Bayesian Learning

Basics | Computation | Approximation | Futures

Shakir Mohamed

@shakir_za

DeepMind

Intentions

Have fun | To get to know you | To hear from you | To learn from how you think

Topics in Bayesian Theory

Topics on Ethics and Social Impact

Tricks for Manipulating Probabilities

Raise your interest in a Critical Bayesian approach to Machine Learning

1

Bayesian | Basics

Shakir Mohamed

@shakir_za

Outcomes

1. Concepts in Probability and Bayesian Analysis
2. Separation of Model, Inference and Algorithm
3. Bayesian Applications and Values

Take a minute to think of your answer.

What is Probability?

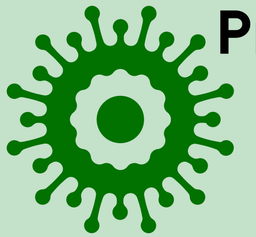
Afterwards, raise hand/unmute/share your answer.
Or write in channel `#Lec_bayesian_inference_mohamed`

Probability

Some Definitions for probability

Statistical Probability
Frequency ratio of items

Logical Probability
Degree of confirmation of a hypothesis based on logical analysis



Probability as Propensity
Probability used for predictions

Subjective Probability
Probability as a degree of belief

Probability is sufficient for the task of reasoning under uncertainty

Probability

Probability as a Degree of Belief

Probability is a measure of the belief in a proposition **given** evidence.
A description of a state of knowledge.

No such thing as **the probability** of an event, since the value depends on the evidence used.

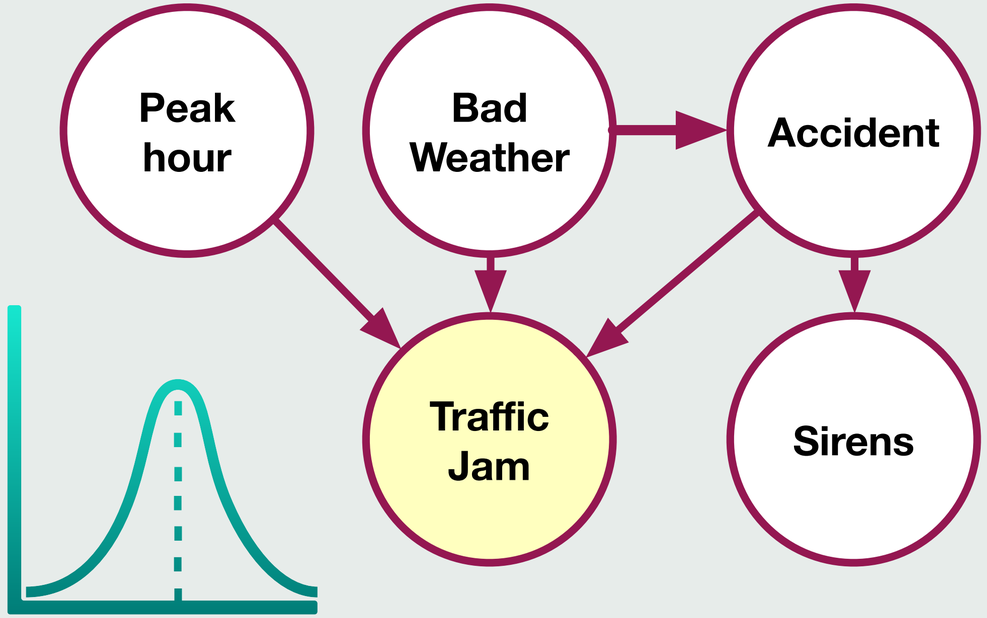
Inherently subjective in that it depends on the believer's information

Different observers with different information will have different beliefs.

Probabilistic Models

Model: Description of the world, of data, of potential scenarios, of processes.

A probabilistic model writes out these models using the language of probability



prob(traffic Jam)

prob(sirens | Accident)

prob(peak hour | Traffic Jam)

Most models in machine learning are probabilistic.

Probabilistic models let you learn probability distributions of data.

You can choose what to learn: Just the mean. Or the entire distribution.

Probabilistic Quantities

Probability $p(\mathbf{x}) \qquad p^*(\mathbf{x}) \qquad q(\mathbf{x})$

Conditions $p(\mathbf{x}) \geq 0 \qquad \int p(\mathbf{x}) d\mathbf{x} = 1$

Bayes Rule $p(\mathbf{z}|\mathbf{x}) = \dfrac{p(\mathbf{x}|\mathbf{z})p(\mathbf{z})}{p(\mathbf{x})}$

Parameterisation $p_\theta(\mathbf{x}|\mathbf{z}) \equiv p(\mathbf{x}|\mathbf{z};\boldsymbol{\theta})$

Expectation $\mathbb{E}_{p_\theta(\mathbf{x}|\mathbf{z})}[f(\mathbf{x};\boldsymbol{\phi})] = \int p_\theta(\mathbf{x}|\mathbf{z}) f(\mathbf{x};\boldsymbol{\phi}) d\mathbf{x}$

Gradient $\nabla_\phi f(\mathbf{x};\boldsymbol{\phi}) = \dfrac{\partial f(\mathbf{x};\boldsymbol{\phi})}{\partial \boldsymbol{\phi}}$

Take a minute to think of your answer.

What is Bayesian Statistics?

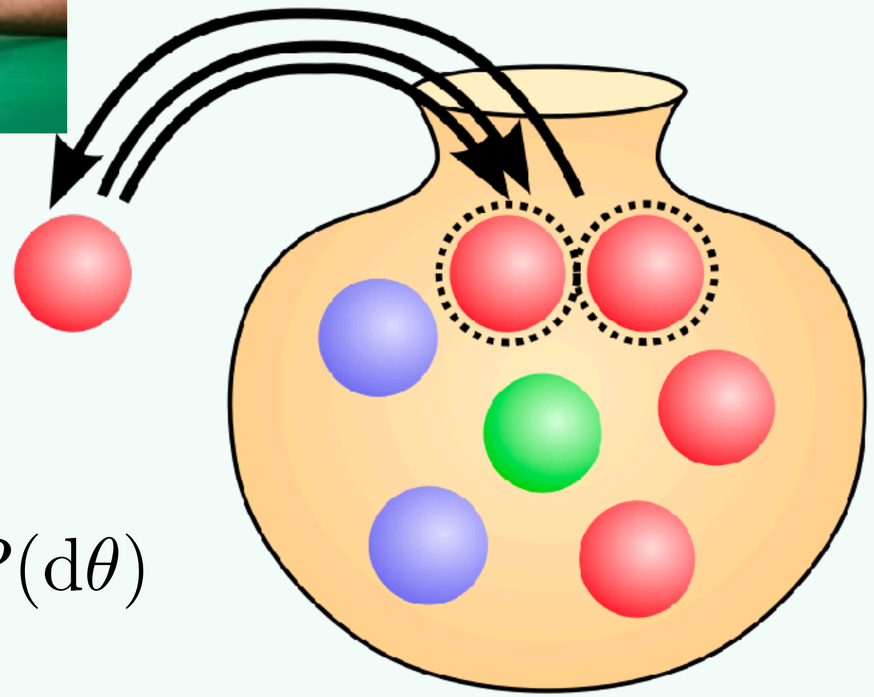
Afterwards, raise hand/unmute/share your answer.
Or write in channel `#Lec_bayesian_inference_mohamed`

Probability of a Sequence

Exchangeable sequence of events

$$p(1,1,1,0,0) = p(1,0,1,0,1)$$

$$p(x_1, \ldots, x_n) = p(x_{\pi_1}, \ldots, x_{\pi_n})$$



For **infinite exchangeability**, the joint probability is invariant to permutation of the indices.

De Finetti's Theorem

$$p(x_1, \ldots, x_N) = \int \prod_{n=1}^{N} p(x_n|\theta) P(\mathrm{d}\theta)$$

Explains why we have parameters and priors, and the power of averaging.

Model-based Bayesian

$$p(x_1, \ldots, x_N) = \int \prod_{n=1}^{N} p(x_n|\theta) P(\mathrm{d}\theta)$$

For an exchangeable sequence of events

Doesn't say anything about the dimension of parameters, so can be infinite.

Do computations and inversions of these quantities using Bayes' Rule

$$p(x_1, \ldots, x_N) = \int \prod_{n=1}^{N} p(x_n|\theta)\, p(\theta)\, \mathrm{d}\theta$$

There is a parameter θ

The data $x_1, \ldots x_n$ is then conditionally independent

There is a likelihood $p(x \mid \theta)$

There is a distribution P on θ, if there's a density then a prior $p(\theta)$

Model-based approach (as opposed to an empirical approach) since we represent the sequence using a parameterised model.

Gives a bridge between Bayesian and Frequentist views of probability.

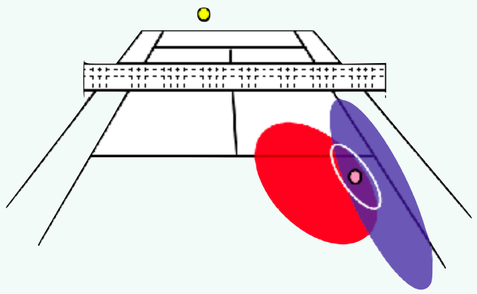
Bayesian Analysis

Bayesian approach follows the idea that all components of a model should be probabilistic and be described by probability distributions

Bayes' Rule

$$\underbrace{p(z|y)}_{\text{Posterior}} = \frac{\overbrace{p(y|z)}^{\text{Likelihood}}\ \overbrace{p(z)}^{\text{Prior}}}{\int p(y,z)dz}$$

Marginal likelihood/ Model evidence



Bayesian analysis is an approach to modelling that follows:

- Decide on a priori beliefs.
- Posit an explanation of how the observed data is generated, i.e. provide a probabilistic description.
- Allows for recursive updating in the light of new evidence.

Rule for inverse probabilities.
Go from prior states of knowledge to new states based on evidence.

