Bayesian Inverse Problems

Fehmi Cirak

University of Cambridge

UKACM-SEMNI Autumn School

Data-Centric Engineering in Computational Mechanics

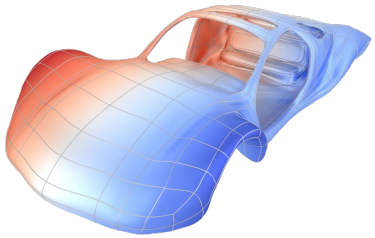
16–19 September 2024

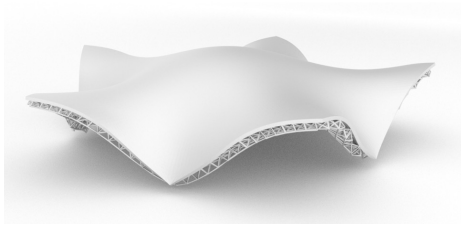
Who Are We?

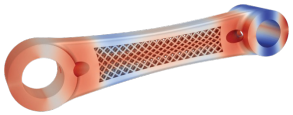
- Lecturer
  - Prof Fehmi Cirak
    - PhD in Computational Mechanics at University of Stuttgart (Germany)
    - Postdoc & Senior Scientist at Caltech (USA)
    - In Cambridge since 2006
- Hands-on Session Leaders
  - Yiren Wang (2nd year PhD, Tongji University, China)
  - Shehara Perera(2nd year PhD, University Moratuwa, Sri Lanka & Imperial College, UK)
  - Oguzhan Yuksel (4th year PhD, Middle Eastern Technical University, Turkey & University of Manchester, UK)

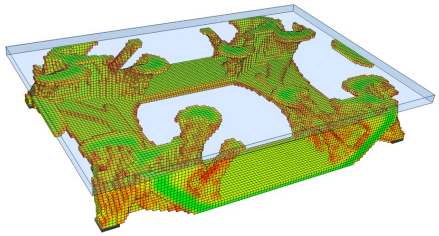
Where Are We Based?

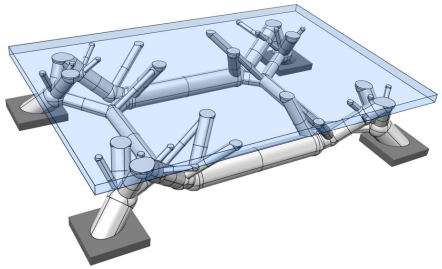
Our Research Themes







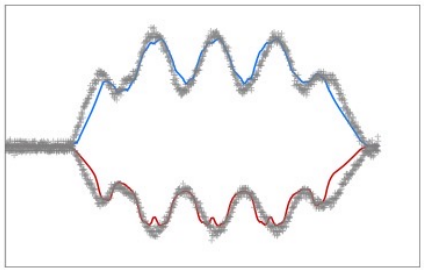


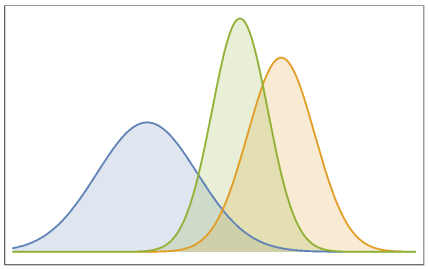


Simulation

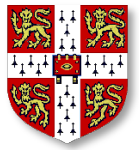
Geometry

Data & Sensors





Today's Schedule

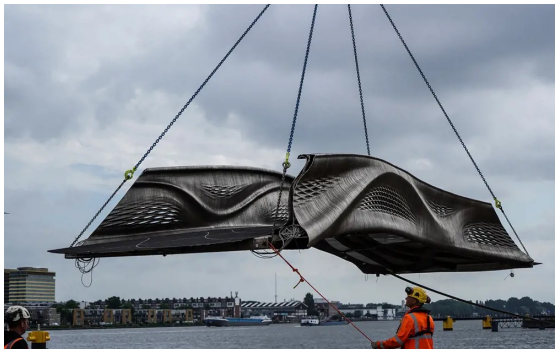


- 09:00–09:50 Bayesian Inverse Problems
  - 10:00–11:00 Hands-on Session
- 11:15–12:10 Gaussian Process Surrogates
  - 13:30–14:30 Hands-on Session
- 14:45–15:35 Statistical Finite Elements
  - 15:45–16:45 Hands-on Session
- 16:45–17:00 Summary & Discussion

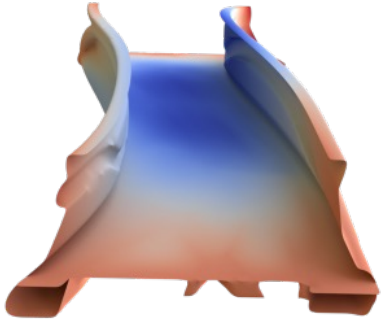
Asking questions by typing in Q&A or raising hand both are fine

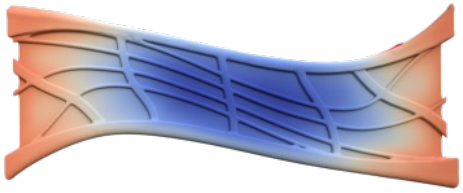
Forward and Inverse Problems

- World's first 3D printed bridge



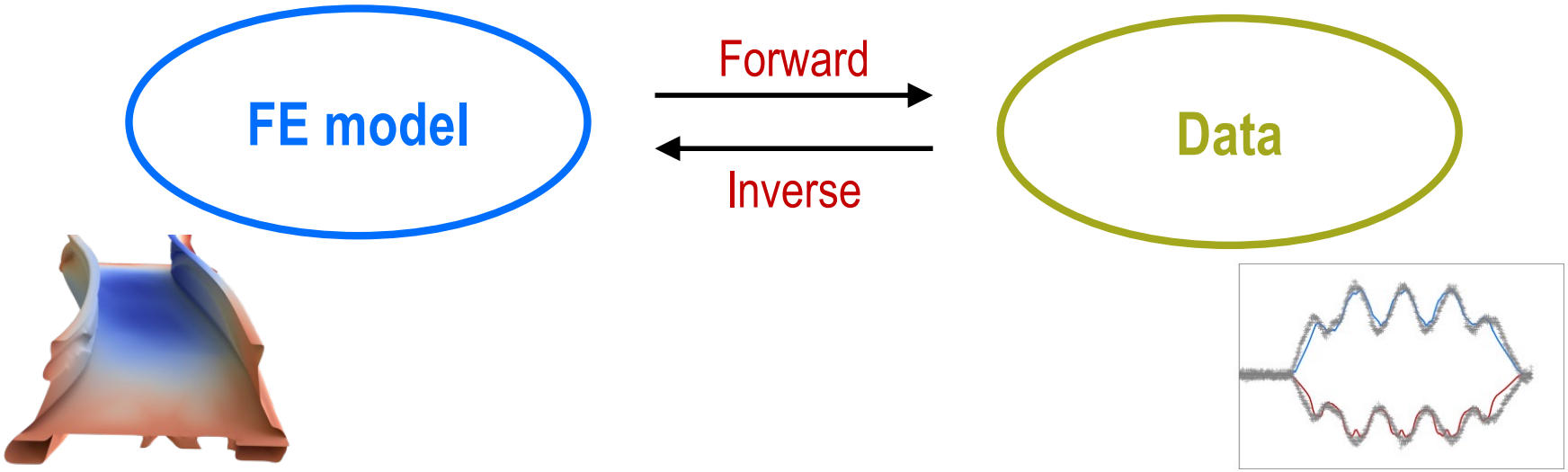
  - Finite element model





- Forward problem
  - Predict the displacements, stresses, etc of the actual structure
- Inverse problem
  - Infer FE model parameters from measured displacements, stresses, etc.

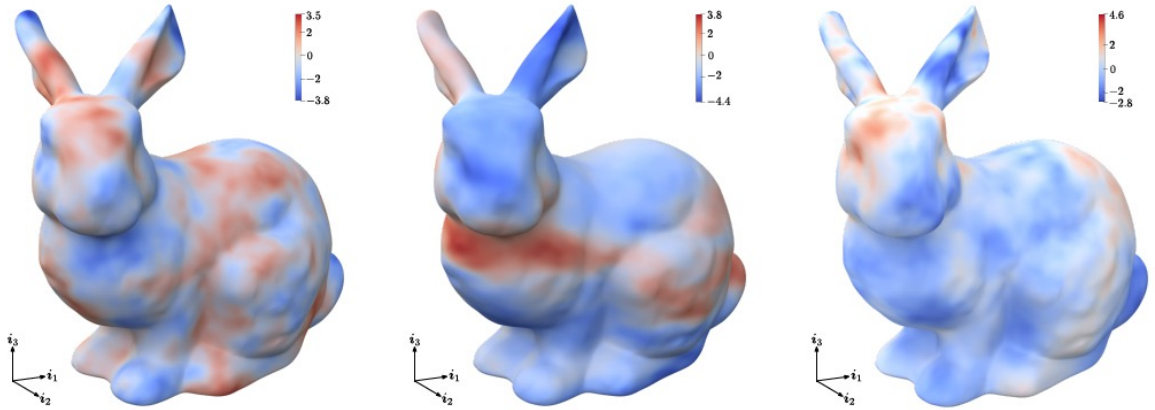
Forward and Inverse Problems



- Forward problems usually well-posed
  - Solution is unique and stable
    - Displacement and stress fields are unique and depend continuously on forcing
- Inverse problems often ill-posed
  - Parameters inferred from measurements usually not unique and stable
    - Same sensor displacement/strain can be obtained with different Young's modulus distributions

