Introduction to Bayesian inference

+Incorporating a Modeling Workflow

Matthew Heffernan (McGill University)
4th JETSCAPE School

July 27, 2021

Based on recent work in: 2104.08621 [arXiv:physics.ed-ph]

McGill

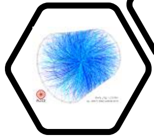
JETSCAPE

Plan for this session

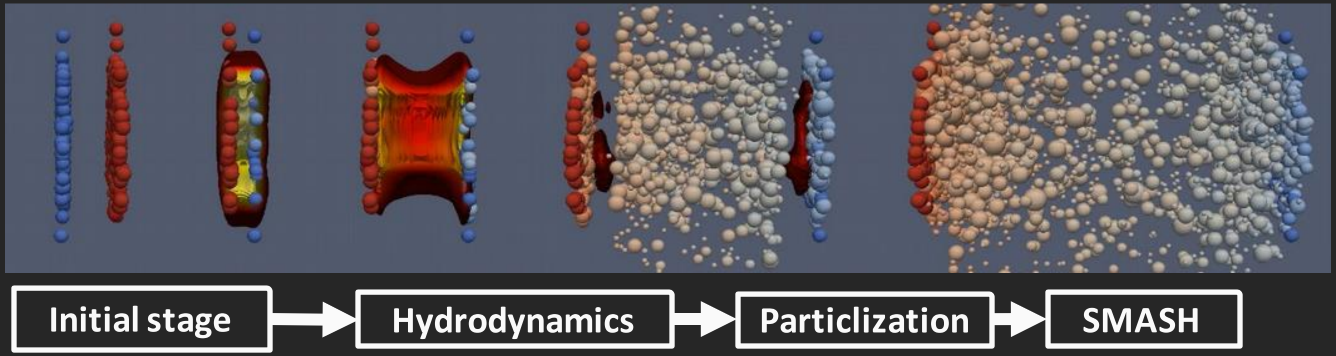
Today:

- 1-hour general introduction to Bayesian Inference and workflow
- 2-hour hands-on session (Jupyter Notebook)
- Ask questions in #jul27-bayes-overview

Big Picture

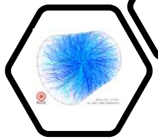


- We have a model of some physical process, say a relativistic heavy ion collision
- We have experimental measurements of what we believe to be the same process

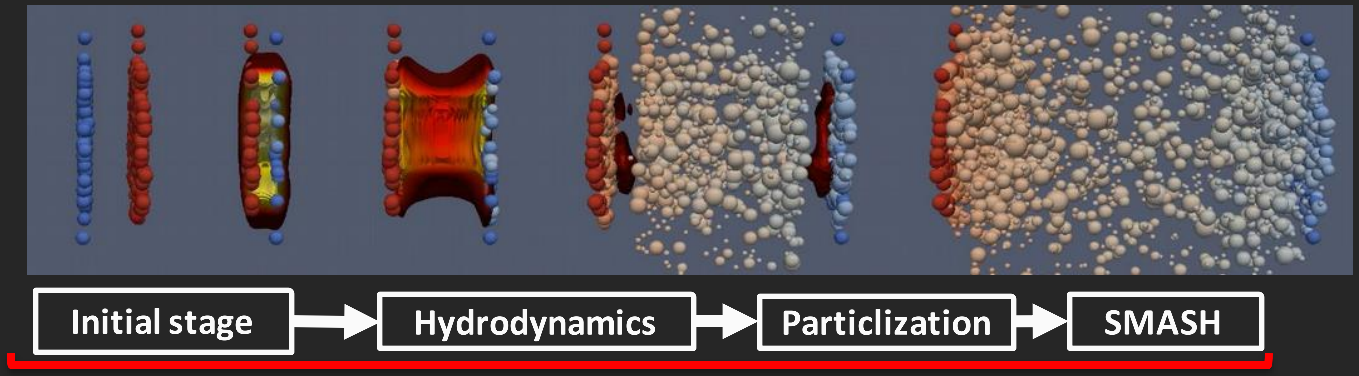


How can we learn about the physics by using a model?

Big Picture



- We have a model of some physical process, say a relativistic heavy ion collision
- We have experimental measurements of what we believe to be the same process

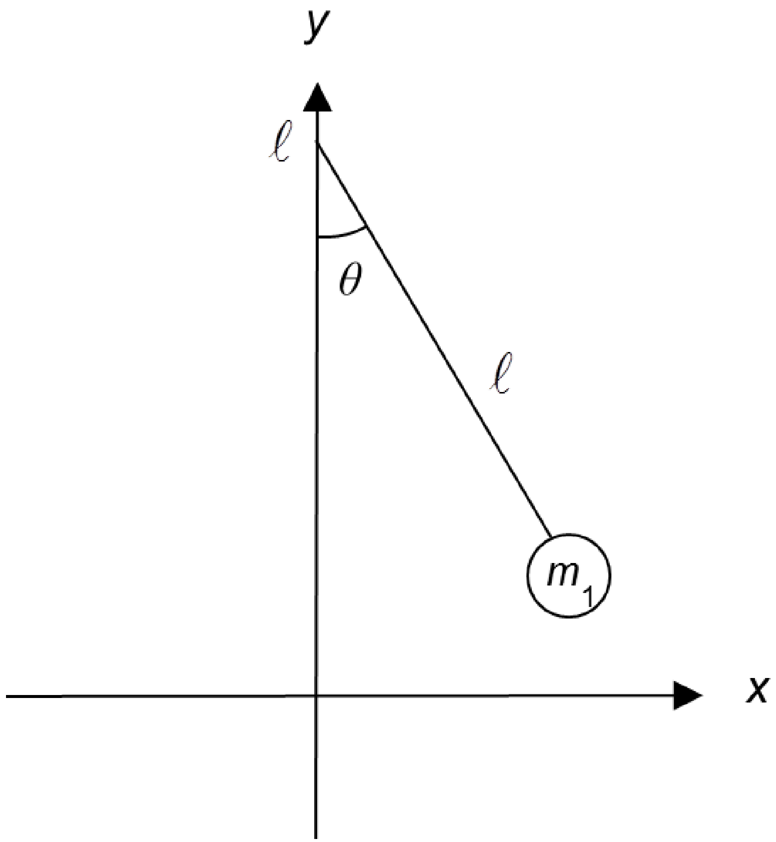


Model parameters

Constrain model parameters via measurements

Big picture: Hands on session

- Heavy ion collisions: computationally expensive and theoretically complex
- In this session: a 1 parameter, 1 observable problem
- Familiar physics: The simple pendulum
    - Can we infer gravitational acceleration given the period?
- Use this example to demonstrate how to approach Bayesian modeling
    - Ensure robust, trustworthy analysis



Questions I hope to answer

- Haven't we all been comparing models with data all our lives? Why do I suddenly need Bayesian methods?
- How do I input my theoretical knowledge into a Bayesian study?
- What is a workflow and why do I need it?
- How does this work in practice?
- Where can I find more details?