Bayesian Analysis

Interested in reasoning about two important quantities

Evidence

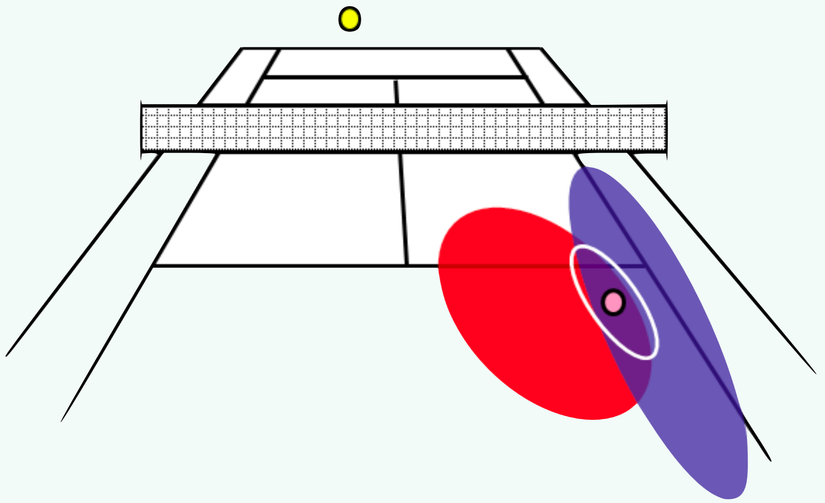
$$p(y|\mathbf{x}) = \int p(y|h(\mathbf{x});\boldsymbol{\theta})p(\boldsymbol{\theta})d\boldsymbol{\theta}$$

Posterior

$$p(\boldsymbol{\theta}|y,\mathbf{x}) \propto p(y|h(\mathbf{x});\boldsymbol{\theta})p(\boldsymbol{\theta})$$

- In Bayesian analysis, things that are *not. observed* must be integrated over - averaged out.
- This makes computation difficult.
- Integration is the central operation.

Intractable Integrals: Will often see this phrasing.
- Don't know the integral in closed form
- Very high-dimensional quantities and can't compute (e.g., using quadrature)



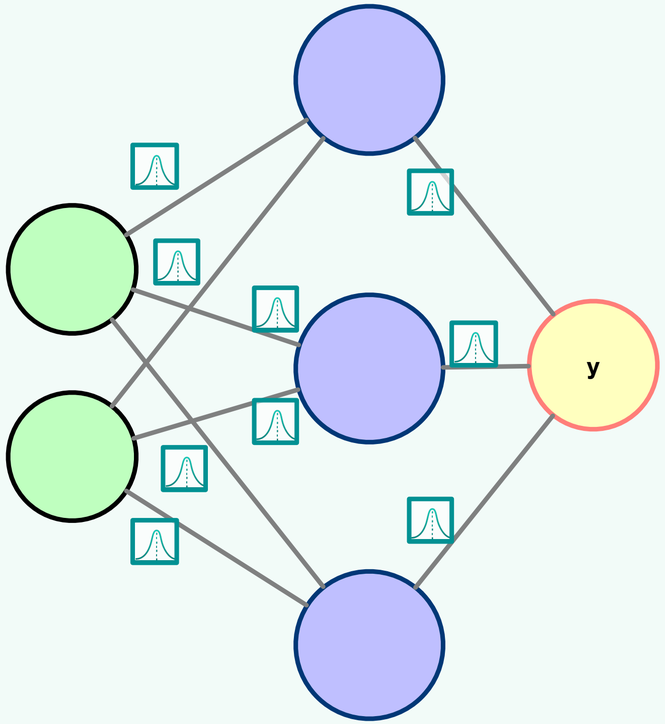
Regression and Classification

Probabilistic models over functions

Prior $p(\boldsymbol{\theta}) = \mathcal{N}(\boldsymbol{\theta}|0, \mathbf{I})$

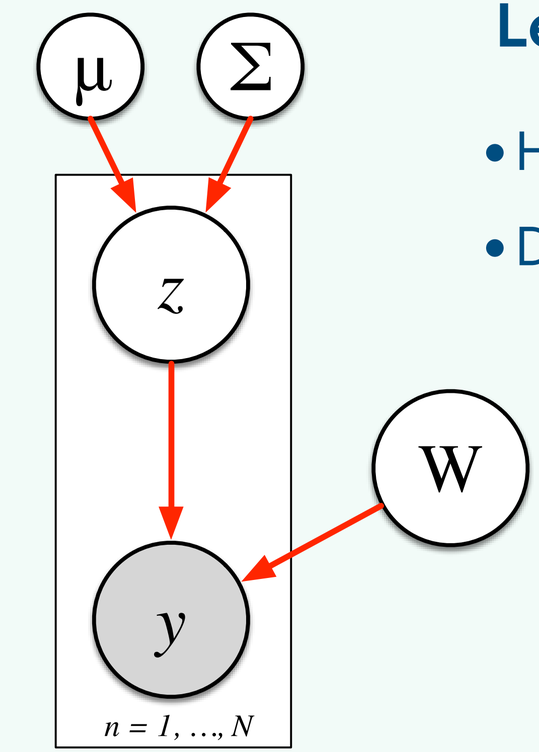
Observation model $p(y|\mathbf{x}, \boldsymbol{\theta}) = Categorical(\pi(\mathbf{x}; \boldsymbol{\theta}))$

Posterior $p(\boldsymbol{\theta}|y, \mathbf{x})$



- Make predictions of future based on past correlations.
- Ways of learning distributions over functions and maintaining uncertainty over functions.
- This allows many types of models, from linear models, deep networks, splines, etc.
- Many ways to learn the posterior distribution.

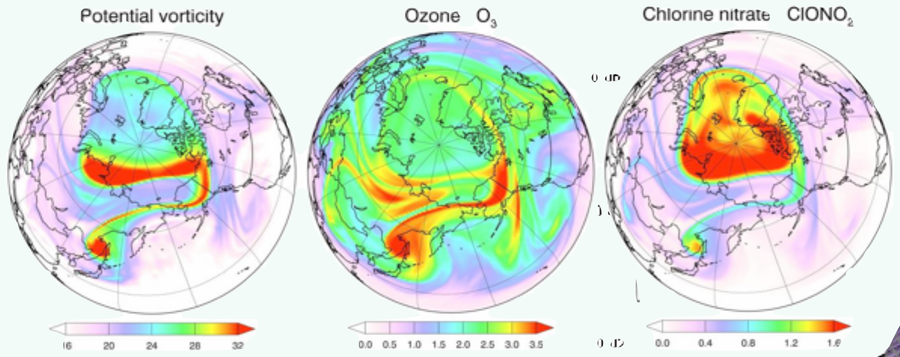
Density Estimation

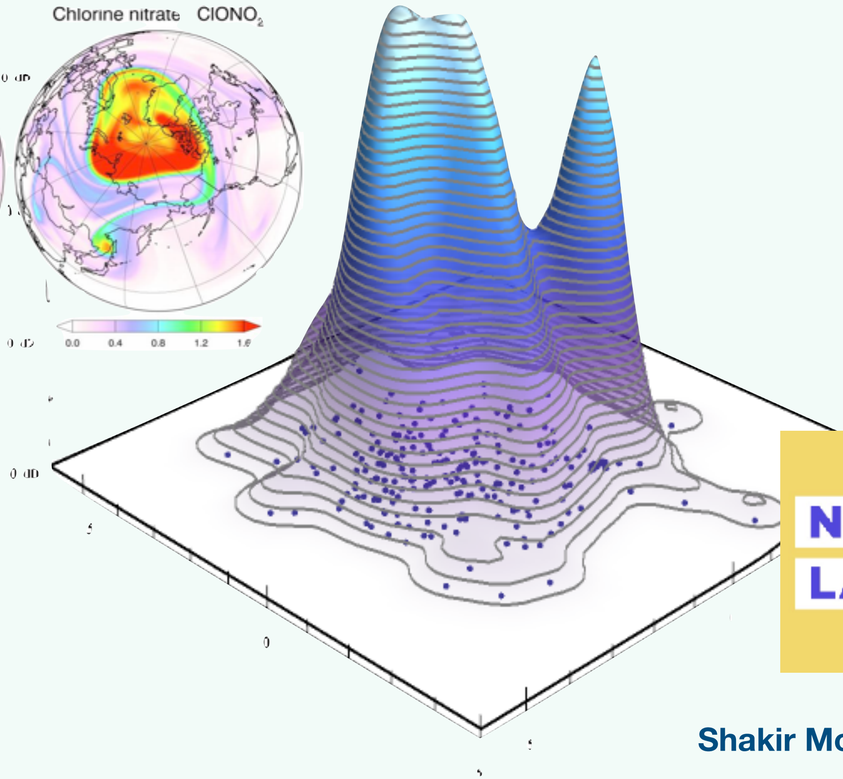


Factor Analysis / PCA

Learn probability distributions over the data itself

- How can you learn from data without any labels. Structure of the data.
- Deep Generative Models and Unsupervised learning.





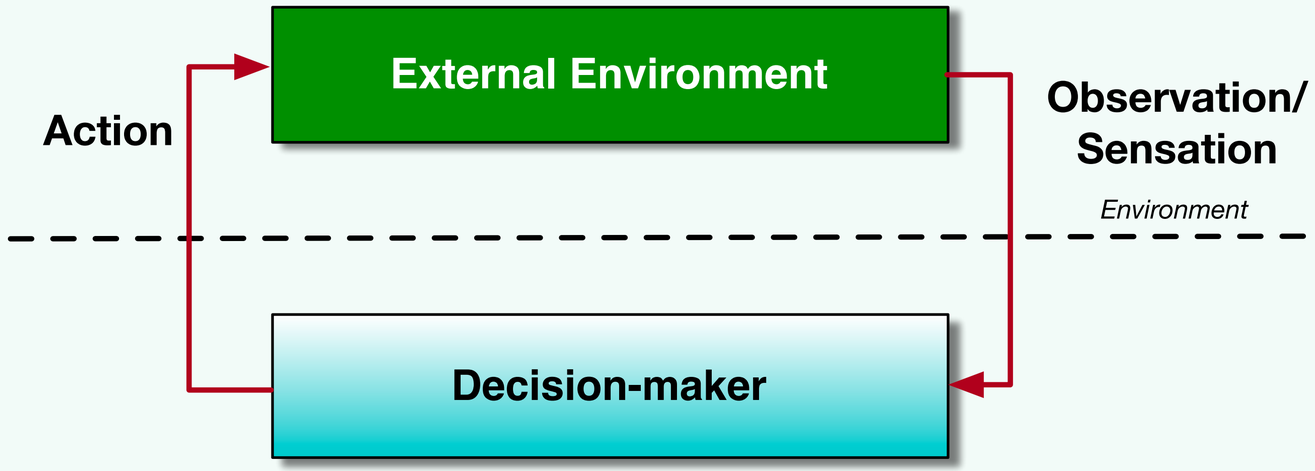
NO LABELS

Prior $z \sim \mathcal{N}(z|\mu, \Sigma)$

Observation model $y \sim \mathcal{N}(y|Wz, \sigma_y^2 I)$

Decision-making

Probabilistic models of environments and actions



Setup is common in experimental design, causal learning, reinforcement learning.