Regularisation (Big Picture)

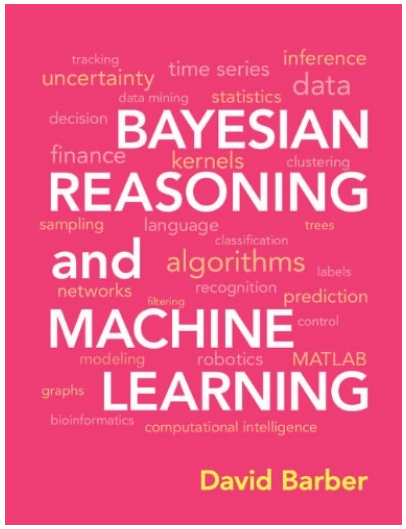
- Ill-posedness of inverse problems can be alleviated by regularisation
  - Sought model parameters must satisfy certain additional assumptions
    - E.g., the Young's modulus field must have a certain length-scale and distribution

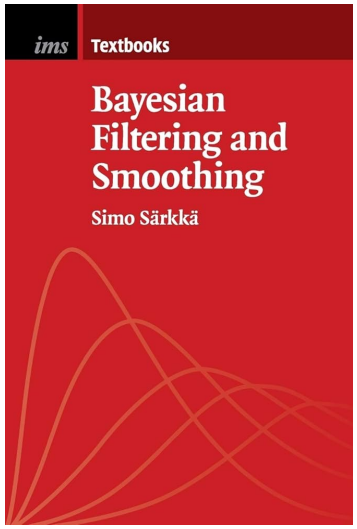


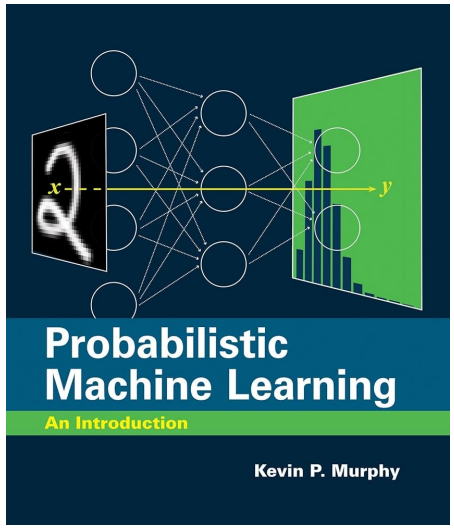
- Bayesian statistics provides a consistent framework for regularising inverse problems
  - Model parameters are assumed random variables with a corresponding probability density
  - Assumed *prior probably density* encode all existing knowledge before considering data
  - *Posterior probability density* determined by updating the prior probability density using the Bayes rule and taking into account observation data

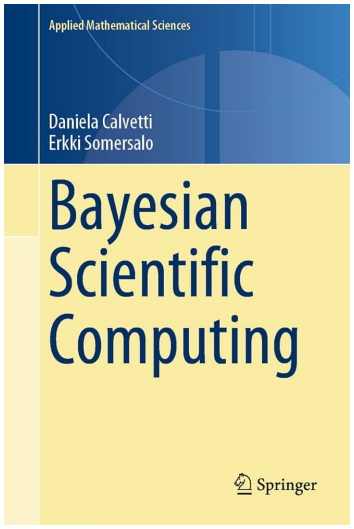
Bayesian Statistics

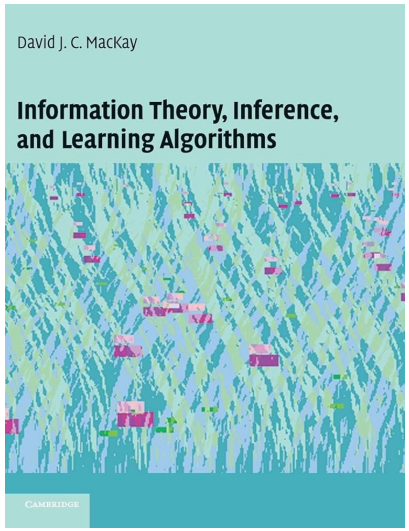
- Extensively used in probabilistic machine learning and computing

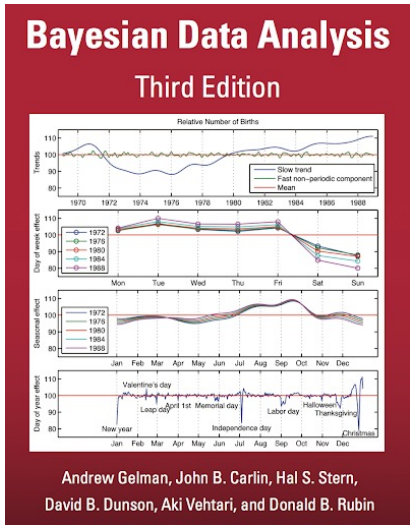


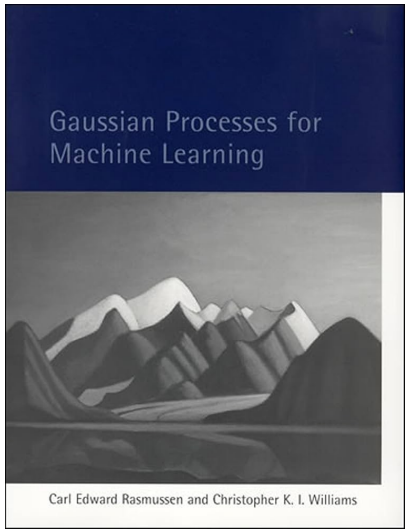


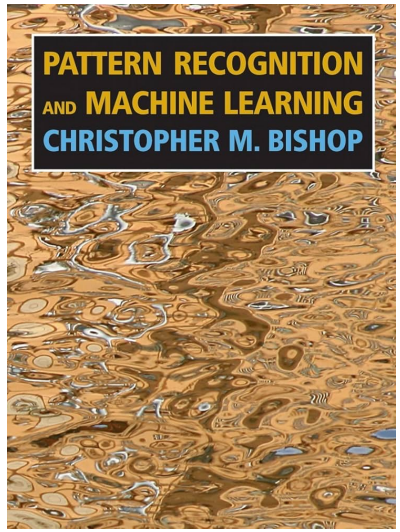












Outline

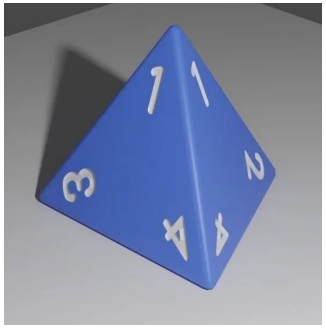
- Probability refresher
  - Discrete random variables
    - Probability mass function, summation and product rule
  - Continuous random variables
    - Probability density function, marginalisation and product rule
    - Gaussian probability densities and properties
- Bayesian formalism for solving inverse problems
  - Linear regression
  - Boundary value problems
  - Variational approximate inference

Probability Refresher – Discrete

Definitions – Four-sided Tetrahedral Dice

- Each throw (experiment) results in one of the four outcomes
- Sample space

$$\Omega = \{1, 2, 3, 4\}$$



- The outcome $x \in \Omega$ is a random variable
- Outcome of probabilities must add up to 1

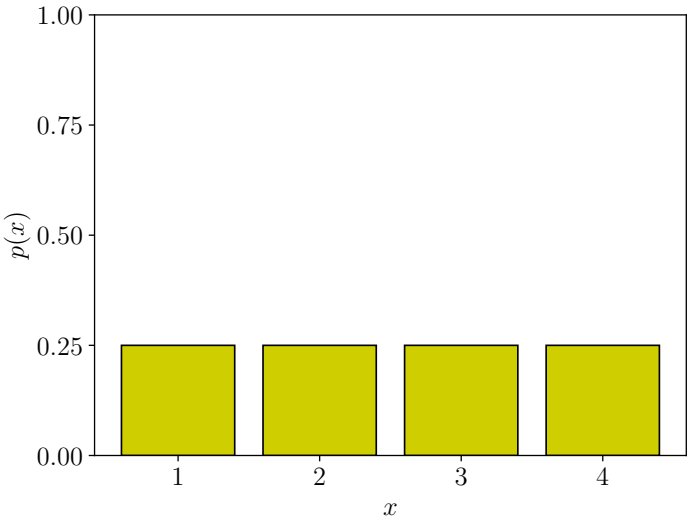
$$p(x=1) + p(x=2) + p(x=3) + p(x=4) = 1$$

  - $p(x)$ is called probability mass function, we will call it probability density

- Die is fair. hence

$$p(x=1) = p(x=2) = p(x=3) = p(x=4) = \frac{1}{4}$$

- An event is a subspace of the sample space Ω
  - Event space is the set of all subsets