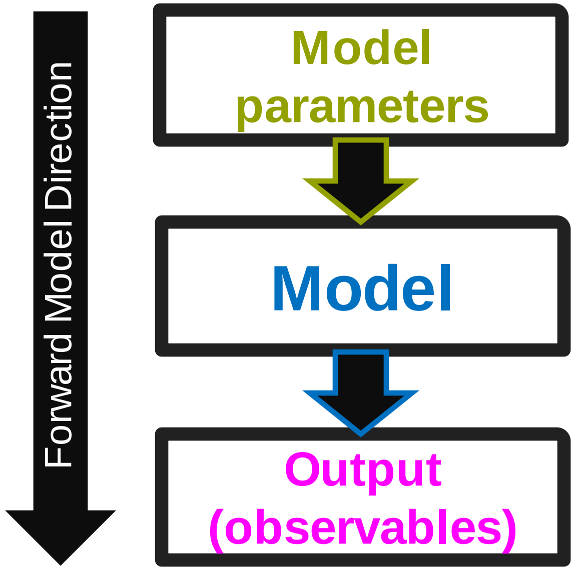
What is Bayesian inference?

The Forward Problem

Modelling has **three main ingredients**:

1. A **model**: Theoretical description of the relevant process(es)
2. Model **parameter(s)**: Quantities poorly constrained by theory, but contain information and can be used to fit model predictions to measurements
3. Model **output(s)** or **observable(s)**: Prediction(s) for the result of a process that can be compared to measurements

Generally well-defined: Given a set of model parameters, what are the model outputs?



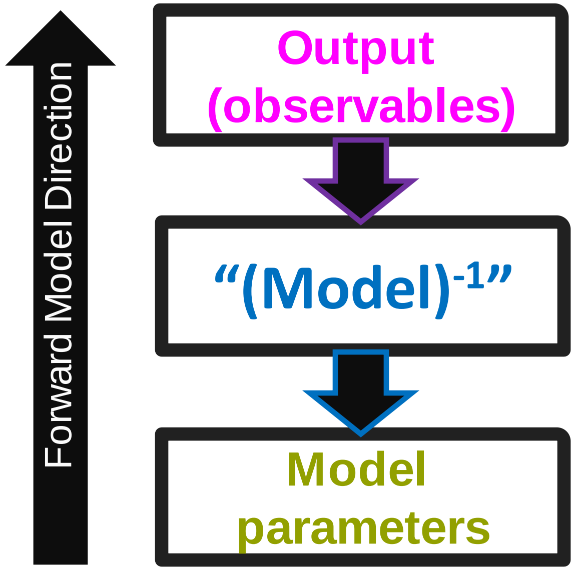
The Inverse Problem

Modelling has **three main ingredients**:

1. A **model**: Theoretical description of the relevant process(es)
2. Model **parameter(s)**: Quantities poorly constrained by theory, but contain information and can be used to fit model predictions to measurements
3. Model **output(s)** or **observable(s)**: Prediction(s) for the result of a process that can be compared to measurements

Generally ill-defined:

What are the model parameters that result in given set of model outputs?



Matthew Heffernan | McGill University

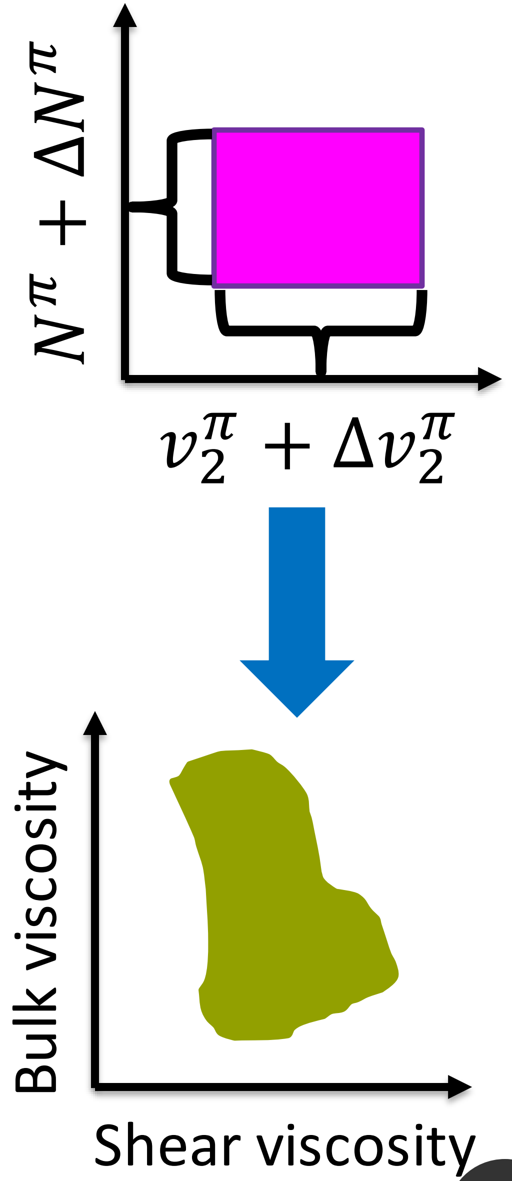
Inverse Problem:

Mapping Observables to Parameters

Generally ill-defined:

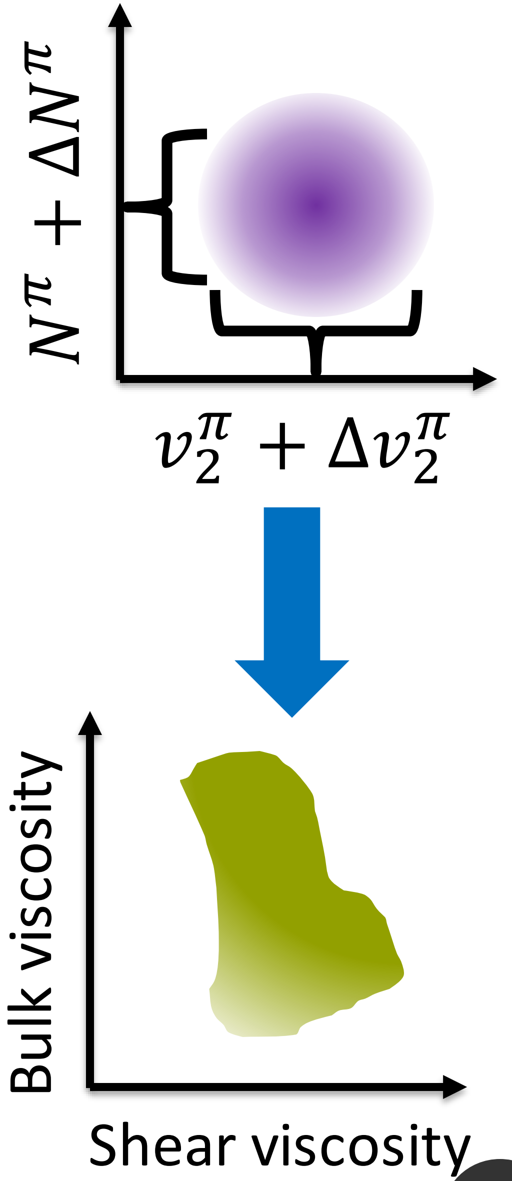
What are the model parameters that result in given set of model outputs?