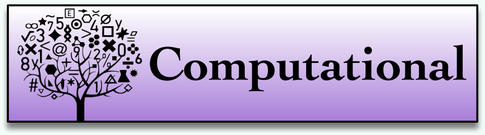
Prior over actions $a \sim p(a)$

Interaction only $u(s, a) \sim \text{Environment}(a)$

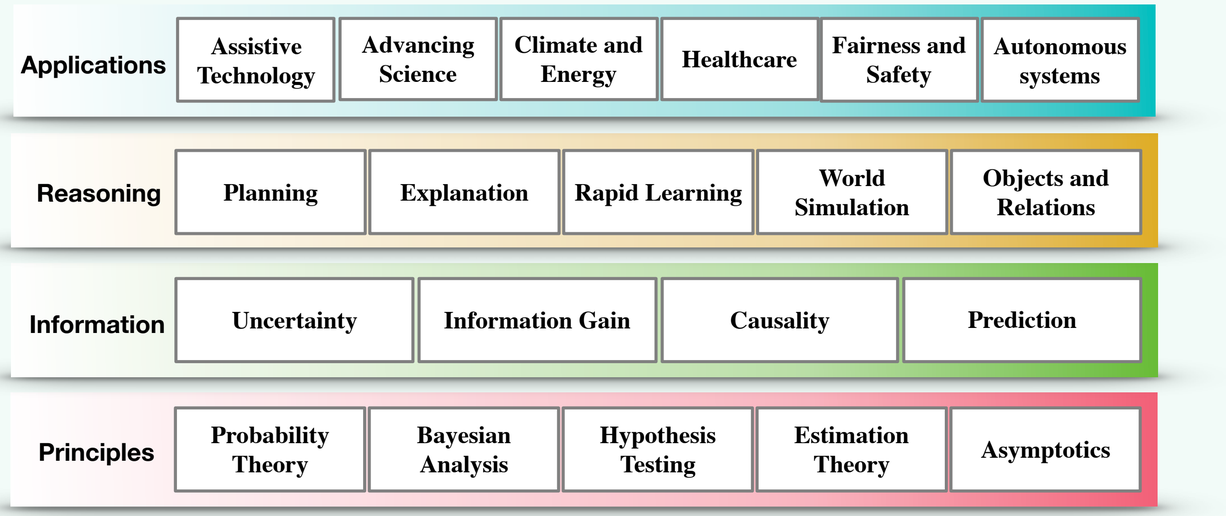
Reward/Utility $p(R(s)|a) \propto \exp(u(s, a))$

A Poetics of Machine Learning

Bayesian approach is also an Interpretive Approach to Modelling and Data Analysis

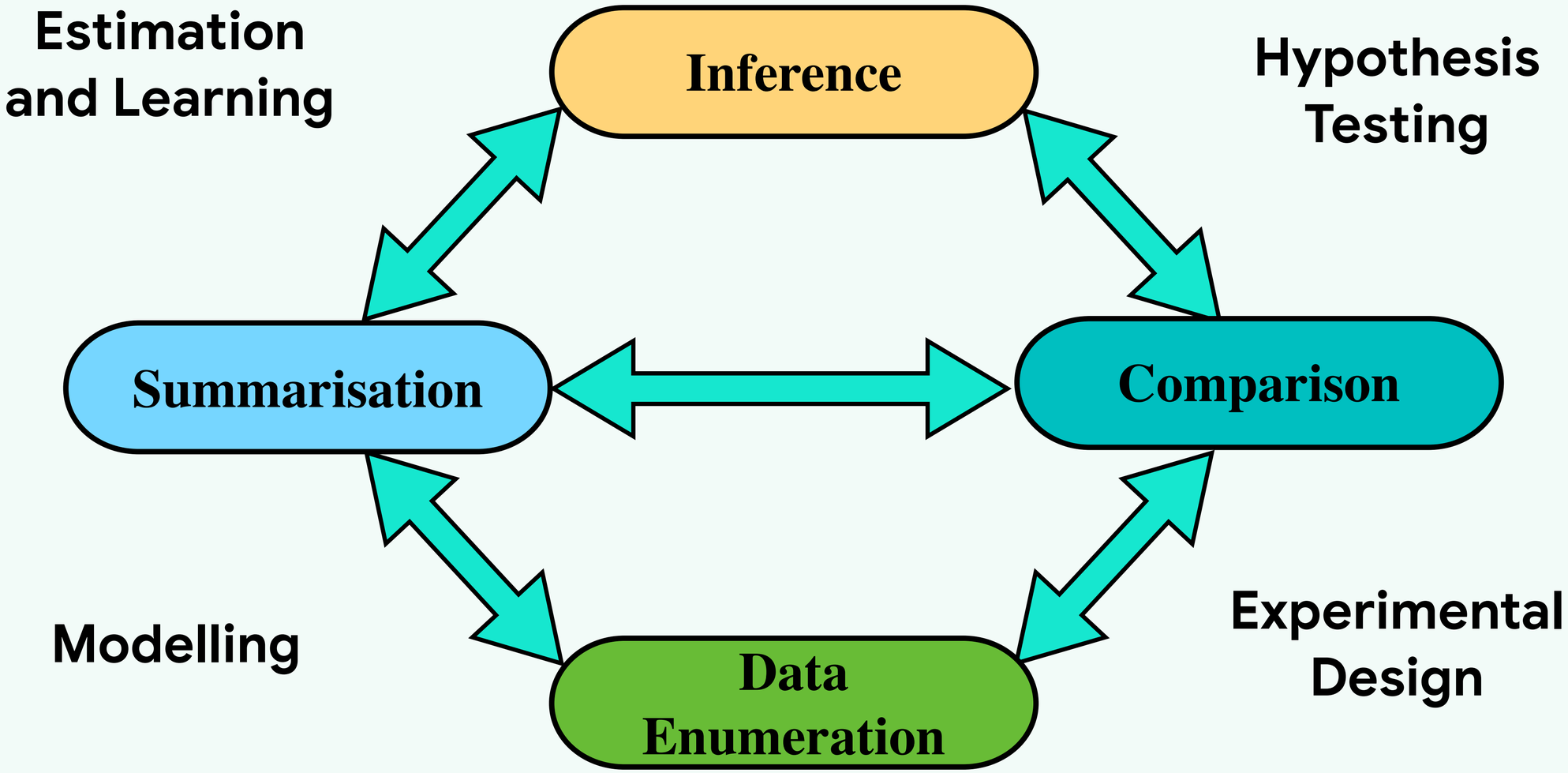


Marr's Levels of Analysis

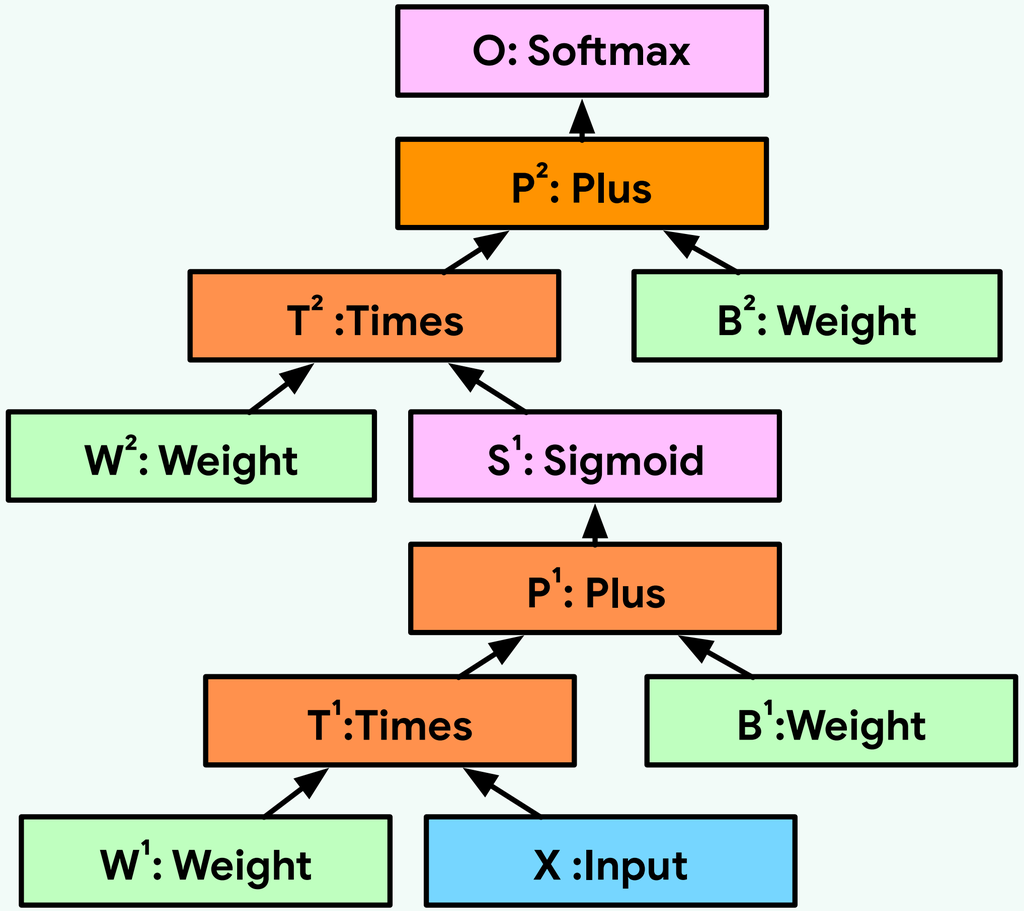


- We exist within and across many interpretive communities in ML.
- Interpretive frameworks we use not only describes our work, but also produce the work it describes.
- Become aware of the interpretive frames we use.