Definitions & Rules –1–

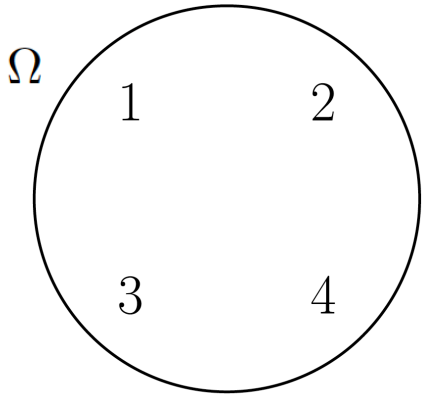
- Consider the two events
  - Outcomes larger than one: $L = \{ 2, 3, 4 \}$

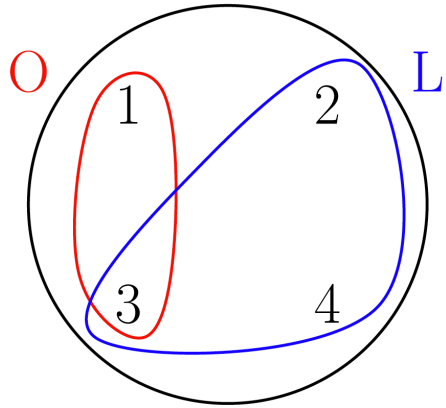
$$p(L) = p(x = 2) + p(x = 3) + p(x = 4) = \frac{3}{4}$$

    - Probability of event is the sum of probabilities of all outcomes
  - Outcomes that are odd: $O = \{ 1, 3 \}$

$$p(O) = p(x = 1) + p(x = 3) = \frac{1}{2}$$

  - Intersection oft the two events: $p(L \cap O) = p(x = 3) = \frac{1}{4}$
  - Union of two events: $p(L \cup O) = p(x = 1) + p(x = 2) + p(x = 3) + p(x = 4) = 1$





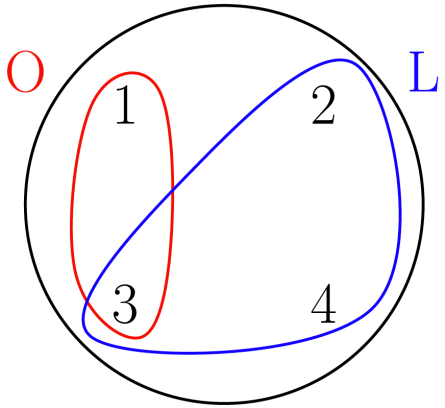
Definitions & Rules –2–

- Definition of conditional probabilities
  - Probability of L occurring given O has occurred

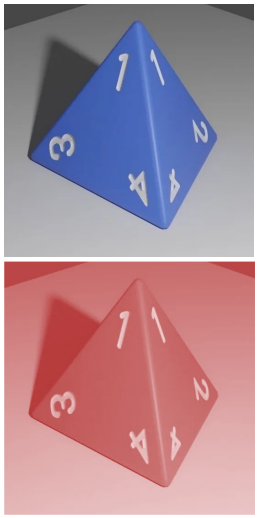
$$p(L|O) = \frac{p(L \cap O)}{p(O)} \qquad p(L|O) = \frac{1/4}{1/2} = \frac{1}{2}$$

  - Probability of O occurring given L has occurred

$$p(O|L) = \frac{p(O \cap L)}{p(L)} = \frac{1/4}{3/4} = \frac{1}{3}$$



Two Four-sided Tetrahedral Die



- Two dice in each throw
  - Two random variables x and y
  - Sample space consists of all outcomes

$$\Omega = \{(1,1),\ (1,2),\ (1,3),\ (1,4),\ (2,1),\ (2,2),\ (2,3),\ (2,4),\\ (3,1),\ (3,2),\ (3,3),\ (3,4),\ (4,1),\ (4,2),\ (4,3),\ (4,4)\}\,.$$

  - Random variables independent and dice are fair

$$p(x=1 \cap y=1) = p(x=1 \cap y=2) = \cdots = \frac{1}{16}$$

    - Probability table

$p(x,y)$	$y=1$	$y=2$	$y=3$	$y=4$
$x=1$	$1/16$	$1/16$	$1/16$	$1/16$
$x=2$	$1/16$	$1/16$	$1/16$	$1/16$
$x=3$	$1/16$	$1/16$	$1/16$	$1/16$
$x=4$	$1/16$	$1/16$	$1/16$	$1/16$

Rules

- Two rules of which will used throughout the day
  - Marginalisation /summation rule: $p(x) = \sum_y p(x, y)$
    - Marginal probability densities are simply the probabilities of x and y, respectively

$p(x,y)$	$y=1$	$y=2$	$y=3$	$y=4$	$p(x)$
$x=1$	$1/16$	$1/16$	$1/16$	$1/16$	$1/4$
$x=2$	$1/16$	$1/16$	$1/16$	$1/16$	$1/4$
$x=3$	$1/16$	$1/16$	$1/16$	$1/16$	$1/4$
$x=4$	$1/16$	$1/16$	$1/16$	$1/16$	$1/4$
$p(y)$	$1/4$	$1/4$	$1/4$	$1/4$	1

  - Product rule: $p(x \cap y) = p(y|x)\, p(x) \qquad p(y \cap x) = p(x|y)\, p(y)$
    - Note the product rule is simply the definition of conditional probability
- Bayes rule

$$p(x|y) = \frac{p(y|x)\, p(x)}{p(y)} \qquad p(x|y) = \frac{p(y|x)\, p(x)}{\sum_x p(y,x)} = \frac{p(y|x)\, p(x)}{\sum_x p(y|x)p(x)}$$

  - Follows from product rule: $p(x \cap y) = p(x|y)p(y) = p(y|x)p(x)$

Example – Bayesian Inference

- Continuing with the tetrahedral die consider the new random variable $z = x + y$
- If we know $z = 5$ what is the probability $p(x, y|z = 5)$?
  - Bayes rule

$$p(x, y|z = 5) = \frac{p(z = 5|x, y)p(x, y)}{\sum_{x,y} p(z = 5|x, y)p(x, y)}$$

  - Conditional density

$p(z = 5\|x, y)$	$y = 1$	$y = 2$	$y = 3$	$y = 4$
$x = 1$	0	0	0	1
$x = 2$	0	0	1	0
$x = 3$	0	1	0	0
$x = 4$	1	0	0	0

  - Hence, $p(x, y|z = 5) = \frac{p(z = 5|x, y)p(x, y)}{p(z = 5)} = \frac{p(z = 5|x, y)p(x, y)}{4/16}$
    - E.g., $p(x = 1, y = 2|z = 5) = \frac{0}{4/16} = 0$, $p(x = 2, y = 3|z = 5) = \frac{1/16}{4/16} = \frac{1}{4}$

Continuous Random Variables/Vectors

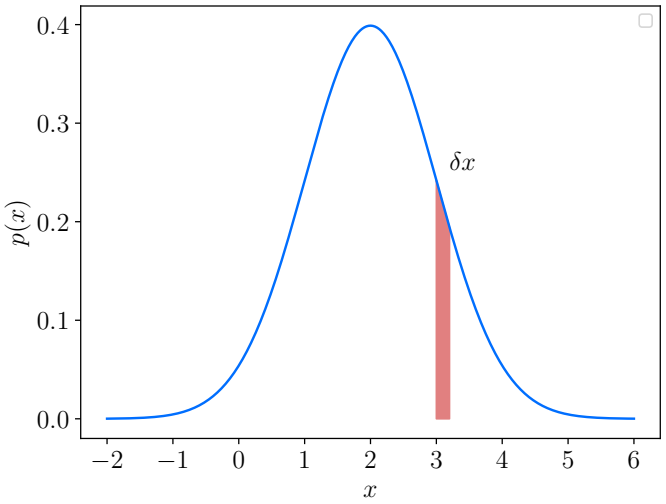
- Random variable $x \in \mathbb{R}$
  - Probability that x is in the interval $(\tilde{x} + \delta x)$

$$p(x \in (\tilde{x}, \tilde{x} + \delta x)) = \int_{\tilde{x}}^{\tilde{x}+\delta x} p(x)\, dx$$

  - Probability density function must satisfy

$$p(x) \geq 0$$

$$\int_{-\infty}^{\infty} p(x)\ \mathrm{d}x = 1$$



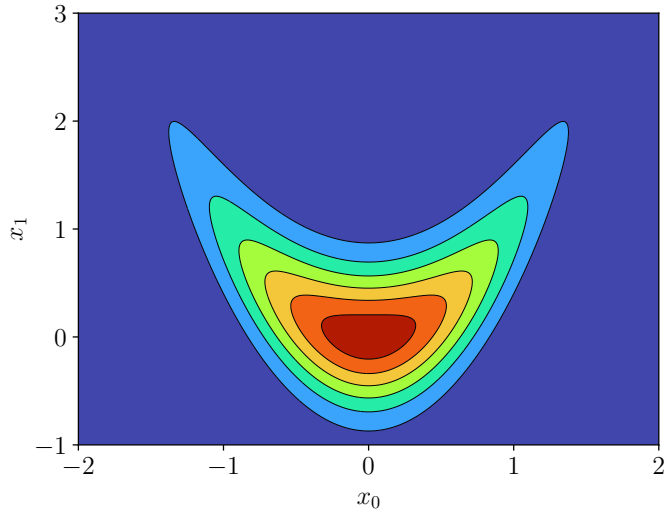
- Random vector $\boldsymbol{x} \in \mathbb{R}^d$
  - Probability that $\boldsymbol{x} \in \Omega \subseteq \mathbb{R}^d$

$$p(\boldsymbol{x} \in \Omega) = \int_{\Omega} p(\boldsymbol{x})\ \mathrm{d}\boldsymbol{x}$$