Observables (output) → **$(\text{Model})^{-1}$** → **Model parameters**



Inverse Problem:

Mapping an observed probability distribution to a distribution of parameters



Observables are not known exactly: probability distributions

Can use Bayes’ Theorem to *infer* this mapping

Bayes' Theorem is a way to solve probabilistic inverse problems

Introduction to Bayesian inference

Bayes' Theorem:
Connecting the Forward and Inverse Problems

$$p(H|d,I) = \frac{p(d|H,I)p(H,I)}{p(d,I)}$$

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

Components of Bayes Theorem

- Posterior (Inverse): Probability of the hypothesis "H" **posterior to comparison** with data "d"
- Likelihood (Forward): Probability of the data "d" given the hypothesis "H"
- Prior: Probability of H **prior to comparison** with the data "d", only informed by other expectations we have, e.g. theoretical constraints on the quantity "H"
- Bayes' Evidence: Probability of data given the model. Often treated as a normalization constant, is key used in model selection and averaging.

- Notation:
  - p(A): Probability of, or degree of belief in, A
  - A and B: Propositions or statements
  - |: Conditionality, i.e. A|B,C means "A given (B and C)"
  - I: Other information, can include theoretical expectations such as the model

Bayes' Theorem:
Connecting the Forward and Inverse Problems

$$p(H|d,I) = \frac{p(d|H,I)p(H,I)}{p(d,I)}$$

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

Components of Bayes Theorem:
Posterior (Inverse Problem)

- Posterior (Inverse): Probability of the hypothesis "H" **posterior to comparison** with data "d"
- Hypothesis: a particular set of parameters
- Represents what we know about the hypotheses after (*posterior* to) comparison of the model outputs to data

- Notation:
  - p(A): Probability of, or degree of belief in, A
  - A and B: Propositions or statements
  - |: Conditionality, i.e. A|B,C means "A given (B and C)"
  - I: Other information, can include theoretical expectations such as the model

Bayes' Theorem:
Connecting the Forward and Inverse Problems

$$p(H|d,I) = \frac{p(d|H,I)\,p(H,I)}{p(d,I)}$$

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

- Notation:
  - p(A): Probability of, or degree of belief in, A
  - A and B: Propositions or statements
  - |: Conditionality, i.e. A|B,C means “A given (B and C)”
  - I: Other information, can include theoretical expectations such as the model

Components of Bayes Theorem:

Likelihood (Forward Problem)

- Likelihood: Probability of the data “d” given the hypothesis “H”
- Calculates how likely the data is based on the model prediction, model uncertainty, and data uncertainty.
- Gaussian likelihood assumes the errors are normally distributed in both model and data (may not be the case).

Bayes' Theorem:
Connecting the Forward and Inverse Problems

$$p(H|d,I) = \frac{p(d|H,I)\,p(H,I)}{p(d,I)}$$

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

Components of Bayes Theorem:

Likelihood (Forward Problem)

- Likelihood: Probability of the data "d" given the hypothesis "H"
- Calculates how likely the data is based on the model prediction, model uncertainty, and data uncertainty.
- Y_{model} is the model prediction of the observables

$$p(d|H,I) = \frac{1}{\sqrt{2\pi}\sigma}\prod_{i}^{M}\exp\left(-\frac{(d_i - Y_{\text{model}})^2}{2\sigma^2}\right)$$

$$\ln p(d|H,I) = -\frac{1}{2}\sum_{i=1}^{M}\left[\ln(2\pi\sigma^2) + \frac{(d_i - Y_{\text{Model}})^2}{\sigma^2}\right]$$

$$\sigma^2 = \sigma^2_{model} + \sigma^2_{data}$$

Bayes' Theorem:
Connecting the Forward and Inverse Problems

$$p(H|d,I) = \frac{p(d|H,I)\,p(H,I)}{p(d,I)}$$

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

Components of Bayes Theorem:
Likelihood (Forward Problem)

- Likelihood: Probability of the data "d" given the hypothesis "H"
- Generalizing to higher dimensions is straightforward
- Because this involves model calculations, the likelihood is the most computationally expensive component to calculate: a variety of methods exist to help

$$\ln p(d|H,I) = -\frac{1}{2}\sum_{i=1}^{M}\left[\ln(2\pi\sigma^2) + \frac{(d_i - Y_{\text{Model}})^2}{\sigma^2}\right]$$

$$\ln\left[p\left(d \mid H, I\right)\right] = -\frac{1}{2}\ln\left[(2\pi)^n \det\Sigma\right] - \frac{1}{2}\Delta\mathbf{y}^T\Sigma^{-1}\Delta\mathbf{y}$$

Length of the data: n

Difference between data and prediction: $\Delta\mathbf{y} = \mathbf{d} - \mathbf{Y}$

Covariance Matrix: Σ

Bayes' Theorem:
Connecting the Forward and Inverse Problems

$$p(H|d,I) = \frac{p(d|H,I)\,p(H,I)}{p(d,I)}$$

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

Components of Bayes Theorem:

Prior

- Prior: Probability of H **prior to comparison** with the data "d", only informed by other expectations
- This represents what we know before (*prior* to) comparison with data
- Examples: A parameter must be positive definite; is unlikely to be outside of a certain range

- Notation:
  - p(A): Probability of, or degree of belief in, A
  - A and B: Propositions or statements
  - |: Conditionality, i.e. A|B,C means "A given (B and C)"
  - I: Other information, can include theoretical expectations such as the model

Bayes' Theorem:
Connecting the Forward and Inverse Problems

$$p(H|d,I) = \frac{p(d|H,I)p(H,I)}{p(d,I)}$$

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

- Notation:
  - p(A): Probability of, or degree of belief in, A
  - A and B: Propositions or statements
  - |: Conditionality, i.e. A|B,C means “A given (B and C)”
  - I: Other information, can include theoretical expectations such as the model

Components of Bayes Theorem:
Bayes’ Evidence

- Bayes’ Evidence: Probability of data given the model. Often treated as a normalization constant, key to model selection and averaging.
- Large normalization constant is needed when the likelihood is large, small normalization constant is needed when the likelihood is small

Bayes' Theorem:
A Simple Example

A Cloudy Day

- In JETSCAPEville, 40% of all rainy days have cloudy mornings.
- 30% of all mornings in JETSCAPEville are cloudy and in July, it has typically rained in JETSCAPEville on 5 out of 31 days (~16.13%).
- Today in July in JETSCAPEville, the morning was cloudy. Based on this information, what is the probability of rain today?

$$p(\text{rain}|\text{morning clouds})$$

$$p(A|B) = \frac{p(B|A)p(A)}{p(B)}$$

$$p(\text{rain}|\text{morning clouds}) = \frac{p(\text{morning clouds}|\text{rain})p(\text{rain})}{p(\text{morning clouds})}$$

$$p(\text{morning clouds}|\text{rain}) = 0.4$$

$$p(\text{rain}) = 5/31$$

$$p(\text{morning clouds}) = 0.3$$

$$p(\text{rain}|\text{morning clouds}) = \frac{0.4 \times 0.1613}{0.3}$$

$$p(\text{rain}|\text{morning clouds}) = 0.2151$$

Questions?