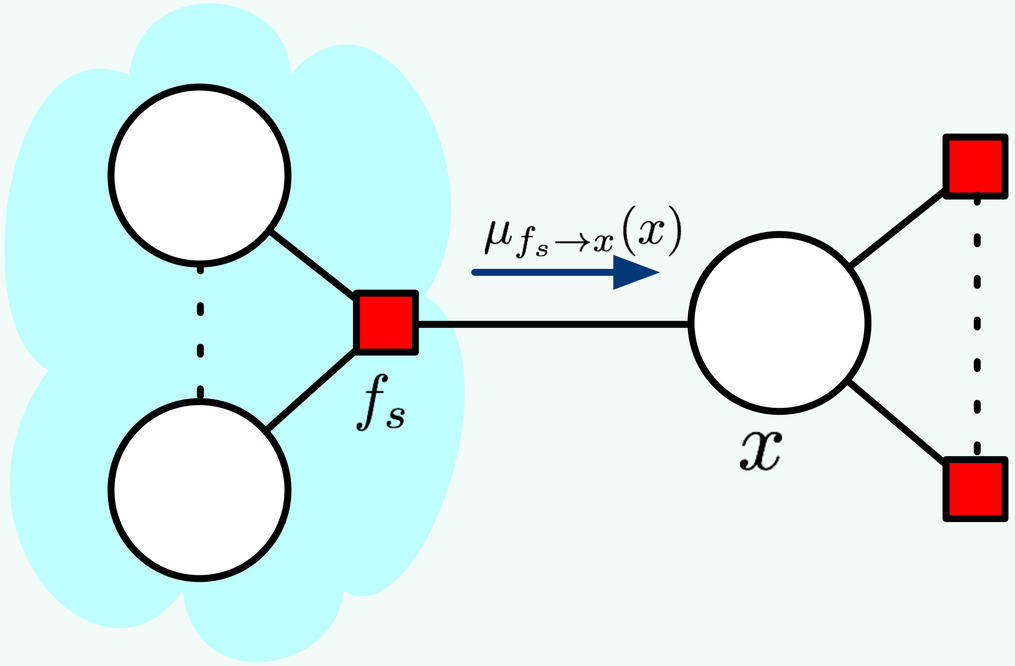
Statistical Operations



Architecture and Loss

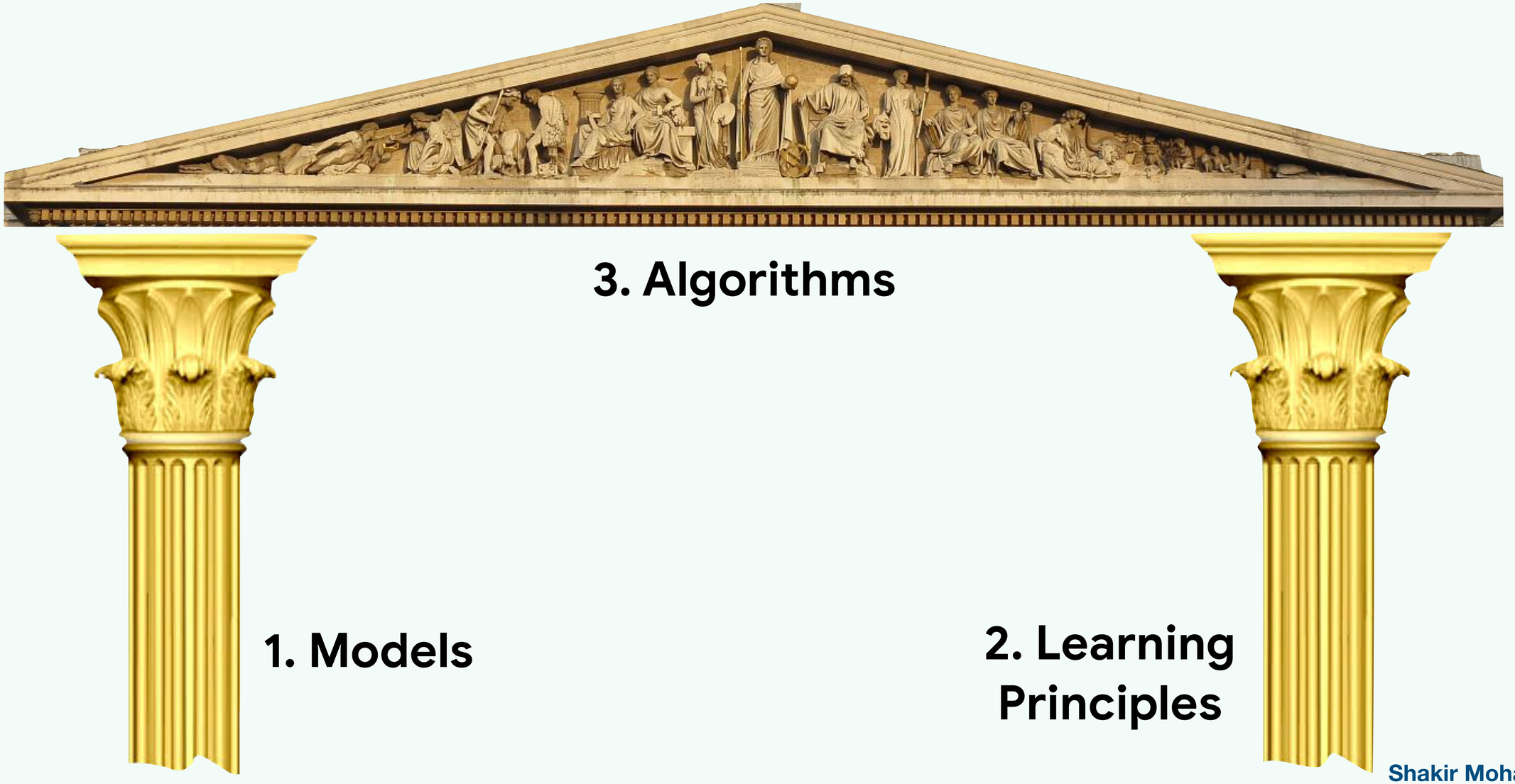


1. Computational Graphs



2. Error propagation

Model-Inference-Algorithm



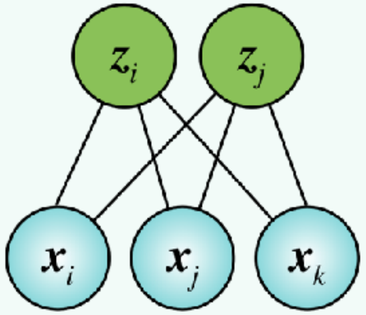
3. Algorithms

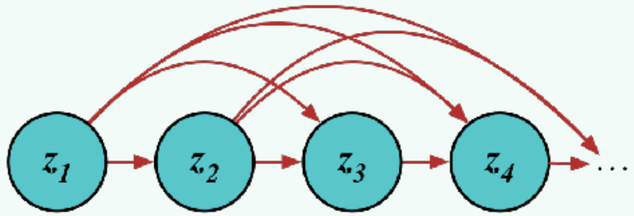
1. Models

2. Learning Principles

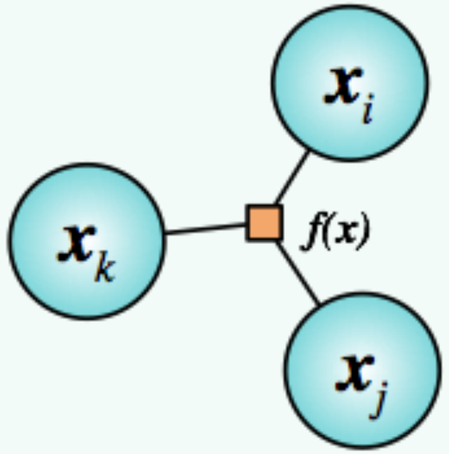
Models

Directed and Undirected

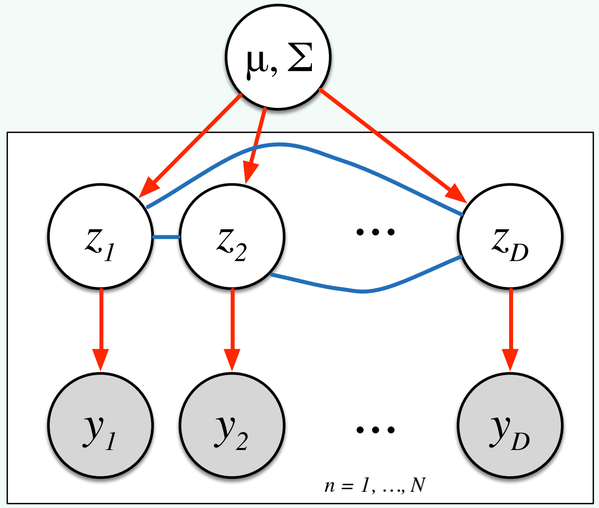




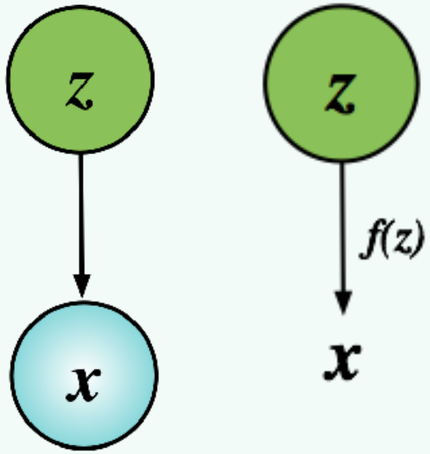
Fully-observed



Parametric, Non-parametric And semi-parametric



Latent Variable



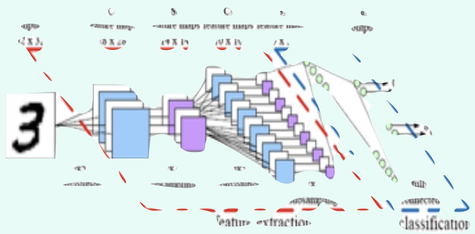
Learning Principles

- **Statistical Inference**
  - **Direct**
    - Laplace approximation
    - Maximum a posteriori
    - Cavity Methods
    - Expectation Maximisation
    - Noise Contrastive
    - Maximum Likelihood
    - Variational Inference
    - Integr. Nested Laplace Approx
    - Markov chain Monte Carlo
    - Sequential Monte Carlo
  - **Indirect**
    - Two Sample Comparison
    - Approx Bayesian Computation
    - Max Mean Discrepency
    - Method of Moments
    - Transportation methods

Algorithms

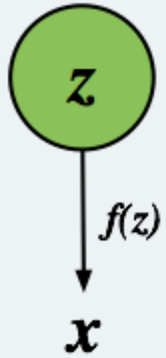
A given model and learning principle can be implemented in many ways.