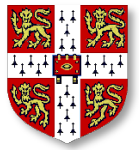
  - Probability density function must satisfy

$$p(\boldsymbol{x}) \geq 0$$

$$\int_{-\infty}^{\infty} p(\boldsymbol{x})\ \mathrm{d}\boldsymbol{x} = 1$$



Expectation, Variance and Covariance



- Random variable $x \in \mathbb{R}$
  - Expectation, or mean,

$$\mathbb{E}[x] = \int_{-\infty}^{\infty} x p(x)\, \mathrm{d}x := \overline{x}$$

  - Variance

$$\mathrm{var}[x] = \mathbb{E}[(x - \overline{x})^2] = \mathbb{E}[x^2] - \overline{x}^2$$

- Random vector $\boldsymbol{x} \in \mathbb{R}^d$
  - Expectation, or mean, vector

$$\mathbb{E}[\boldsymbol{x}] = \int_{-\infty}^{\infty} \boldsymbol{x} p(\boldsymbol{x})\, \mathrm{d}\boldsymbol{x} := \overline{\boldsymbol{x}}$$

    - Mean of components
  - Covariance matrix

$$\mathrm{cov}[\boldsymbol{x}, \boldsymbol{x}] = \mathbb{E}\left[(\boldsymbol{x} - \overline{\boldsymbol{x}})(\boldsymbol{x} - \overline{\boldsymbol{x}})^{\mathsf{T}}\right] = \mathbb{E}\left[\boldsymbol{x}\boldsymbol{x}^{\mathsf{T}}\right] - \overline{\boldsymbol{x}}\,\overline{\boldsymbol{x}}^{\mathsf{T}}$$

    - Matrix of dimension $d \times d$
    - Matrix entries capture the variance, covariance between the components, i.e. $\mathrm{cov}[x_i, x_j]$

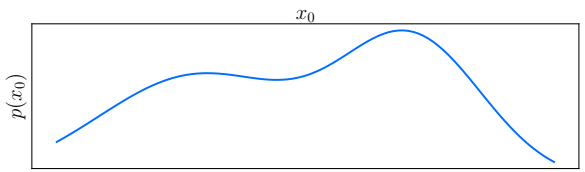
Marginalisation and Product Rules

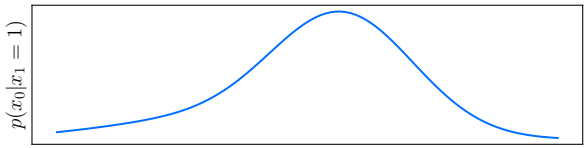
- Consider, e.g., the random vector
  - Marginalisation rule

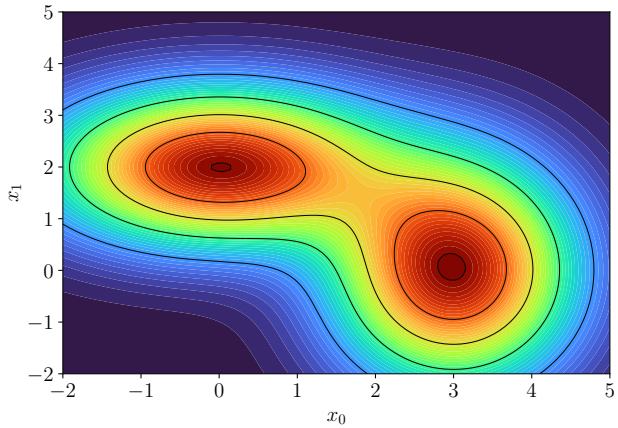
$$p(x_0) = \int_{-\infty}^{\infty} p(x_0, x_1)\ \mathrm{d}x_1$$

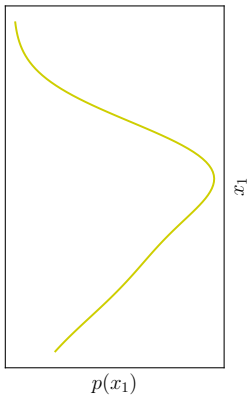
  - Product rule

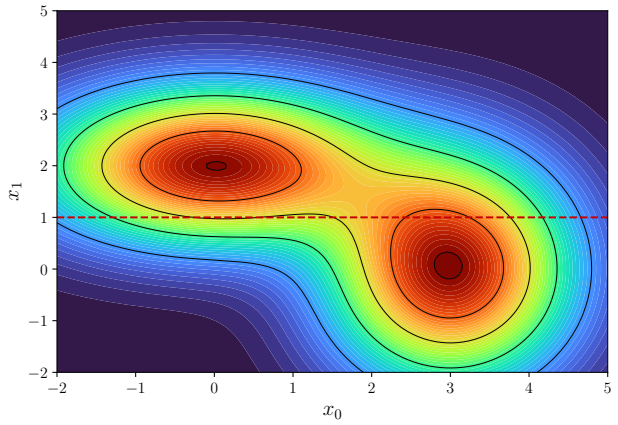
$$p(x_0, x_1) = p(x_0|x_1)p(x_1)$$









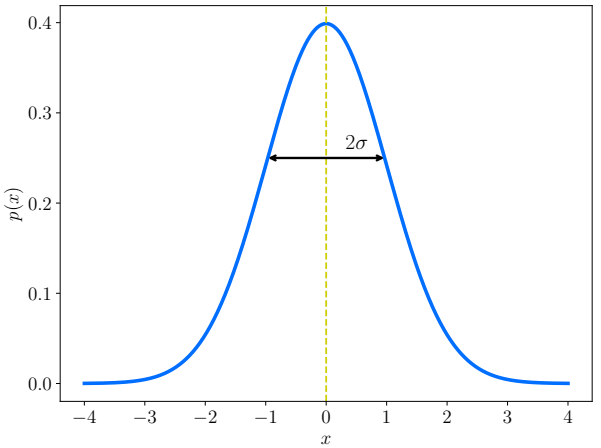


Gaussian Probability Densities

- Univariate Gaussian probability density for $x \in \mathbb{R}$

$$p(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x-\overline{x})^2}{2\sigma^2}\right)$$

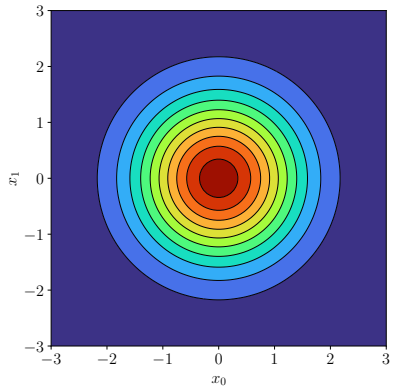
  - Standard deviation: $\sigma = \sqrt{\text{var}[x]}$



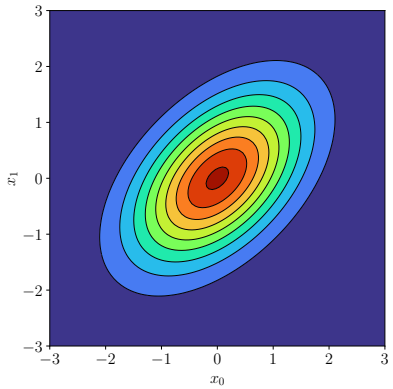
- Bivariate Gaussian probability density for $\boldsymbol{x} \in \mathbb{R}^2$

$$p(\boldsymbol{x}) = \frac{1}{(2\pi)^{d/2}\det(\boldsymbol{C})^{1/2}} \exp\left(-\frac{1}{2}(\boldsymbol{x}-\overline{\boldsymbol{x}})^{\mathsf{T}}\boldsymbol{C}^{-1}(\boldsymbol{x}-\overline{\boldsymbol{x}})\right)$$

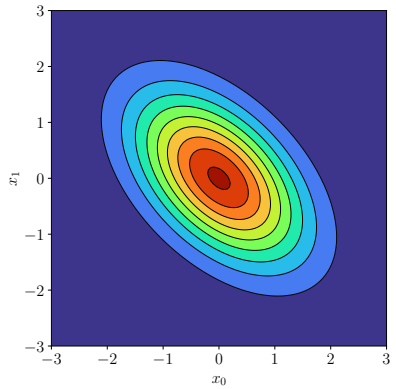
  - Dependence of probability density function on covariance matrix $\boldsymbol{C} = \text{cov}[\boldsymbol{x}, \boldsymbol{x}]$



$$\boldsymbol{C} = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}$$



$$\boldsymbol{C} = \begin{pmatrix} 1 & 0.5 \\ 0.5 & 1 \end{pmatrix}$$



$$\boldsymbol{C} = \begin{pmatrix} 1 & -0.5 \\ -0.5 & 1 \end{pmatrix}$$

Properties of Gaussians

- Gaussian probability densities are very easy to work with
- Consider, e.g., Gaussian random variables

$$x \sim \mathcal{N}\left(\overline{x}, \sigma_x^2\right), \quad y \sim \mathcal{N}\left(\overline{y}, \sigma_y^2\right)$$

  - Sum of two Gaussian variables is a Gaussian variable

$$x + y \sim \mathcal{N}\left(\overline{x} + \overline{y}, \sigma_x^2 + \sigma_y^2\right)$$