When is it reasonable to use Bayesian inference?

Ideally suited to the problem when:

- An **accurate quantification of uncertainty** is critical to the result
- **Theoretical expectations** add constraint
- Models with **many parameters** are constrained with **many measurements**
- **Comparing** complex models
- **Making broad statements** about models that aren't justified by first-principles theory, e.g. "No model of type X can reproduce the data"
- Advanced: No single model is best suited to data (Bayesian model mixing and averaging)

Why is this better than more familiar Frequentist methods?

Bayesian techniques:

- Are more intuitively interpreted - it's how most scientists *already interpet plots*
- Are simpler and easier to teach, turn around, and immediately apply in research
- Do not rely on complicated formula
- Make assumptions explicit and clear without hiding them in mathematical abstraction

What Bayesian Inference is and what it isn't

Avoid sacrificing accuracy for precision

Bayesian inference is:

- A method to systematically compare a model to data
- A method to quantify uncertainty
- Reliable **with** rigorous validation
- Trustworthy **with reporting** diagnostics

Bayesian inference isn't:

- A simple best-fit answer
- A way to drive uncertainties to 0
- Reliable **without** rigorous validation
- Trustworthy **without reporting** diagnostics

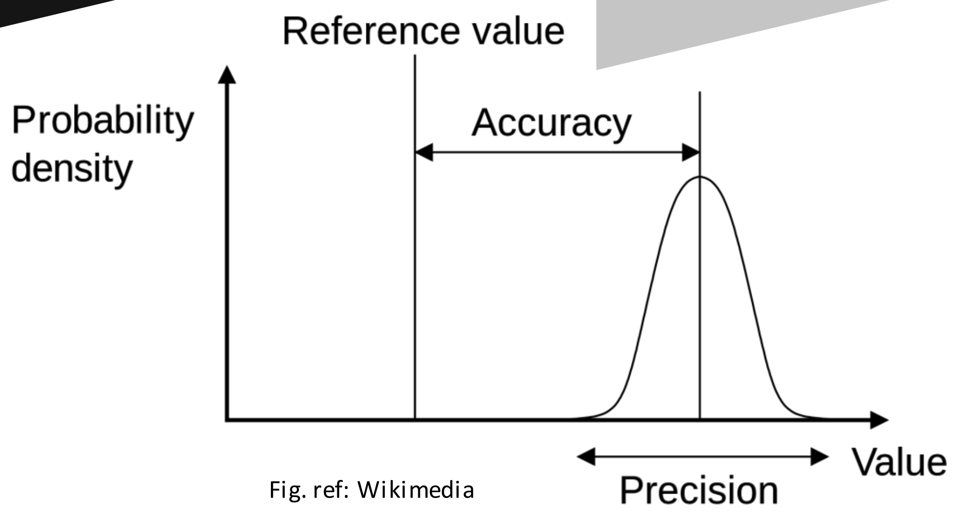


Fig. ref: Wikimedia

Measurements and predictions as probability distributions

Crude simplifying assumption:
all uncertainties are normally distributed

Data are often averages over ensembles:

- mean value
- statistical uncertainty on this mean;
- additional systematic uncertainties (not normally distributed)

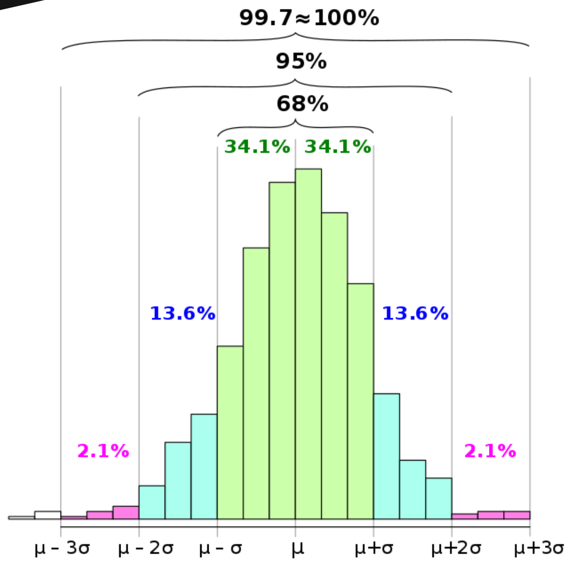


Fig. ref: Wikimedia

Models have uncertainty:

- Statistical: e.g. averaging over collisions, finite number of particles
- Numerical: e.g. interpolation uncertainty
- Systematic: e.g. Approximations

Markov Chain Monte Carlo:
A forgetful walk through parameter space

Models take in specific parameter values, not distributions: how do we connect these?

MCMC

- "Walkers" step through the parameter space
- Probability of accepting the next step:
  p(next location) / p(current location)
- After enough steps, walkers forget where they started
- Distribution of steps corresponds to samples drawn from distribution of interest ("target distribution")
- Distribution of samples can be used to estimate properties of the underlying target distribution

2. Forget starting location

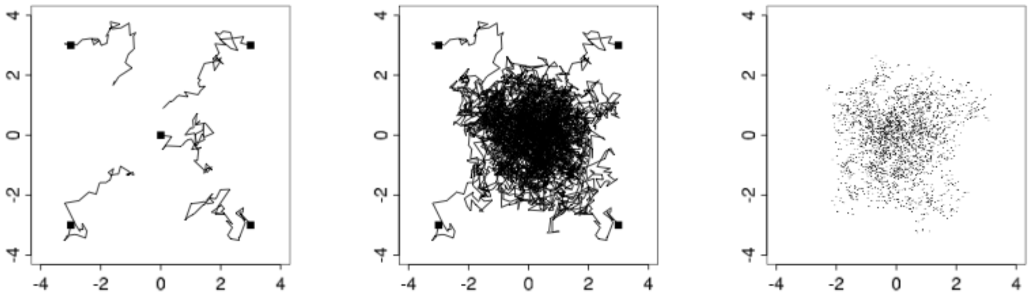


Fig. ref: Gelman, Bayesian Data Analyis

1. Start off and explore

3. Walk according to target

For a more detailed introduction: https://github.com/mrhheffernan/bayes-tutorial/blob/master/mcmc/MCMC-intro-and-diagnostics.ipynb

Questions?