Convolutional neural network + penalised maximum likelihood



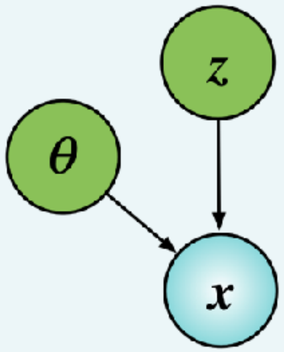
- Optimisation methods (SGD, Adagrad)
- Regularisation (L1, L2, batchnorm, dropout)

Implicit Generative Model + Two-sample testing



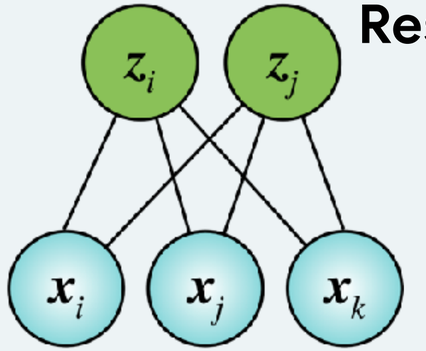
- Unsupervised-as-supervised learning
- Approximate Bayesian Computation (ABC)
- Generative adversarial network (GAN)

Latent variable model + variational inference



- VEM algorithm
- Expectation propagation
- Approximate message passing
- Variational auto-encoders (VAE)

Restricted Boltzmann Machine + maximum likelihood



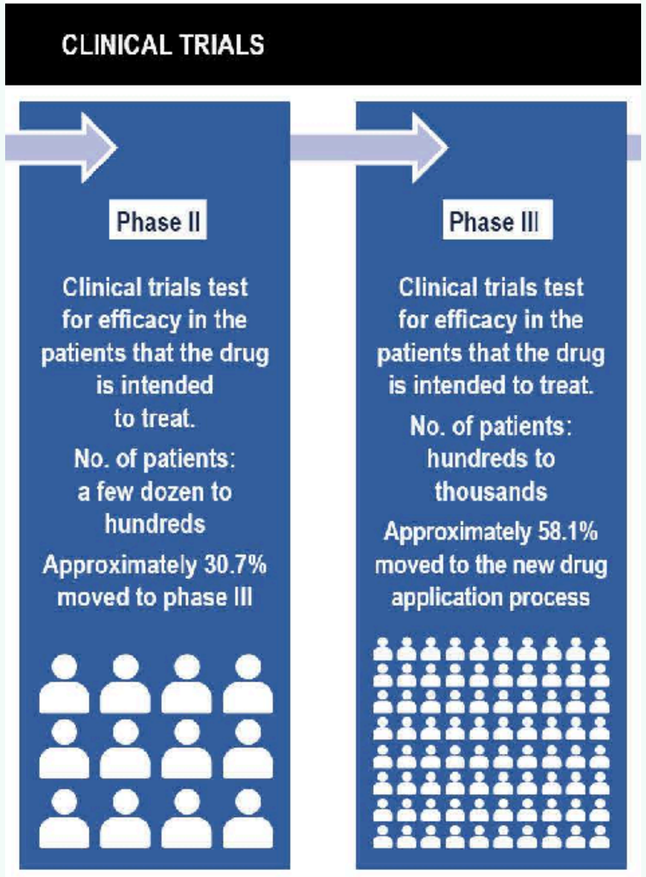
- Contrastive Divergence
- Persistent CD
- Parallel Tempering
- Natural gradients

Bayesian Applications

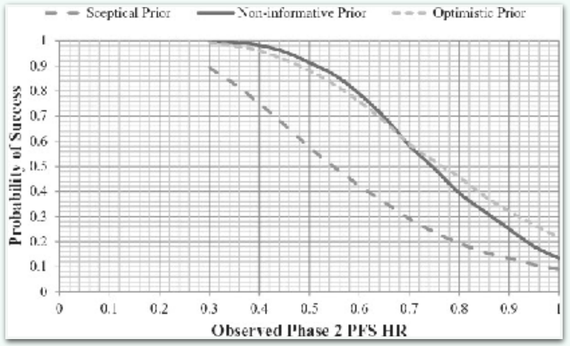
- Building models
- Choice of priors and where they come from.
- Including prior knowledge
- Doing Bayesian computations
- Reporting uncertainties
- Accounting for noise and stationarity
- Loss, Risk and Utility
- Uncertainty and Calibration
- Evaluating and comparing models.
- Adapting to different amounts of evidence

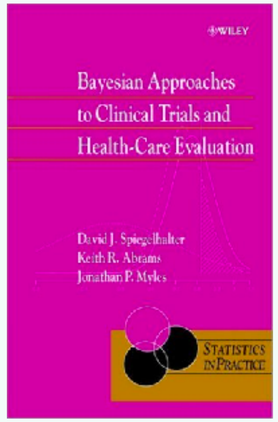
Continuing Clinical Trials

Informing the decision to move from Stage 2 to 3 trials



- Sensitive decision since it requires considerable commitments.
- Inform the decision with predictions of the **probability of success** at the Stage 3 testing.
- Simplest models use Bayesian Linear regression.





Bayesian Considerations: Choice of models, priors, how to compute posterior distributions, how to report variability in predictions.

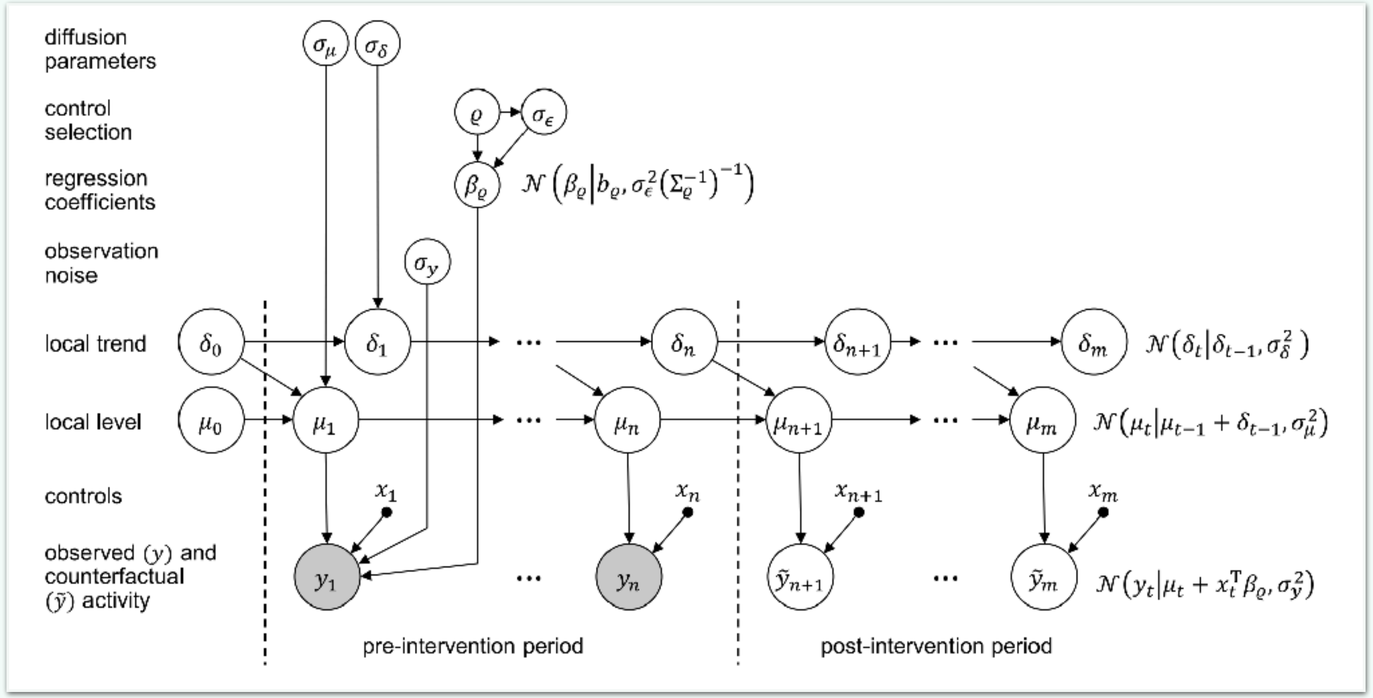
Causal Inference in Structural Time Series

Estimate the causal effect of a designed intervention on a time series.

Clicks generated during an ad campaign, deterioration in clinical interventions. Posterior predictive density over counterfactual responses to understand causal impact of an intervention.

Structural time series include:

- Latent Gaussian models (Kalman Filters),
- state-space models, stochastic RNNs,
- (non-)linear temporal models, ARIMA
- autonomous and non-autonomous systems,
- semi-MDPs and POMDPs.