  - Product of a Gaussian variable with a constant is a Gaussian variable

$$cx \sim \mathcal{N}\left(c\overline{x}, c^2\sigma_x^2\right)$$

- Same relationships hold for Gaussian random vectors

$$\boldsymbol{x} \sim \mathcal{N}\left(\overline{\boldsymbol{x}}, \boldsymbol{C}_x\right), \quad \boldsymbol{y} \sim \mathcal{N}\left(\overline{\boldsymbol{y}}, \boldsymbol{C}_y\right)$$

$$\boldsymbol{x} + \boldsymbol{y} \sim \mathcal{N}\left(\overline{\boldsymbol{x}} + \overline{\boldsymbol{y}}, \boldsymbol{C}_x + \boldsymbol{C}_y\right)$$

$$\boldsymbol{A}\boldsymbol{x} \sim \mathcal{N}\left(\boldsymbol{A}\overline{\boldsymbol{x}}, \boldsymbol{A}\boldsymbol{C}_x\boldsymbol{A}^{\mathsf{T}}\right),$$

Bayesian Statistics

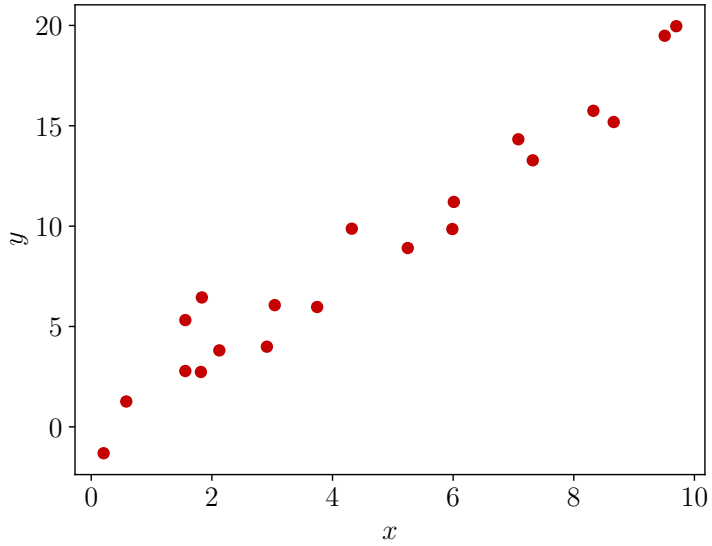
- We mentioned so far mainly the frequentist interpretation of probability
  - Proportion of events after infinite number trials
- Many events of interest not repeatable, e.g., collapse of a bridge or weather on a given day
- In Bayesian statistics probability specifies the belief about the certainty of an event, i.e. certain: 1 and uncertain: 0
  - Different observers may assign different probabilities to the same event
  - Rules of probability are correct irrespective of the assigned probabilities
- Assigned probabilities are updated using the Bayes rule as new data becomes available

$$p(\boldsymbol{w}|\mathcal{D}) = \frac{p(\mathcal{D}|\boldsymbol{w})p(\boldsymbol{w})}{p(\mathcal{D})} \qquad \text{posterior} = \frac{\text{likelihood} \times \text{prior}}{\text{evidence}}$$

  - $\mathcal{D}$ is the data and $\boldsymbol{w}$ the random vector of interest
  - Likelihood is the probability of observing $\mathcal{D}$ for a given/fixed $\boldsymbol{w}$
  - Evidence, or marginal likelihood, is the probability of observing $\mathcal{D}$ with the chosen model

Bayesian Inference

Linear Regression – Example



- Observations

$$\mathcal{D} = \{(x_i, y_i)\}_{i=0}^{n-1} \qquad \boldsymbol{x} = (x_0, x_1, \ldots, x_{n-1})^\mathsf{T} \qquad \boldsymbol{y} = (y_0, y_1, \ldots, y_{n-1})^\mathsf{T}$$

- Data generating model (line with intercept zero)

$$y = wx + \epsilon$$

  - Prior for random slope : $p(w) = \mathcal{N}(0, \sigma_w^2) = \frac{1}{\sqrt{2\pi\sigma_w^2}} \exp\left(-\frac{w^2}{2\sigma_w^2}\right)$
  - Prescribed random noise: $p(\epsilon) = \mathcal{N}(0, \sigma_\epsilon^2) = \frac{1}{\sqrt{2\pi\sigma_\epsilon^2}} \exp\left(-\frac{\epsilon^2}{2\sigma_\epsilon^2}\right)$

Linear Regression – Bayes Rule

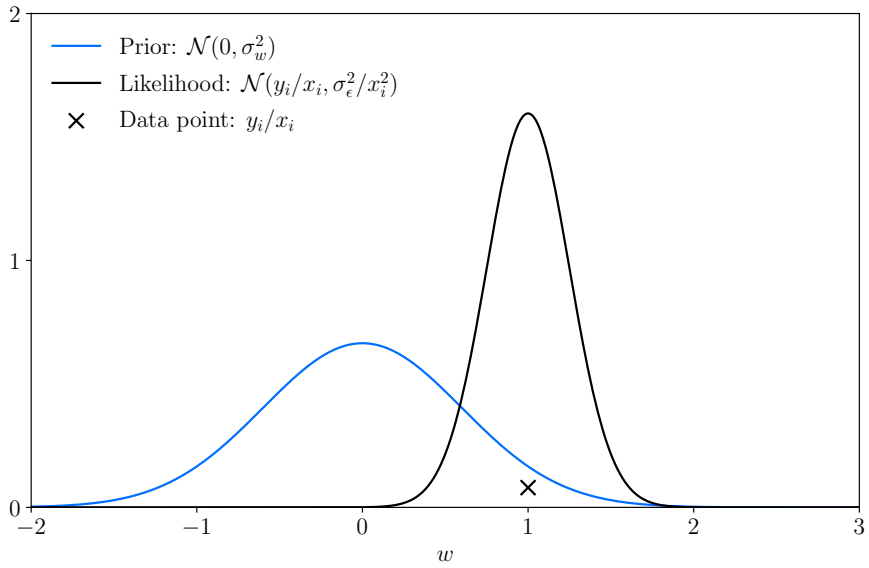
- Bayes rule for updating the prior $p(w)$ in light of a single data point (x_i, y_i)

$$p(w|y_i) = \frac{p(y_i|w)p(w)}{p(y_i)}$$

  - Likelihood and marginal likelihood are easy to derive using standard rules for addition and multiplication

$$p(y_i|w) = \mathcal{N}(wx_i, \sigma_\epsilon^2) = \frac{1}{\sqrt{2\pi\sigma_\epsilon^2}} \exp\left(-\frac{(y_i - wx_i)^2}{2\sigma_\epsilon^2}\right)$$

$$p(y_i) = \int p(y_i|w)p(w)\, \mathrm{d}w = \mathcal{N}(0, x_i^2\sigma_w^2 + \sigma_\epsilon^2)$$



$\sigma_w = 0.6, \sigma_\epsilon = 0.5$

Linear Regression – MAP Estimate

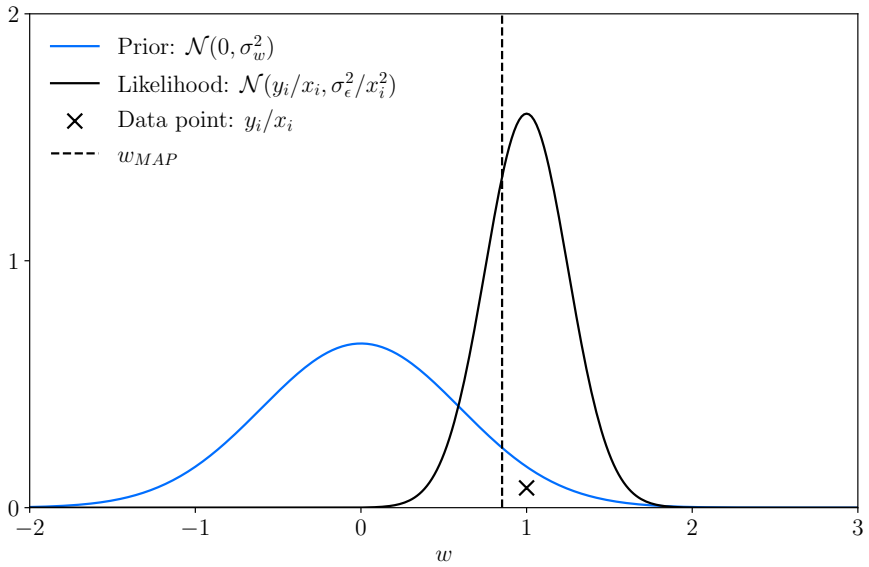
- Maximum a posterior (MAP) estimate for w

$$w_{\mathrm{MAP}} = \arg\max_w p(w|y_i) = \arg\max_w p(y_i|w)p(w)$$

  - Because the logarithm is a convex function

$$w_{\mathrm{MAP}} = \arg\max_w \log p(w|y_i) = \arg\max_w \log p(y_i|w) + \log p(w)$$