Introduction to Bayesian Modeling Workflow

and why you need one

What is a Bayesian Modeling Workflow?

Workflow: A repeatable pattern of steps to complete a task

In complex modeling environments, a workflow ensures a rigorous, repeatable set of steps has been taken to ensure reliable development and analysis

A workflow helps bullet-proofs the analysis
Goal: Trustworthy Inference

Bayesian Modeling Workflow

Step 1: Choose initial model
Goal: Explicitly define the model and prior state of knowledge
Step 2: Prior Predictive checks
Goal: Critically evaluate if the model+prior are consistent with domain knowledge
Step 3: Model validation via fake data simulation
Goal 1: Critically evaluate if the model+prior can recover known inputs
Goal 2: Inspect tools to ensure output is reliable
Step 4: Fitting the Model - Inference with Data
Goal: Extract state of knowledge *posterior* to comparison with data
Safety Check: Inspect tools using available diagnostics
Step 5: Posterior predictive check
Goal: Evaluate if model+posterior is consistent with data and domain knowledge

Bayesian Workflow
Step 1 | Choosing an initial model

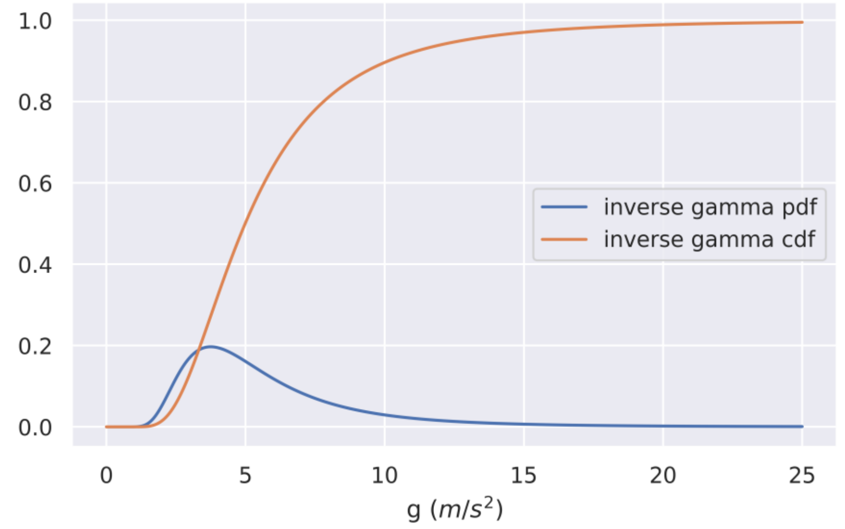
Step 1

Pendulum hands-on preview:
- Model: Exact formula for the period
- Prior: g on Eath is 99% likely to be between g on the Moon and g on Jupiter

$$T = 4\sqrt{\frac{L}{g}}\int_0^{\pi/2} \frac{1}{\sqrt{1-\sin^2(\theta_0/2)\sin^2(\phi/2)}}d\phi$$

Choose an initial model

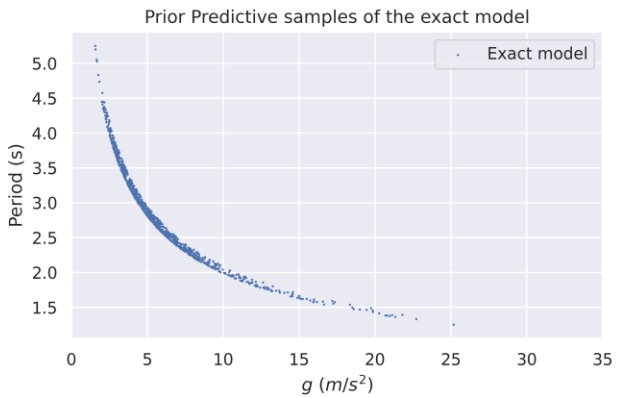
- Either develop a model or motivate the choice of an existing model
- 1.1. Define principled priors motivated by knowledge of the model and the system
  - Justify specific choices and why they are reasonable

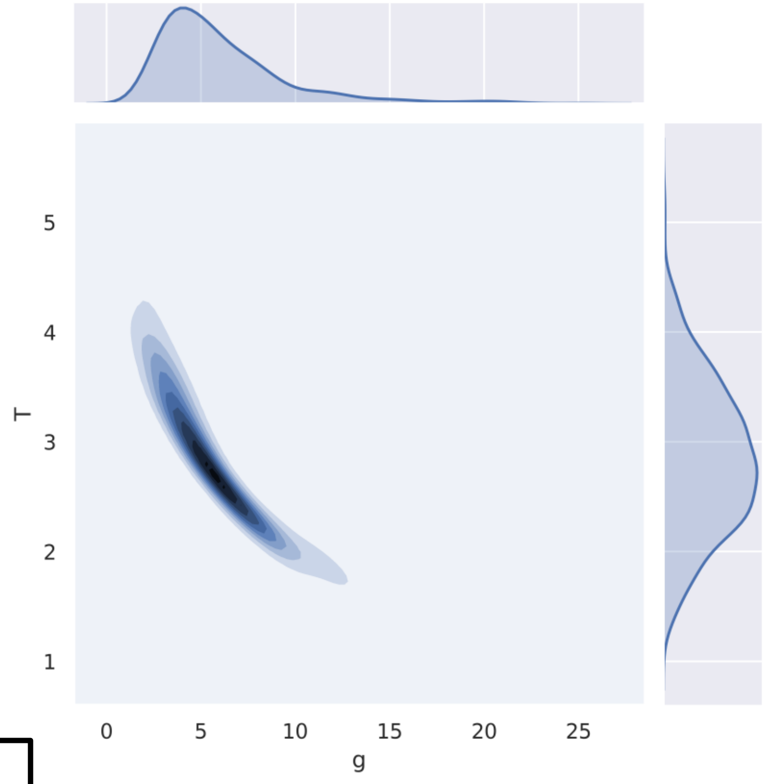


Bayesian Workflow
Step 2 | Prior Predictive Checks

Step 2

Prior Predictive checks
- Are predictions made by the model+prior consistent with domain knowledge?
  - If yes: Provisionally accept model+prior.
  - If no: Identify specific contradictions and rectify.