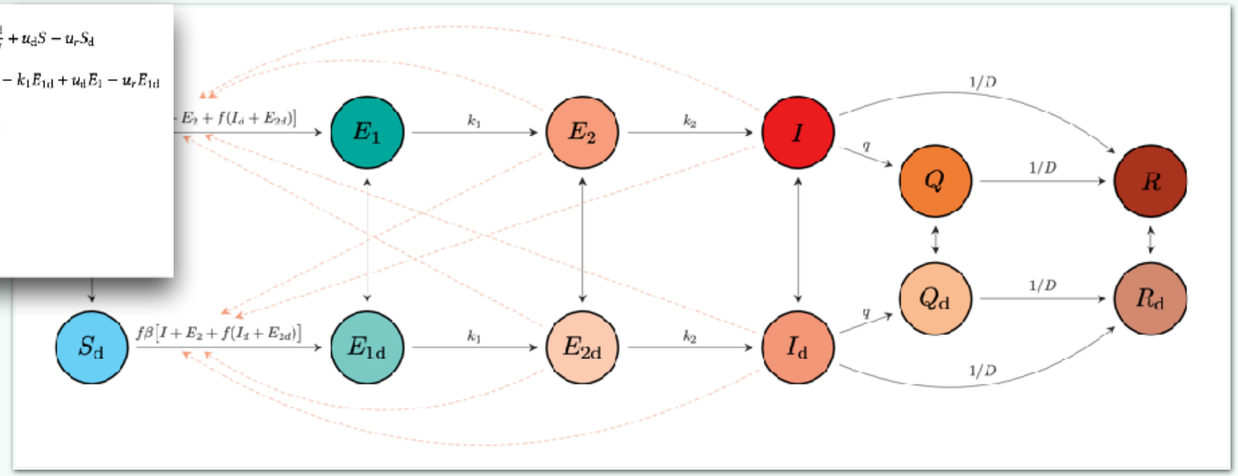


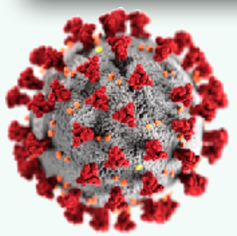
Bayesian Considerations: Understanding of a system at different scales, allowing other sources of knowledge to be included in the model. Will require complex methods for computing posterior distributions efficiently.

Infectious Disease Modelling

Bayesian model of physical distancing for COVID-19

$$\frac{dS_d}{dt} = -f\beta\left[I + E_2 + f(I_d + E_{2d})\right]\frac{S_d}{N} + u_d S - u_r S_d$$
$$\frac{dE_{1d}}{dt} = f\beta\left[I + E_2 + f(I_d + E_{2d})\right]\frac{S_d}{N} - k_1 E_{1d} + u_d E_1 - u_r E_{1d}$$
$$\frac{dE_{2d}}{dt} = k_1 E_{1d} - k_2 E_{2d} + u_d E_2 - u_r E_{2d}$$
$$\frac{dI_d}{dt} = k_2 E_{2d} - qI_d - \frac{I_d}{D} + u_d I - u_r I_d$$
$$\frac{dQ_d}{dt} = qI_d - \frac{Q_d}{D} + u_d Q - u_r Q_d$$
$$\frac{dR_d}{dt} = \frac{I_d}{D} + \frac{Q_d}{D} + u_d R - u_r R_d.$$





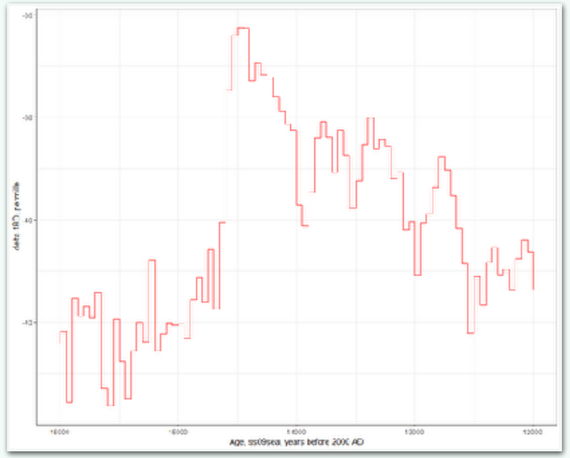
Have **simulators of a process**, like the spread of a disease, and no likelihood models. We would still want to understand variability of models and report uncertainty.

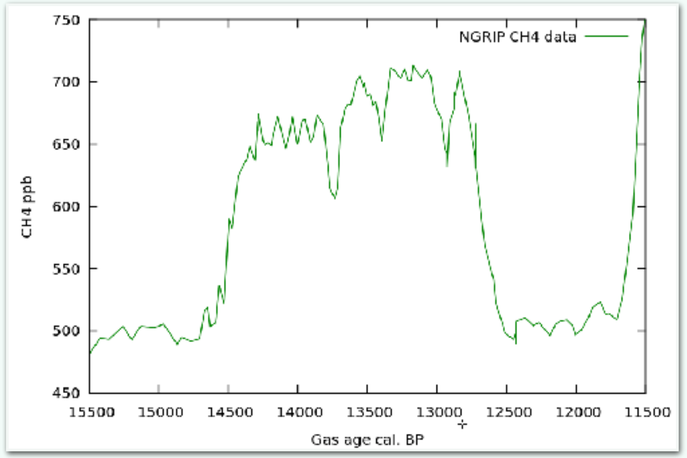
Bayesian Considerations: intractable likelihoods, simulation-based approaches for Bayesian analysis, need for expert definitions of similarity, efficient use of evidence.

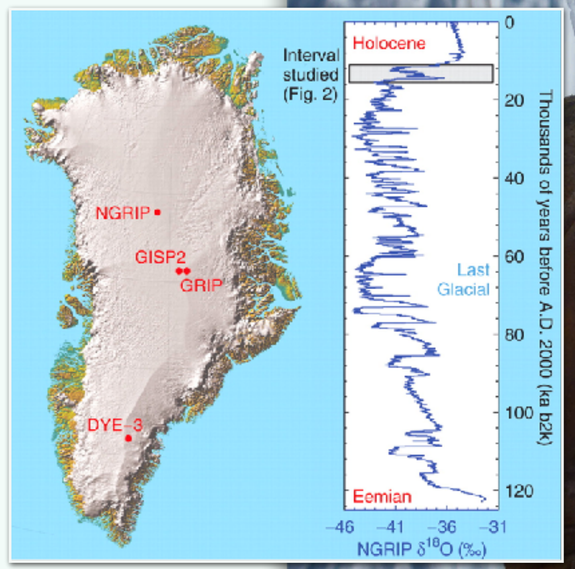
Layer-counting in Ice Cores

Estimating historical time from ice cores

Concentration of particles, chemicals and gasses in ice cores are a continuous record of climatic and environmental information; used widely in climate change analysis.



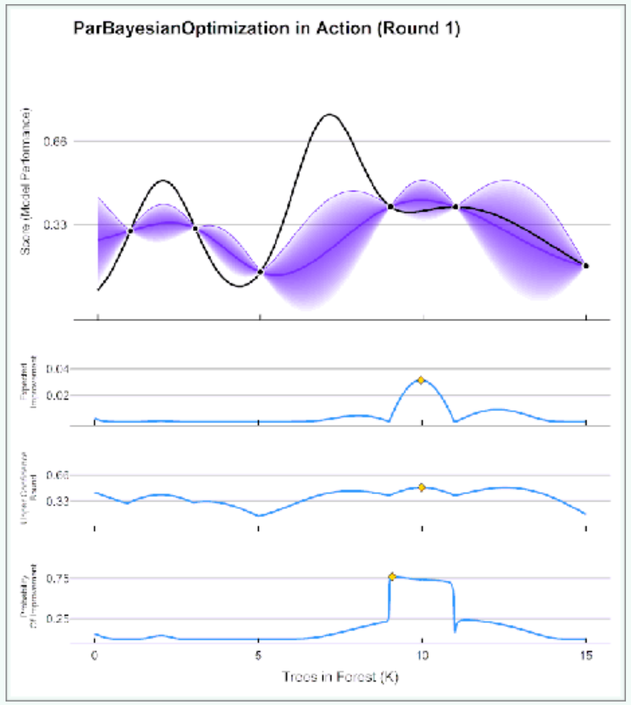




Bayesian Considerations: Model that accounts for periodicity of the sedimentation process, account for lack of stationarity, noise in the observations.