  - Can be minimized with standard gradient-based algorithms, like LBFGS



$\sigma_w = 0.6, \sigma_\epsilon = 0.5$

Linear Regression – Posterior

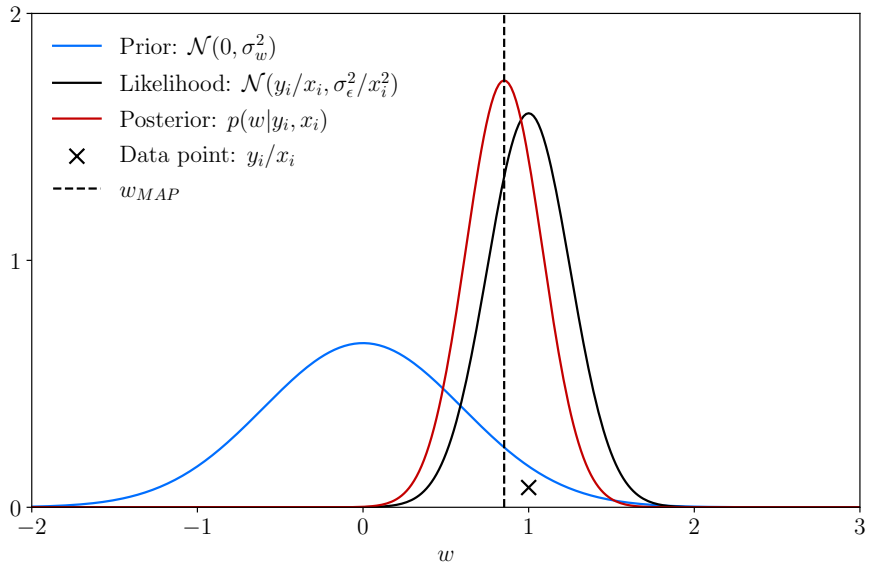
- Or, more explicitly

$$w_{\text{MAP}} = \arg\max_w \left(-\frac{(y_i - wx_i)^2}{2\sigma_\epsilon^2} - \frac{w^2}{2\sigma_w^2} \right) = \frac{x_i y_i \sigma_w^2}{\sigma_\epsilon^2 + x_i^2 \sigma_w^2}$$

  - For $\sigma_\epsilon = 0$ this yields $w = y_i/x_i$ and for $\sigma_\epsilon \to \infty$ it yields $w \to 0$

- Easy to show that posterior is a Gaussian as well
  - Product of two Gaussian probability densities is a (non-normalised) Gaussian

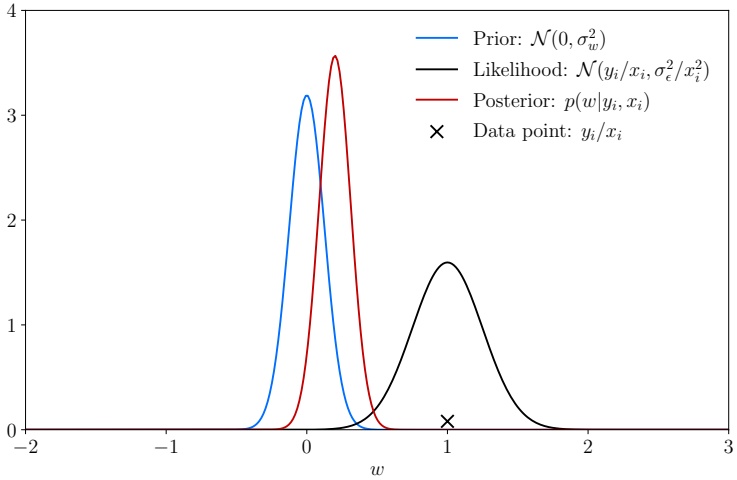
$$p(w|y_i) = \mathcal{N}(w_{|y}, \sigma_{w|y}) \qquad \sigma_{w|y}^2 = \left(\frac{1}{\sigma_w^2} + \frac{x_i^2}{\sigma_\epsilon^2} \right)^{-1} \qquad w_{|y} = \sigma_{w|y}^2 \frac{x_i y_i}{\sigma_\epsilon^2}$$



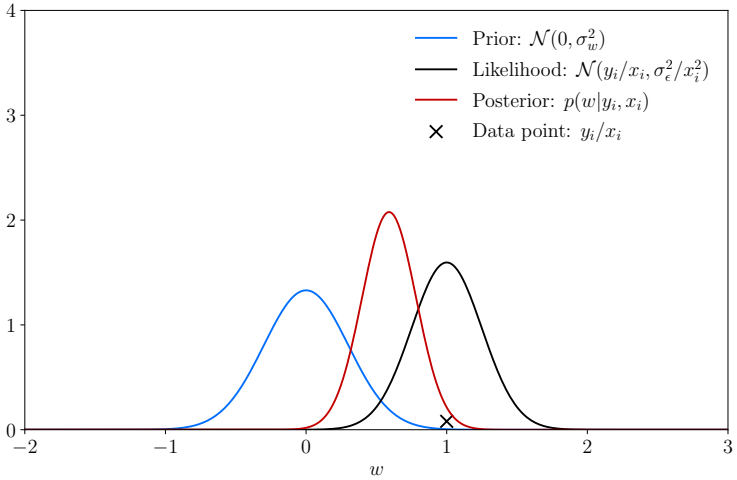
$\sigma_w = 0.6, \sigma_\epsilon = 0.5$

Linear Regression – Parameter Study

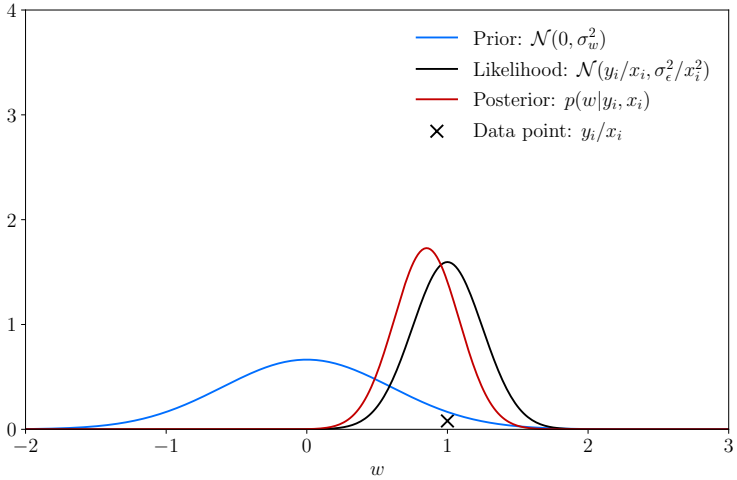
- Influence of σ_w and σ_ϵ on the posterior



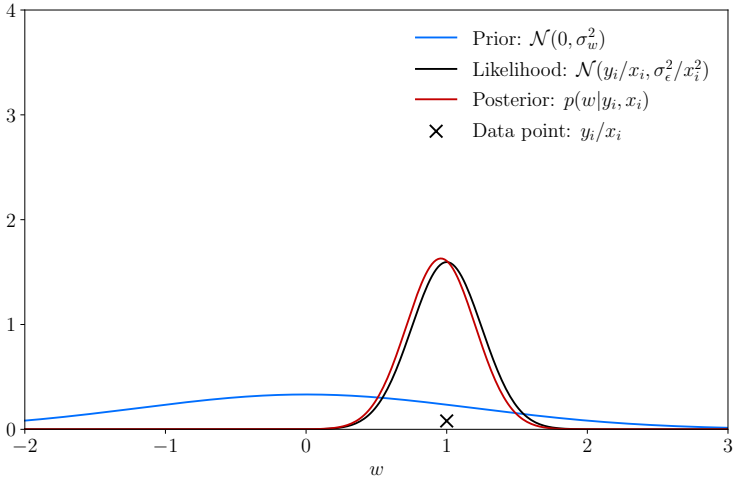
$\sigma_w = 0.125, \sigma_\epsilon = 0.5$



$\sigma_w = 0.3, \sigma_\epsilon = 0.5$



$\sigma_w = 0.6, \sigma_\epsilon = 0.5$



$\sigma_w = 1.2, \sigma_\epsilon = 0.5$

Linear Regression – More Observations

- Observations are assumed independent and identically distributed (as usual)

$$\mathcal{D} = \{(x_i, y_i)\}_{i=0}^{n-1} \qquad \boldsymbol{x} = (x_0, x_1, \ldots, x_{n-1})^\mathsf{T} \qquad \boldsymbol{y} = (y_0, y_1, \ldots, y_{n-1})^\mathsf{T}$$

- Data likelihood

$$p(\boldsymbol{y}|w) = \prod_{i=0}^{n-1} p(y_i|w) = \prod_{i=0}^{n-1} \mathcal{N}(wx_i, \sigma_\epsilon^2) = \frac{1}{(2\pi\sigma_\epsilon^2)^{(n-1)/2}} \exp\left(-\frac{1}{2\sigma_\epsilon^2}\sum_{i=0}^{n-1}(y_i - wx_i)^2\right)$$