Pendulum hands-on preview:
- Good coverage of a reasonable range for periods and g

Bayesian Workflow
Step 3 | Model Validation

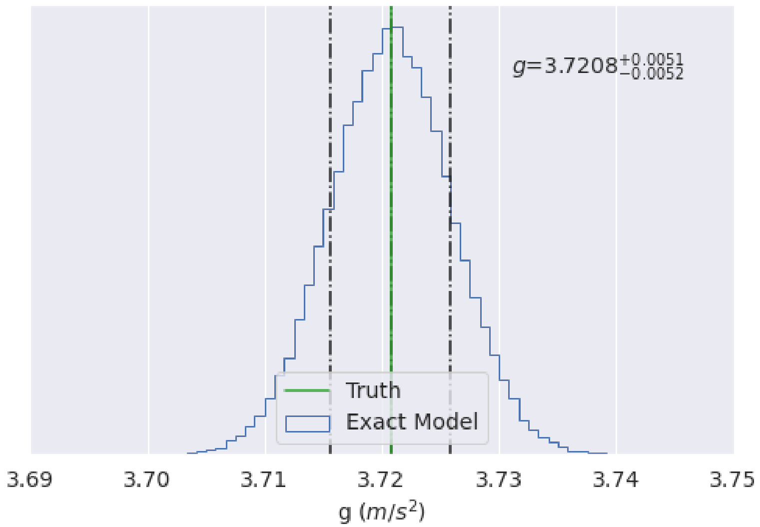
Step 3

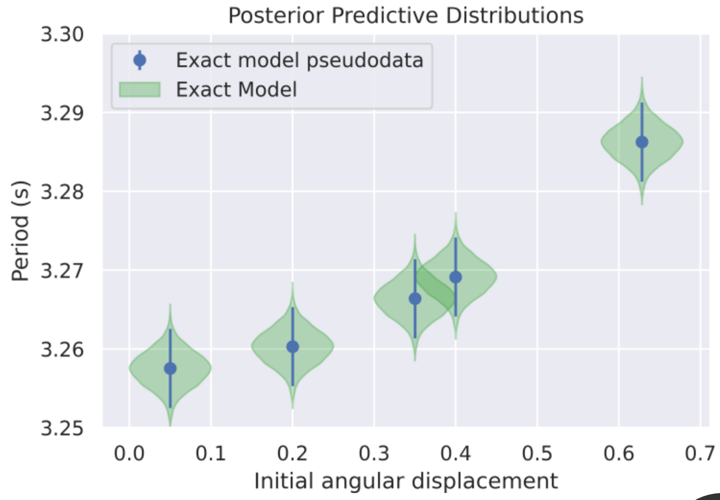
Model validation via fake data simulation
(a.k.a. Closure tests, Empirical coverage tests)

- 3.1. Can the model and prior recover a known value in idealized data using the available tools? Do diagnostics indicate that the numerics are reliable?
  - If yes: Provisionally accept computation and proceed.
  - If no: Identify specific issues with input recovery and rectify by modifying priors or model. If modifying priors or the model, return to step 1.

- 3.2. Posterior predictive check: Does the fake data posterior produce predictions consistent with the fake data?
  - If yes: Provisionally accept computation and proceed.
  - If no: Identify source of discrepancy and rectify.

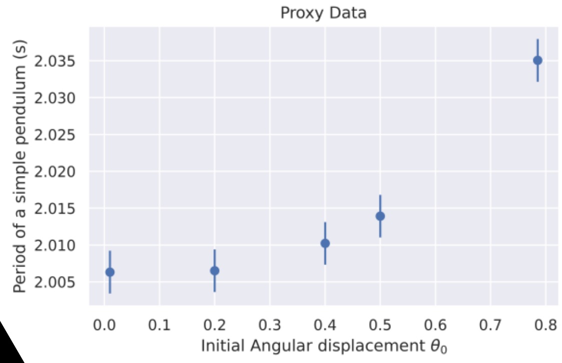
Pendulum hands-on preview:
- Fake data generated with g on Mars
- Recovers the result!

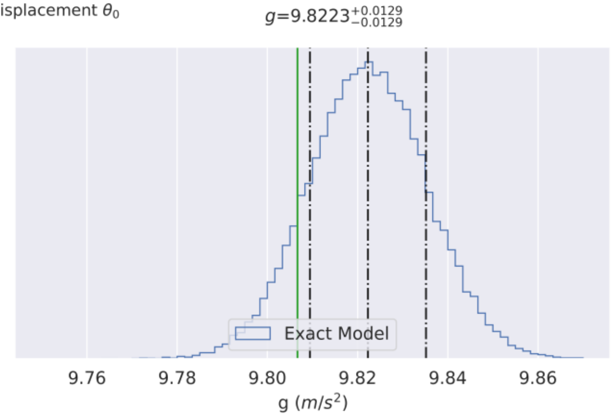




Bayesian Workflow
Step 4 | Inference with Data

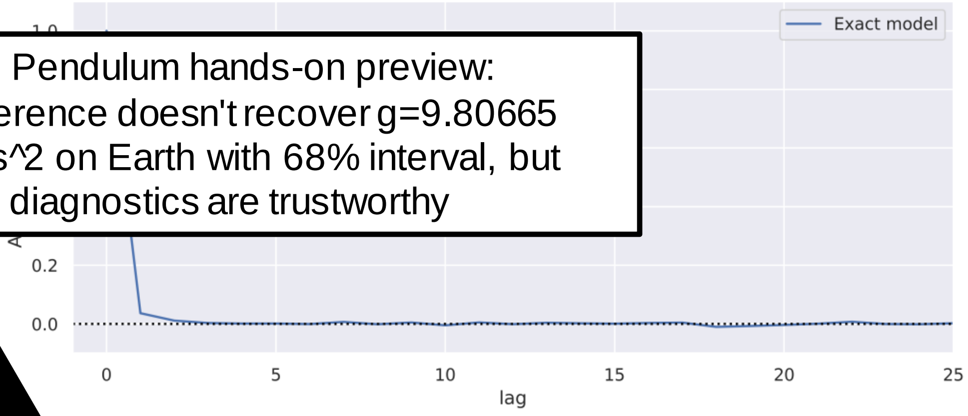
Step 4





Fitting the Model - Inference with Data

- Do the model+prior clearly recover a value consistent with domain knowledge?
- Do diagnostics indicate that the numerics have converged?
  - If yes: Provisionally accept computation and proceed.
  - If no: Modify priors, model, investigate deficiencies in numerical tools



Pendulum hands-on preview:

- Inference doesn't recover g=9.80665 m/s^2 on Earth with 68% interval, but the diagnostics are trustworthy