Experimental Design

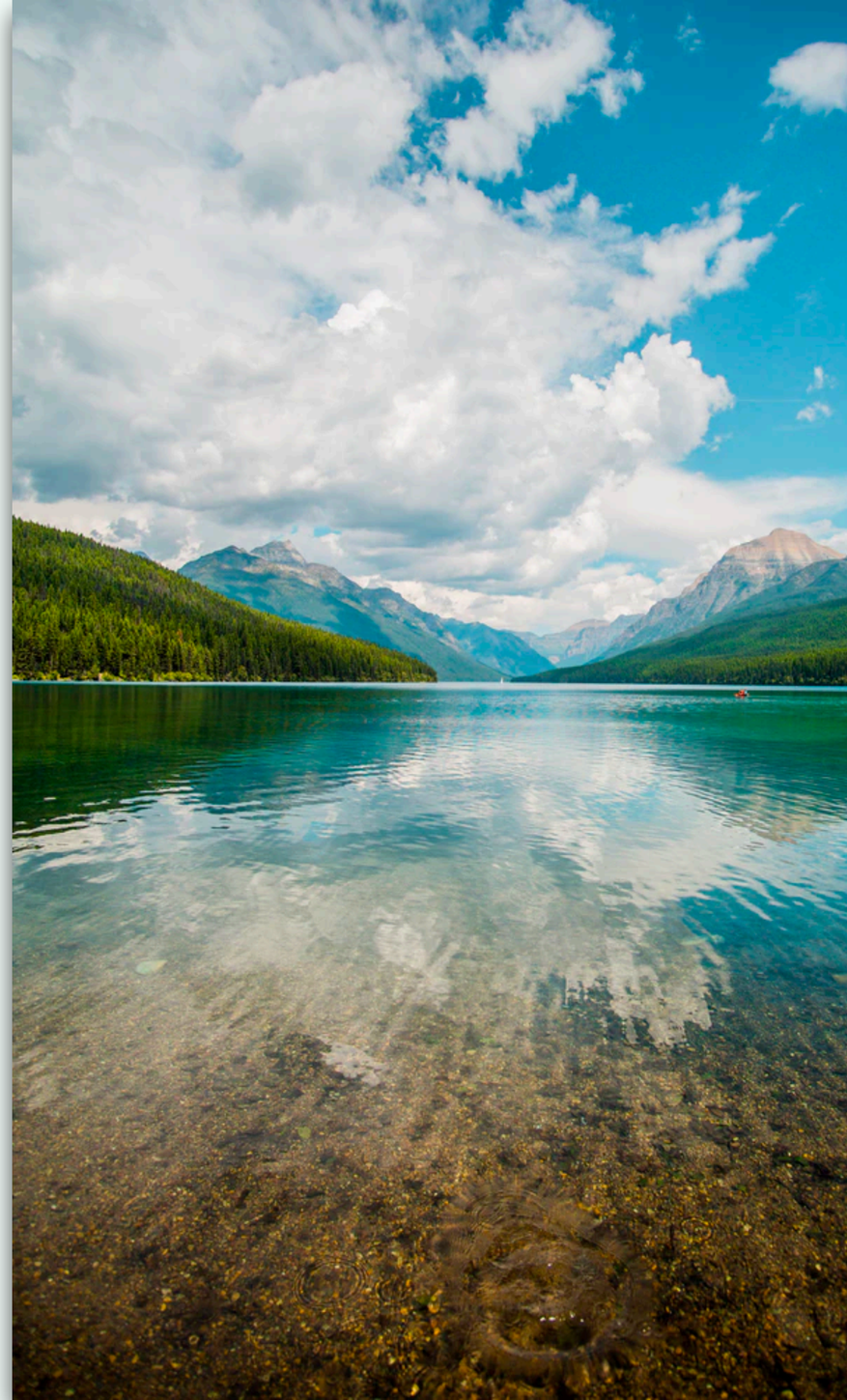
Efficiently placing sensors across an area



We can use estimation of uncertainty from a posterior distribution, to search through a parameter space for the most informative test points.

This comes up in the placement of sensors in a system, or in the search for parameters.

Bayesian Considerations: importance of uncertainty estimation and calibration, Risk minimisation and utility functions.



Contextual Values

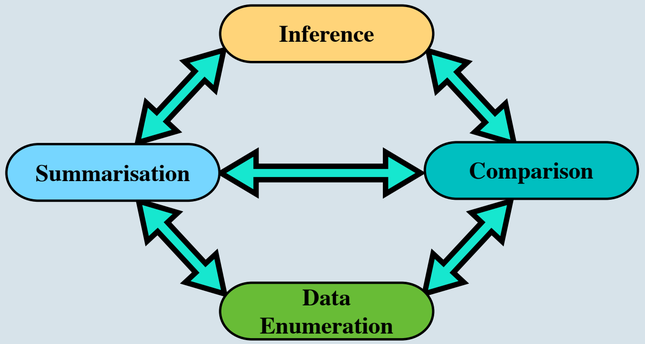
Social

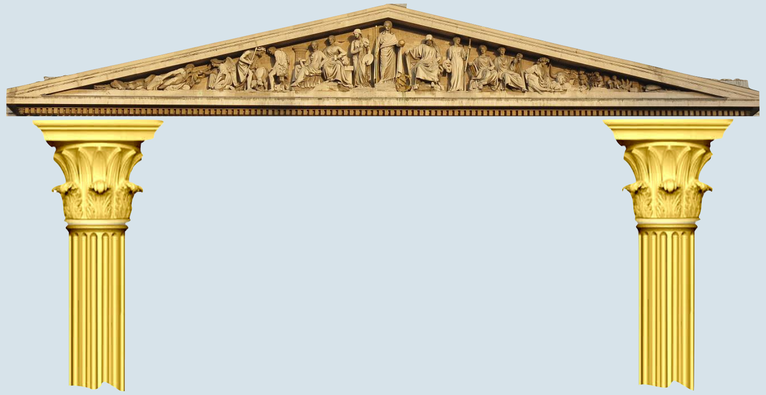
Cultural

Political

Economic

Epistemic concerns and values





Computational

Algorithmic

Implementation

Marr's Levels of Analysis

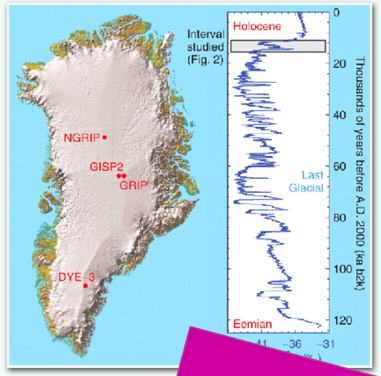
Applications	Assistive Technology	Advancing Science	Climate and Energy	Healthcare	Fairness and Safety	Autonomous systems
Reasoning	Planning	Explanation	Rapid Learning	World Simulation	Objects and Relations	
Information	Uncertainty	Information Gain	Causality	Prediction		
Principles	Probability Theory	Bayesian Analysis	Hypothesis Testing	Estimation Theory	Asymptotics	

Statistical Probability

Frequency ratio of items

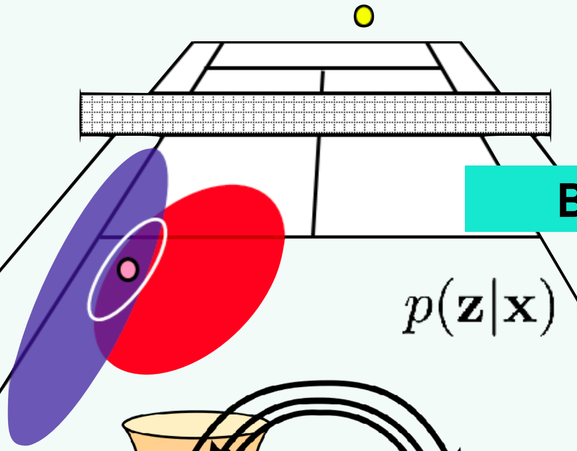
Subjective Probability

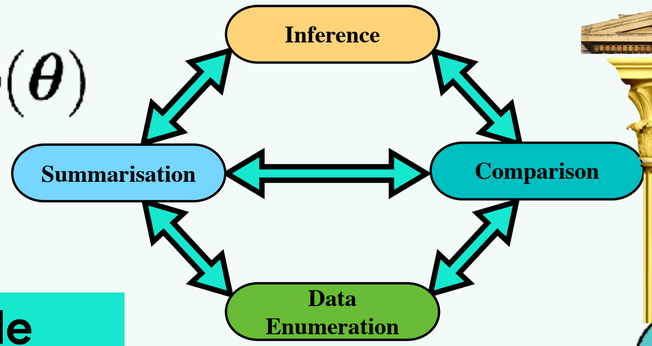
Probability as a degree of belief

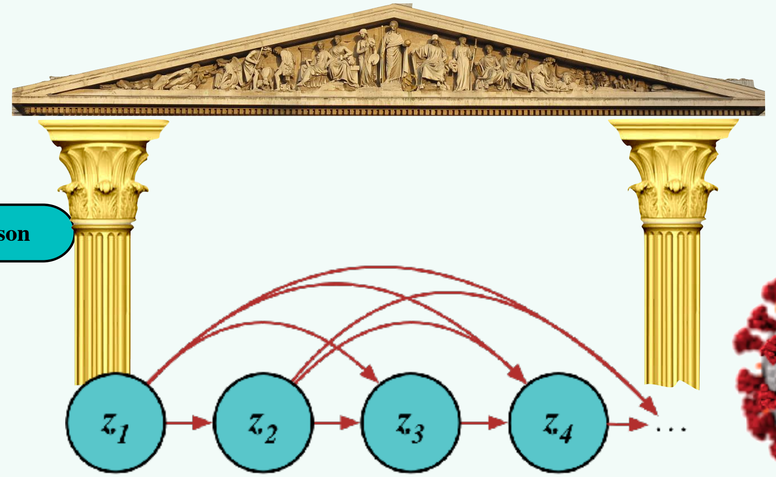


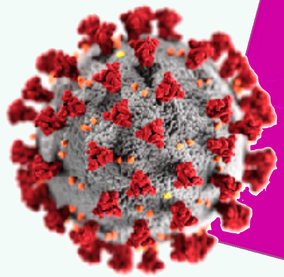
Posterior

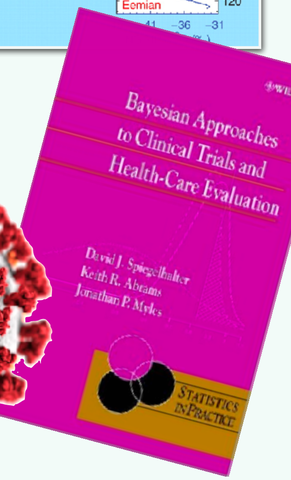
$$p(\boldsymbol{\theta}|y,\mathbf{x}) \propto p(y|h(\mathbf{x});\boldsymbol{\theta})p(\boldsymbol{\theta})$$





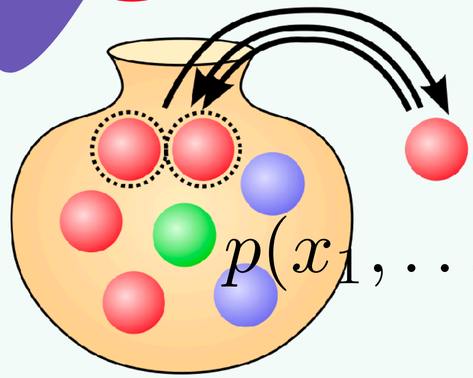


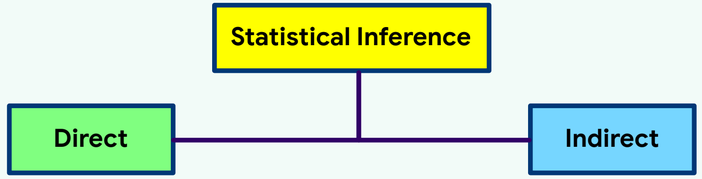




Bayes Rule

$$p(\mathbf{z}|\mathbf{x}) = \frac{p(\mathbf{x}|\mathbf{z})p(\mathbf{z})}{p(\mathbf{x})}$$





$$p(x_1,\ldots,x_N) = \int \prod_{n=1}^{N} p(x_n|\theta)P(\mathrm{d}\theta)$$

Some papers and books

Cheeseman, P.C., 1985, August. In Defense of Probability. In IJCAI (Vol. 2, pp. 1002-1009).

