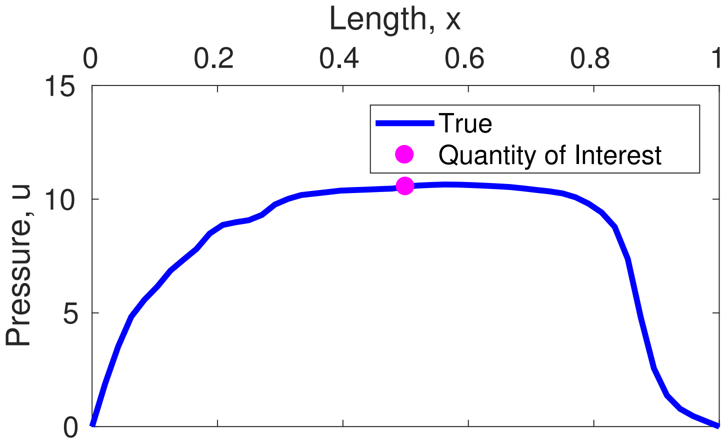
Example 2: (deterministic) log-Permeability



Example 2: Quantity of Interest

Considering the uncertainty in $g(\theta)$, determine the distribution of the Quantity of interest

$$\mathcal{Q} : [0.1, 0.9] \to \mathbb{R}, \theta \mapsto p(1/2).$$

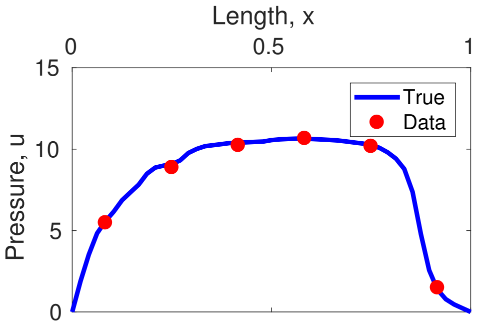


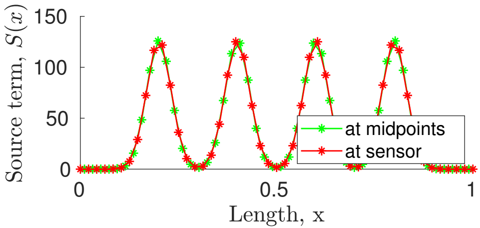
Example 2: Data

The observations are based on the observation operator $\mathcal{O}$, which maps

$$p \mapsto [p(1/12), p(3/12), p(5/12), p(7/12), p(9/12), p(11/12)],$$

given $\theta^{\text{true}} := (4, 0.2, 0.4, 0.6, 0.8)$.





Example 2: Bayesian Setting

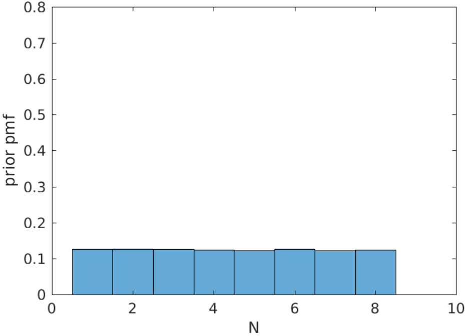
We assume uncorrelated Gaussian noise with variance $\Gamma = 0.4^2$

Prior distribution $\theta \sim \mu_0$. μ_0 is given by the following sampling procedure:

1. Sample $N \sim \mathrm{Unif}\{1, ..., 8\}$
2. Sample $\xi \sim \mathrm{Unif}[0.1, 0.9]^N$
3. Set $\xi := \mathrm{sort}(\xi)$
4. Set $\theta := (N, \xi_1, ..., \xi_N)$

Compare prior and posterior and their uncertainty propagation (Estimations with standard Monte Carlo/Importance Sampling using $J = 10000$.)

Example 2: Prior



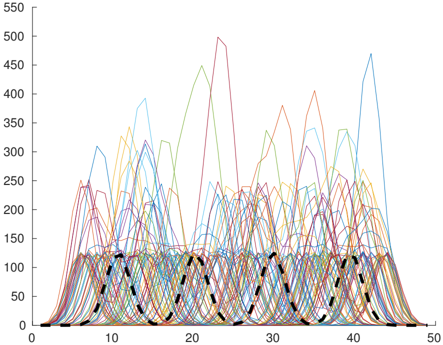


Figure: Prior distribution of N (left) and 100 samples of the prior distribution of the Source terms