Bayesian Workflow

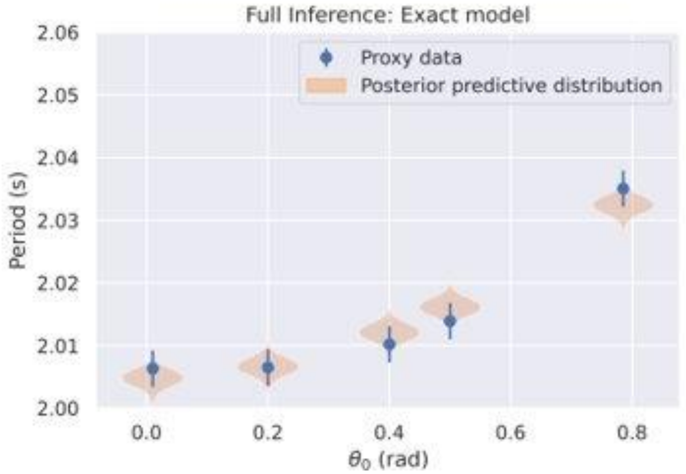
Step 5 | Posterior Predictive Checks

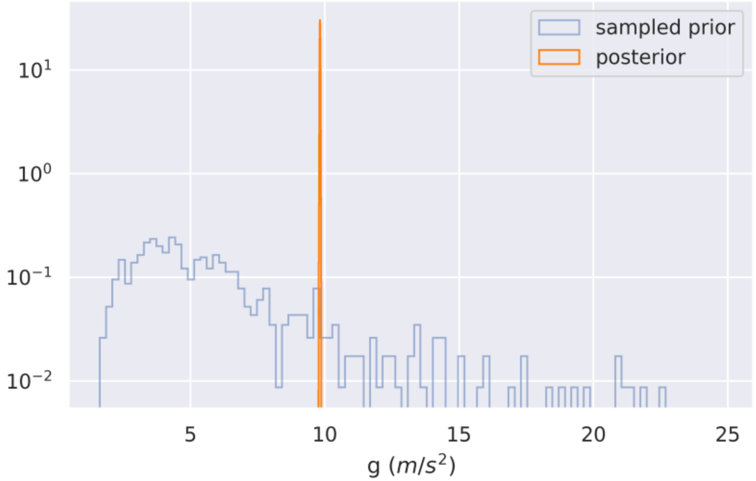
Step 5

- Does the posterior produce predictions consistent with the data? Are any deficiencies interpretable or expected given the model?
  - If yes: Accept computation. If within scope, improve model and repeat.
  - If no: Identify source of discrepancy and rectify. Investigate why this was not caught in model validation.

Pendulum hands-on preview:

- Posterior predictive explains data well
- Most of the constraint comes from the likelihood





Questions?

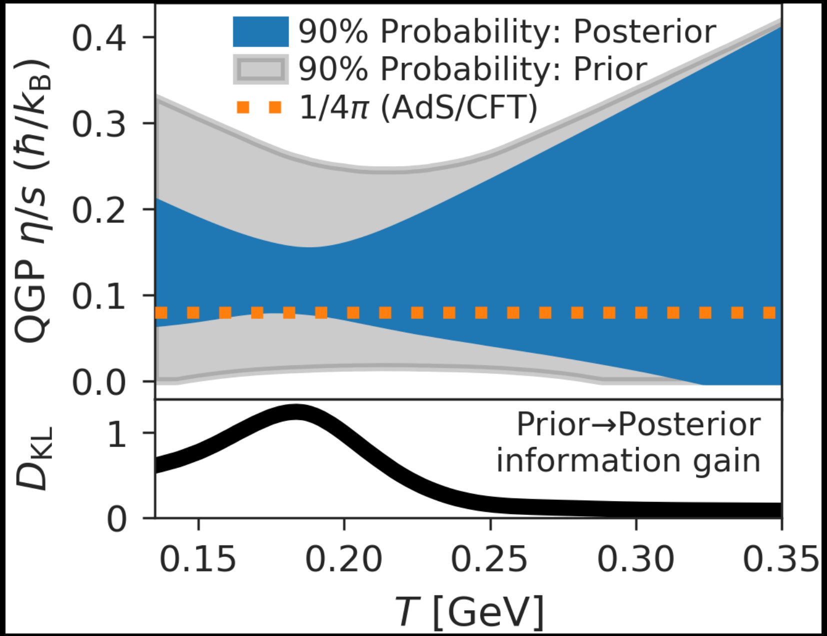
Note on Priors

Some priors are commonly called "non-informative" or "weakly-informative"

- These statements are **not general** and **can only be understood in comparison with the likelihood**

- A prior is weakly-informative when most of the constraint is from the likelihood. If the likelihood gives little constraint, even a maximum entropy prior can be highly informative.

Note on Priors



- What about uniform (flat) priors? Aren't these the most agnostic?
  - Can be parametrization-dependent: May not be weakly informative when transformed
  - Should be used when their features fit the application:
    - **Often not a faithful description of prior knowledge**
    - **Sharp cutoffs can be unrealistic**
    - Maximum Entropy motivations should be well-fit to the specific problem
  - Ways to accurately communicate constraint:
    - Plot prior and posterior together
      - Example: Grey background
      - Good place to put them: Marginal distributions
    - Calculate information gain
    - Be careful, clear, and precise

More guidance: https://github.com/stan-dev/stan/wiki/Prior-Choice-Recommendations

Summary

Summary: Bayesian Inference

- “Forward problem”: Model outputs from parameters; straightforward
- “Inverse problem”: Parameters from the model outputs; challenging

Bayesian inference:

  - Is a method for solving the inverse problem
  - Systematically compares experimental measurements with model calculations
  - Quantifies uncertainties AND incorporates domain knowledge

- Can handle:
  - non-linearity in model
  - any number of observables
  - complex experimental uncertainties
  - complex theoretical uncertainties
  - additional theoretical input through priors
- Gives constraints with quantified uncertainty
  - **Quantifying uncertainty is a goal**: inaccurate arbitrary precision is not

Summary: Bayesian Workflow

Workflows are an algorithm to ensure reliable, robust, and reproducible inference

Complex modeling means that without a workflow, there are many opportunities for things to fall through the cracks

- Get ahead of this and be critical and cautious!

Workflow results in:

- A well-motivated, self-contained, reproducible study
- A clear method for improving models
- Experience with modeling that builds toward a result
- Demonstration of familiarity and mastery with tools
- Many intermediary results:
  - Catch problems long before the end
  - Convince the audience you're right

Outlook: After this session

Know and have examples of:

- when to use Bayesian parameter estimation
- how to construct a reasonable prior
- how to produce predictive distributions
- how to validate a model and how to determine observable sensitivity to parameters
- how to use a common MCMC package
- some easy MCMC diagnostics
- how to decide that it is time to compare to real data
- how to use posterior predictive distributions

Acknowledgements

JETSCAPE Collaboration and the SIMS WG

Special thanks to Jean-François Paquet (Duke), Derek Everett (OSU), Weiyao Ke (UC Berkeley & LBNL) and Dan Liyanage (OSU) for materials, discussions and feedback on this talk

This work was supported in part by the Natural Sciences and Engineering Research Council of Canada

Practical aspect: emulators

with particular thanks to Jean-François Paquet

What if the model is slow?