- Bayes rule for updating the prior $p(w)$ in light of data points $\mathcal{D}$

$$p(w|\boldsymbol{y}) = \frac{p(\boldsymbol{y}|w)p(w)}{p(\boldsymbol{y})}$$

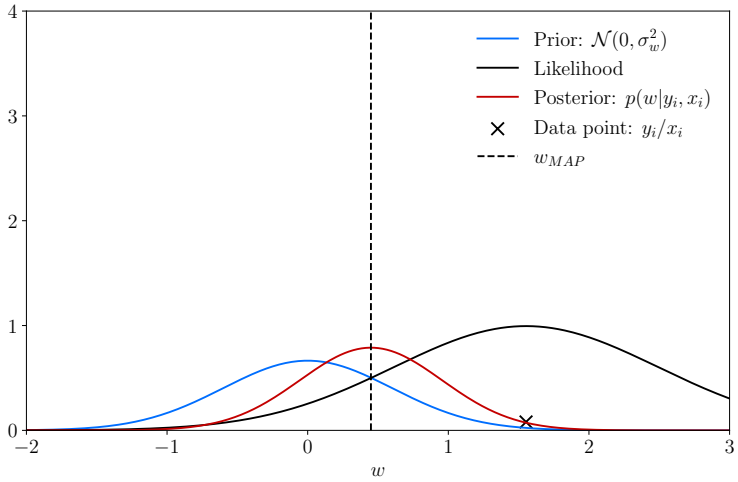
- Easy to show that posterior density is a Gaussian as well
  - Product of two Gaussian probability densities is a (non-normalised) Gaussian

$$p(w|\boldsymbol{y}) = \mathcal{N}\left(w_{|y}, \sigma_{w|y}\right)$$

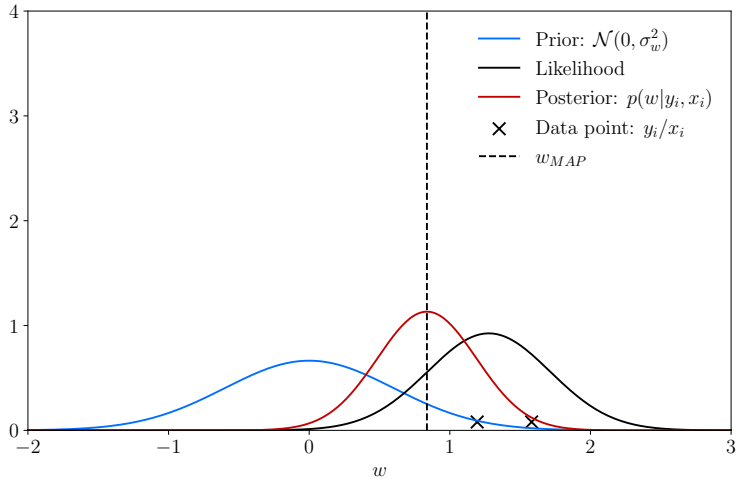
$$w_{|y} = \sigma_{w|y}^2 \frac{\sum_{i=0}^{n-1} x_i y_i}{\sigma_\epsilon^2} \qquad \sigma_{w|y}^2 = \left(\frac{1}{\sigma_w^2} + \frac{\sum_{i=0}^{n-1} x_i^2}{\sigma_\epsilon^2}\right)^{-1}$$

Linear Regression – Parameter Study

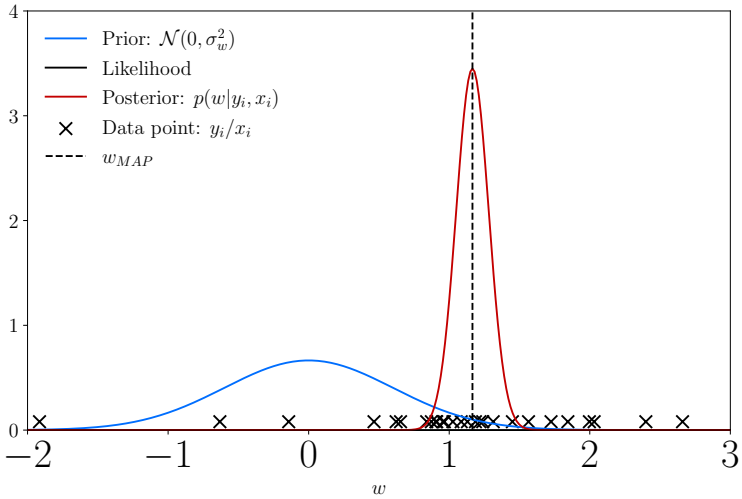
- Influence of the number of observation points n with $\sigma_w = 0.6, \sigma_\epsilon = 0.4$



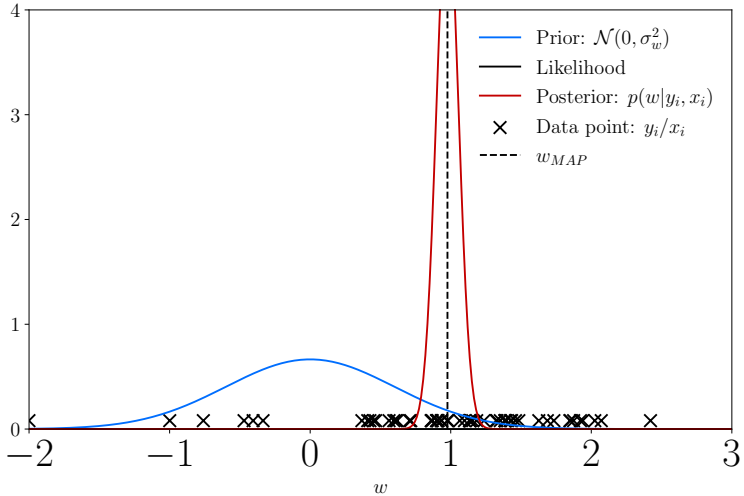
$n = 1$



$n = 2$



$n = 8$



$n = 64$

Linear Regression – Parameter Study

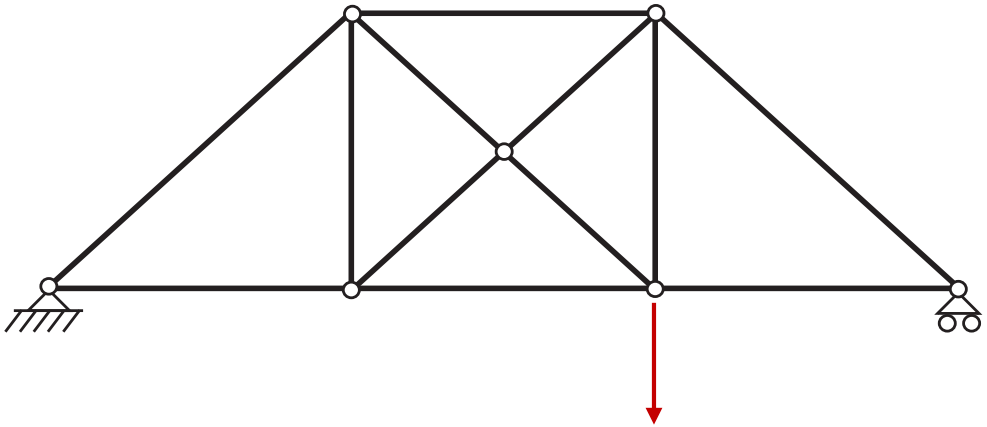
- With increasing n posterior mean converges to sample mean and variance to zero

$$w_{|y} \rightarrow \frac{1}{n}\sum_{i=0}^{n-1}\frac{y_i}{x_i}\,, \qquad \sigma_{w|y} \rightarrow 0$$

  - Variance captures the (epistemic) uncertainty in the estimate of the slope

Inverse Problems in Mechanics

- Problem statement: Infer the parameters a partial differential equation from observed data
- Illustrative example problem: pin-jointed truss structure
  - Axial stiffness of of some or all members is not known
  - Some joint displacements and/or membrane strains are observed



  - Equilibrium: $\boldsymbol{A}(\boldsymbol{w})\boldsymbol{u} = \boldsymbol{f}$
  - Joint displacements: $\boldsymbol{u} = \boldsymbol{A}^{-1}(\boldsymbol{w})\boldsymbol{f}$
    - Forcing $\boldsymbol{f}$ is fixed
    - Unknown member axial stiffness vector $\boldsymbol{w}$ is assumed random

Bayesian Inference –1–

- Observation model

$$\boldsymbol{y} = \boldsymbol{H}\boldsymbol{u}(\boldsymbol{w}) + \boldsymbol{\epsilon}$$

  - Random noise

$$p(\boldsymbol{\epsilon}) = \mathcal{N}(0, \sigma_\epsilon^2 \boldsymbol{I}) = \frac{1}{\sqrt{2\pi\sigma_\epsilon^2}} \exp\left(-\frac{\boldsymbol{\epsilon}^\mathsf{T}\boldsymbol{\epsilon}}{2\sigma_\epsilon^2}\right)$$

  - Axial stiffness prior

$$p(\boldsymbol{w}) = \mathcal{N}(\overline{\boldsymbol{w}}, \boldsymbol{C}) = \frac{1}{\sqrt{2\pi \det(\boldsymbol{C})}} \exp\left(-\frac{1}{2}(\boldsymbol{w} - \overline{\boldsymbol{w}})^\mathsf{T}\boldsymbol{C}^{-1}(\boldsymbol{w} - \overline{\boldsymbol{w}})\right)$$

  - Observation matrix $\boldsymbol{H}$ maps $\boldsymbol{u}$ to observed variables (displacements, strains, …)