Example 2: Posterior

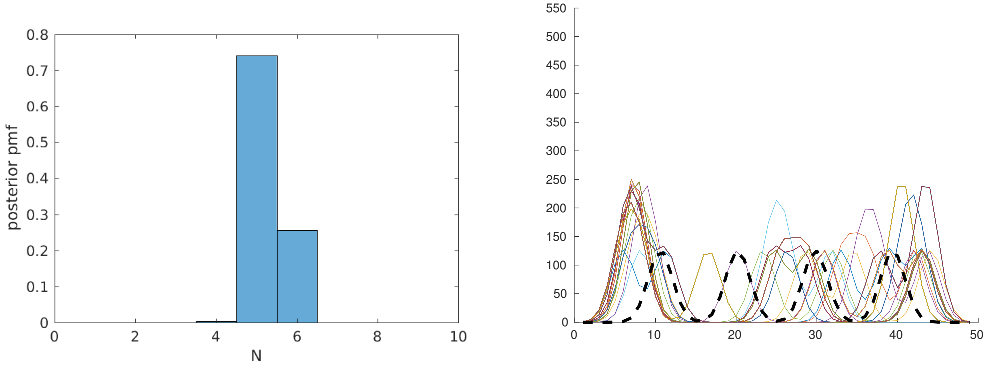


Figure: Posterior distribution of N (left) and 100 samples of the posterior distribution of the Source terms

Example 2: Quantity of Interest

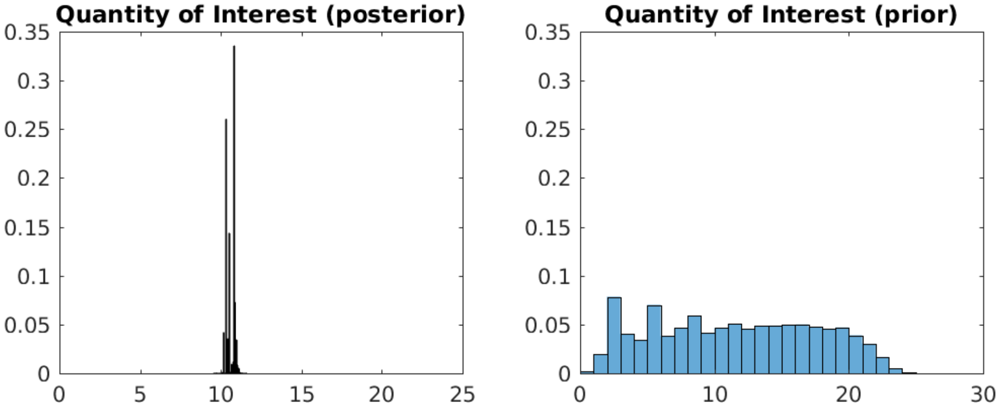


Figure: Quantities of interest, where the source term is distributed according to the prior (left) and posterior (right)

Example 2: Summary

Bayesian estimation is possible in ‘complicated settings’ (such as this transdimensional setting)

Importance Sampling is not very efficient

Outline

Messages to take home

- + Bayesian Statistics can be used to incorporate data into an uncertain model
- + Bayesian Inverse Problems are well-posed and thus a consistent approach to parameter estimation
- + Applying the Bayesian Framework is possible in many different settings, also in ones that are genuinely difficult (e.g. transdimensional parameter spaces)
- - Solving Bayesian Inverse Problems is computationally very expensive

    requires many forward solves
    algorithmically complex

How to learn Bayesian

- Various lectures at TUM:
  - Bayesian strategies for inverse problems, Prof. Koutsourelakis (Mechanical Engineering)
  - Various Machine Learning lectures in CS
- Speak with Prof. Dr. Elisabeth Ullmann or Jonas Latz (both M2)
- GitHub/latz-io
  - A short review on algorithms for Bayesian Inverse Problems
  - Sample Code (MATLAB)
  - These slides

How to learn Bayesian

Various Books/Papers

- Moritz Allmaras et al.- Estimating Parameters in Physical Models through Bayesian Inversion: A Complete Example (2013; SIAM Rev. 55(1))
- Jun Liu - Monte Carlo Strategies in Scientific Computing (2004; Springer)
- Sharon Bertsch McGrayne - The Theory that would not die (2011, Yale University Press)
- Christian Robert - The Bayesian Choice (2007, Springer)
- Andrew Stuart - Inverse Problems: A Bayesian Perspective (2010; in Acta Numerica 19)

Jonas Latz

`Input/Output:  www.latz.io`