"Y(p)" is our model outputs as a function of parameters p.
For a multi-observable model, think of "Y(p)" as a vector containing all the observables

In practice, the model should have a (relatively) smooth dependence on the parameters

Solution: **emulators**

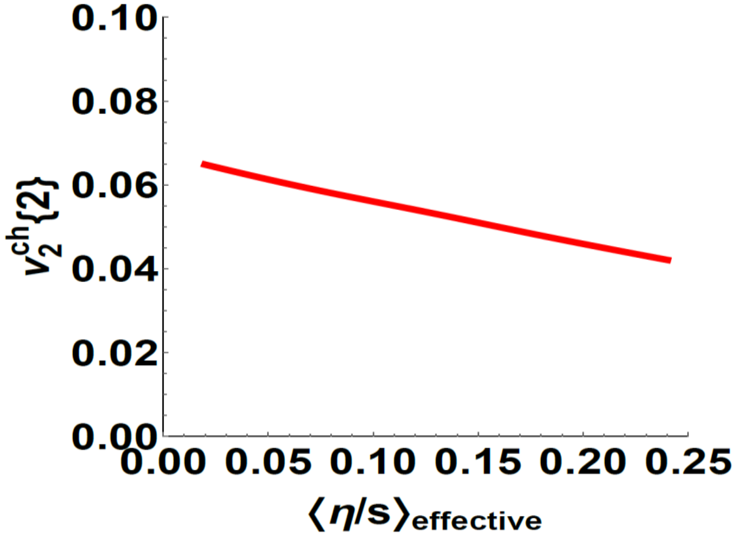
Recall that we already view the prediction of the model "Y(p)" as a probability distribution (which includes theoretical uncertainties)

$$\ln\left[p\left(d \mid H, I\right)\right] = -\frac{1}{2}\ln\left[(2\pi)^n \det\Sigma\right] - \frac{1}{2}\Delta\mathbf{y}^T\Sigma^{-1}\Delta\mathbf{y}$$

Length of the data: n

Difference between data and prediction: $\Delta\mathbf{y}(p) = \mathbf{d} - \mathbf{Y}(p)$

Covariance Matrix: Σ



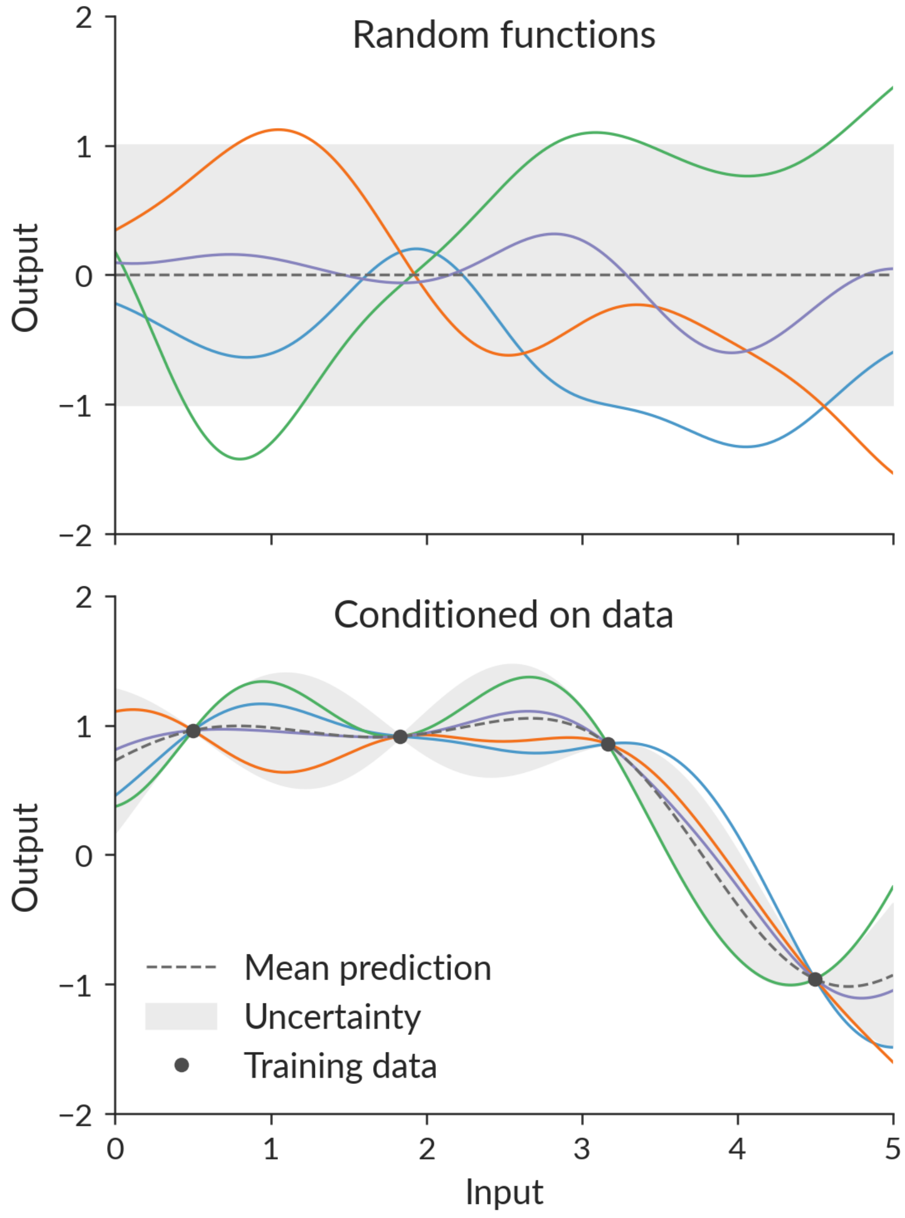
Emulators

Emulators are probability distributions that mimic the model outputs' dependence on the parameters

Constrained by the model at "design points"

Good emulators, like Gaussian process emulators, can estimate the interpolation uncertainty: extremely important!

Gaussian Process emulators (GPs) are just one example: design points are used train GP hyperparameters



Emulators: important additional information

Parameter space:
Range of parameters over which the emulator can mimic the output of the model
(the parameter space is sampled with the Latin hypercube algorithm)

The emulator is not for the model itself,
but for the model outputs (the observables: $v_2, dN/dy, R_{AA}, \ldots$)

For practical reasons, it's not even the model observables that are emulated:
It is linear combinations of model parameters
(identified through principal component analysis)

An independent Gaussian process emulator describes each major linear combinations of observables ("principal components")

Gaussian process emulators: more information

Excellent material from previous JETSCAPE Schools by Jake Coleman and Weiyao Ke

Jake Coleman lecturing at the 2018 JETSCAPE School in Berkeley
https://indico.bnl.gov/event/3958/
https://sites.google.com/a/lbl.gov/jetscape2018/home/school-material/school-preparation

Weiyao Ke lecturing at the 2019 JETSCAPE School in Texas A&M
https://indico.bnl.gov/event/5031/page/115-school-material ("School preparation - Statistical Analysis")
https://github.com/JETSCAPE/STAT/blob/master/WS_Theory_Exe