- Bayes rule for updating the prior $p(\boldsymbol{w})$ in light of a single observation $\boldsymbol{y}$

$$p(\boldsymbol{w}|\boldsymbol{y}) = \frac{p(\boldsymbol{y}|\boldsymbol{w})p(\boldsymbol{w})}{p(\boldsymbol{y})}$$

- Posterior not a Gaussian and not analytically tractable
  - Likelihood depends on $\boldsymbol{A}^{-1}(\boldsymbol{w})$

Bayesian Inference –2–

- Posterior can be sampled using Markov Chain Monte Carlo (MCMC)
  - Many, many steps needed for converge
  - Each MCMC step requires one finite element solve
  - Not feasible for reasonably sized problems
- A MAP estimate can be obtained by maximising the posterior density

$$\boldsymbol{w}_{\mathrm{MAP}} = \arg\max_{\boldsymbol{w}} p(\boldsymbol{y}|\boldsymbol{w})p(\boldsymbol{w})$$

$$\boldsymbol{w}_{\mathrm{MAP}} = \arg\max_{\boldsymbol{w}} \log p(\boldsymbol{y}|\boldsymbol{w}) + \log p(\boldsymbol{w})$$

$$\boldsymbol{w}_{\mathrm{MAP}} = \arg\max_{\boldsymbol{w}} \left(-\frac{1}{2\sigma_\epsilon^2}(\boldsymbol{y} - \boldsymbol{H}\boldsymbol{A}^{-1}(\boldsymbol{w})\boldsymbol{f})^{\mathsf{T}}(\boldsymbol{y} - \boldsymbol{H}\boldsymbol{A}^{-1}(\boldsymbol{w})\boldsymbol{f}) - \frac{1}{2}(\boldsymbol{w} - \overline{\boldsymbol{w}})^{\mathsf{T}}\boldsymbol{C}^{-1}(\boldsymbol{w} - \overline{\boldsymbol{w}}) \right)$$

  - Can be minimised with standard gradient-based algorithms, like LBFGS
  - MAP estimation is equivalent to classical regularised least-squares
  - Only a point estimate with no information about uncertainty

Variational Inference –1–

- Posterior approximated with an assumed trial density by minimising their difference
  - Trial density

$$q(\boldsymbol{w}) = \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$$

  - Components of the mean vector and covariance matrix are the unknown variables
- Kullback-Leibler (KL) divergence measuring the difference between probability densities

$$\mathrm{KL}(q(\boldsymbol{w})||p(\boldsymbol{w}|\boldsymbol{y})) = \int q(\boldsymbol{w}) \log \frac{q(\boldsymbol{w})}{p(\boldsymbol{w}|\boldsymbol{y})} \, \mathrm{d}\boldsymbol{w}$$

  - KL divergence rewritten

$$\begin{aligned}\mathrm{KL}(q(\boldsymbol{w})||p(\boldsymbol{w}|\boldsymbol{y})) &= \int q(\boldsymbol{w}) \log q(\boldsymbol{w}) \, \mathrm{d}\boldsymbol{w} - \int q(\boldsymbol{w}) \log p(\boldsymbol{w}|\boldsymbol{y}) \, \mathrm{d}\boldsymbol{w} \\ &= \int q(\boldsymbol{w}) \log q(\boldsymbol{w}) \, \mathrm{d}\boldsymbol{w} - \int q(\boldsymbol{w}) \left(\log p(\boldsymbol{y}|\boldsymbol{w}) + \log p(\boldsymbol{w})\right) \, \mathrm{d}\boldsymbol{w} + \log p(\boldsymbol{w})\end{aligned}$$

Variational Inference –2–

- Evidence lower bound (ELBO)

$$\log p(\boldsymbol{w}) \geq \int q(\boldsymbol{w}) \left(\log p(\boldsymbol{y}|\boldsymbol{w}) + \log p(\boldsymbol{w})\right) \, \mathrm{d}\boldsymbol{w} - \int q(\boldsymbol{w}) \log q(\boldsymbol{w}) \, \mathrm{d}\boldsymbol{w}$$

$$= \int q(\boldsymbol{w}) \log p(\boldsymbol{y}|\boldsymbol{w}) \, \mathrm{d}\boldsymbol{w} - \mathrm{KL}(q(\boldsymbol{w})||p(\boldsymbol{w})).$$

  - Marginal likelihood $p(\boldsymbol{w})$ constant for a given dataset

- ELBO maximization using stochastic gradient descent

$$\boldsymbol{\mu}, \boldsymbol{\Sigma} = \arg\max_{\boldsymbol{\mu}, \boldsymbol{\Sigma}} \int q(\boldsymbol{w}) \log p(\boldsymbol{y}|\boldsymbol{w}) \, \mathrm{d}\boldsymbol{w} - \mathrm{KL}(q(\boldsymbol{w})||p(\boldsymbol{w}))$$

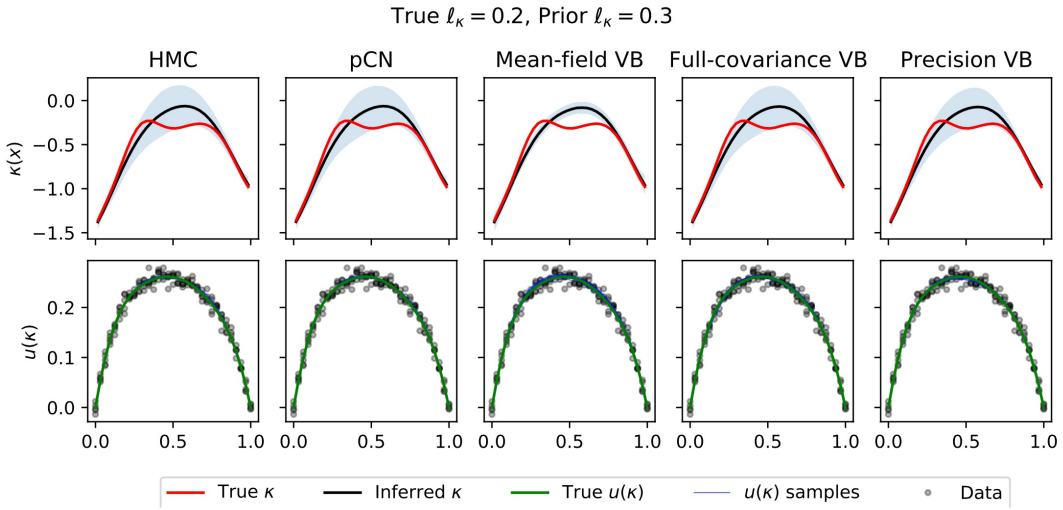
  - KL divergence term is easy to compute when $q(\boldsymbol{w})$ and $p(\boldsymbol{w})$ are Gaussians
  - Expectation term is approximated using Monte Carlo

$$\int q(\boldsymbol{w}) \log p(\boldsymbol{y}|\boldsymbol{w}) \approx \frac{1}{N} \sum_{i=1}^{N} \log p(\boldsymbol{y}|\boldsymbol{w}_i) \qquad \text{with} \quad \boldsymbol{w}_i \sim q(\boldsymbol{w})$$

    - Only very few samples sufficient in practice

Example

- Inferring the Young's modulus distribution in 1D Poisson problem
  - First two columns MCMC and last three columns variational inference



  - Runtimes

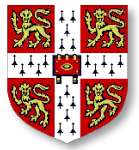
true ℓ_κ	prior ℓ_κ	Time (hours)				
		HMC		MFVB	FCVB	PMVB
0.1	0.1	15.2	(871–3244)	1.1	3.6	2.1
	0.2	11.1	(1043–4006)	0.7	2.7	2.1
	0.3	7.2	(1130–5408)	0.6	2.3	2.0
0.2	0.1	15.2	(1600–4700)	0.6	2.2	1.8
	0.2	10.4	(1067–3468)	0.6	2.3	2.0
	0.3	7.0	(1487–3969)	0.5	1.7	1.8

Table 1: Run-times for different inference schemes in hours for the Poisson 1D problem. For VB methods, $N_{\text{SVI}} = 3$. The column for HMC includes the range of effective sample sizes (ESS) across different components of $\boldsymbol{\kappa}$.

Summary

- Inverse problems are inherently ill-posed
- Bayesian formalism provides an effective approach for regularisation
  - Parameters of interest are treated as random variables
  - Prior knowledge, before seeing data, is encoded in the prior probability density
  - Likelihood of data is determined by the chosen observation model
  - Bayes rule is used to combine prior and data likelihood to compute the posterior density
- In non-liner inverse problems, like involving PDEs, the posterior can only be approximated
  - Variational Bayes provides an elegant optimisation-based for approximating posterior
- References
  - Povala, Kazlauskaite, Febrianto, Cirak, Girolami *Variational Bayesian approximation of inverse problems using sparse precision matrices,* CMAME, 2022
  - Vadeboncoeur, Akyildiz, Kazlauskaite, Girolami, Cirak Fully probabilistic deep models for forward and inverse problems in parametric PDEs*, JCP*, 2023

Today's Schedule



- 09:00–09:50 Bayesian Inverse Problems
  - 10:00–11:00 Hands-on Session
- 11:15–12:10 Gaussian Process Surrogates
  - 13:30–14:30 Hands-on Session
- 14:45–15:35 Statistical Finite Elements
  - 15:45–16:45 Hands-on Session
- 16:45–17:00 Summary & Discussion

Asking questions by typing in Q&A or raising hand both are fine