Jaynes, Edwin T. Probability theory: The logic of science. Cambridge university press, 2003.

Aldous, David J. "Exchangeability and related topics." École d'Été de Probabilités de Saint-Flour XIII—1983. Springer, Berlin, Heidelberg, 1985. 1-198.

Bishop, Christopher M. Pattern recognition and machine learning. springer, 2006.

Barber, David. Bayesian reasoning and machine learning. Cambridge University Press, 2012.

Efron, Bradley. "Maximum likelihood and decision theory." The annals of Statistics (1982): 340-356.

Sabin, Tony, et al. "A quantitative process for enhancing end of phase 2 decisions." Statistics in Biopharmaceutical Research 6.1 (2014): 67-77.

Brodersen, Kay H., et al. "Inferring causal impact using Bayesian structural time-series models." The Annals of Applied Statistics 9.1 (2015): 247-274.

Minter, Amanda, and Renata Retkute. "Approximate Bayesian Computation for infectious disease modelling." Epidemics 29 (2019): 100368.

Anderson, Sean C., et al. "Estimating the impact of COVID-19 control measures using a Bayesian model of physical distancing." medRxiv (2020).

Wheatley, J. J., et al. "Bayesian layer counting in ice-cores: Reconstructing the time scale." The Contribution of Young Researchers to Bayesian Statistics. Springer, Cham, 2014. 121-125.

Krause, Andreas, Ajit Singh, and Carlos Guestrin. "Near-optimal sensor placements in Gaussian processes: Theory, efficient algorithms and empirical studies." Journal of Machine Learning Research 9.Feb (2008): 235-284.

Nissenbaum, Helen. "How computer systems embody values." Computer 34.3 (2001): 120-119.

10mins Break

Bayesian | Computation

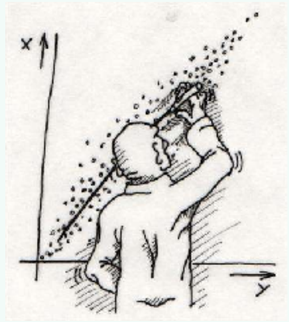
Shakir Mohamed

@shakir_za

Outcomes

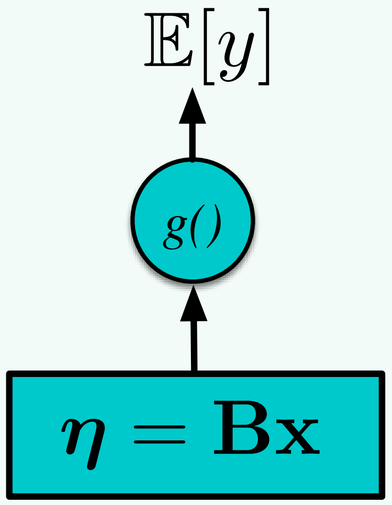
1. Probabilistic Models and Priors
2. Likelihood, Marginalisation, Prediction
3. Inference and Testing

Linear Regression



$$\eta = \mathbf{w}^\top \mathbf{x} + b$$

$$p(y|\mathbf{x}) = p(y|g(\eta);\theta)$$



- The basic function can be any linear function, e.g., affine, convolution.
- g(.) is an **inverse link function** that we'll refer to as an activation function.

Target	Regression	Link	Inv link	Activation
Real	Linear	Identity	Identity	
Binary	Logistic	Logit $\log \frac{\mu}{1-\mu}$	Sigmoid $\frac{1}{1+\exp(-\eta)}$	Sigmoid
Binary	Probit	Inv Gauss CDF $\Phi^{-1}(\mu)$	Gauss CDF $\Phi(\eta)$	Probit
Binary	Gumbel	Compl. log-log $\log(-\log(\mu))$	Gumbel CDF $e^{-e^{-x}}$	
Binary	Logistic		Hyperbolic Tangent $\tanh(\eta)$	Tanh
Categorical	Multinomial		Multin. Logit $\frac{\eta_i}{\sum_j \eta_j}$	Softmax
Counts	Poisson	$\log(\mu)$	$\exp(\nu)$	
Counts	Poisson	$\sqrt{(}\mu)$	ν^2	
Non-neg.	Gamma	Reciprocal $\frac{1}{\mu}$	$\frac{1}{\nu}$	
Sparse	Tobit		max $\max(0;\nu)$	ReLU
Ordered	Ordinal		Cum. Logit $\sigma(\phi_k - \eta)$	

Optimise the negative log-likelihood

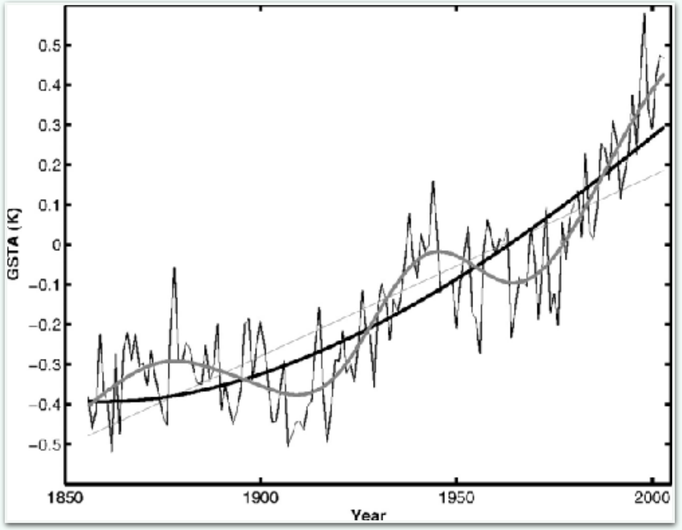
$$\mathcal{L} = -\log p(y|g(\eta);\theta)$$

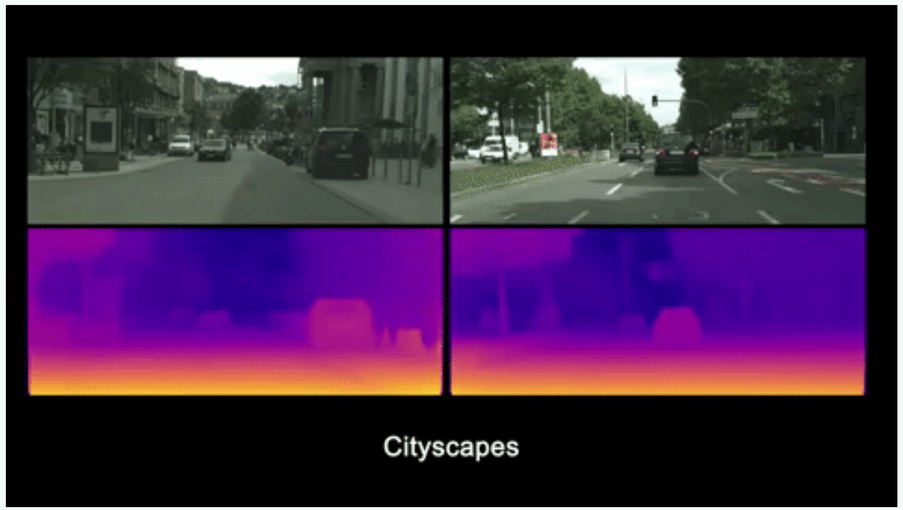
Deep Networks

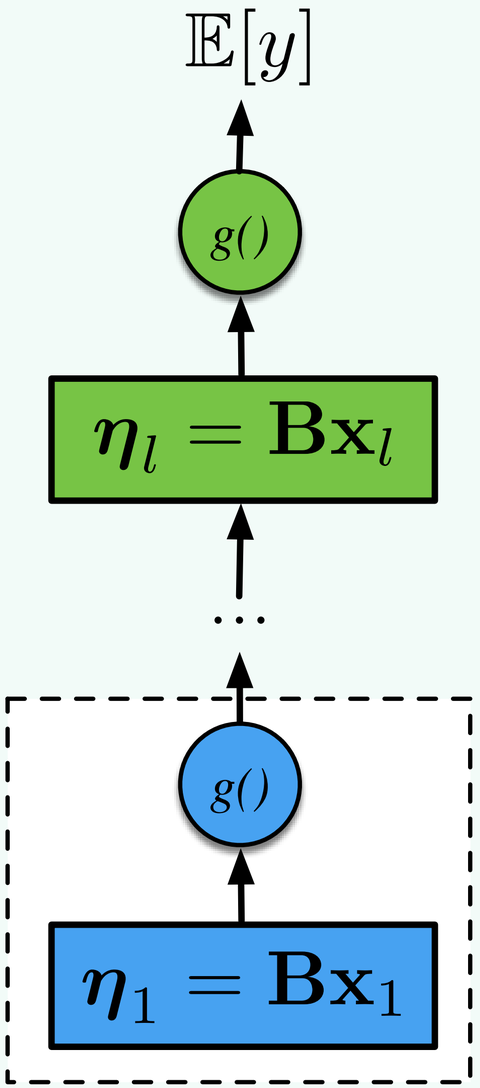
- Recursively compose the basic linear functions.
- Gives a deep neural network.

$$\mathbb{E}[y] = h_L \circ \ldots \circ h_l \circ h_0(\mathbf{x})$$

A general, flexible framework for building
non-linear, parametric models



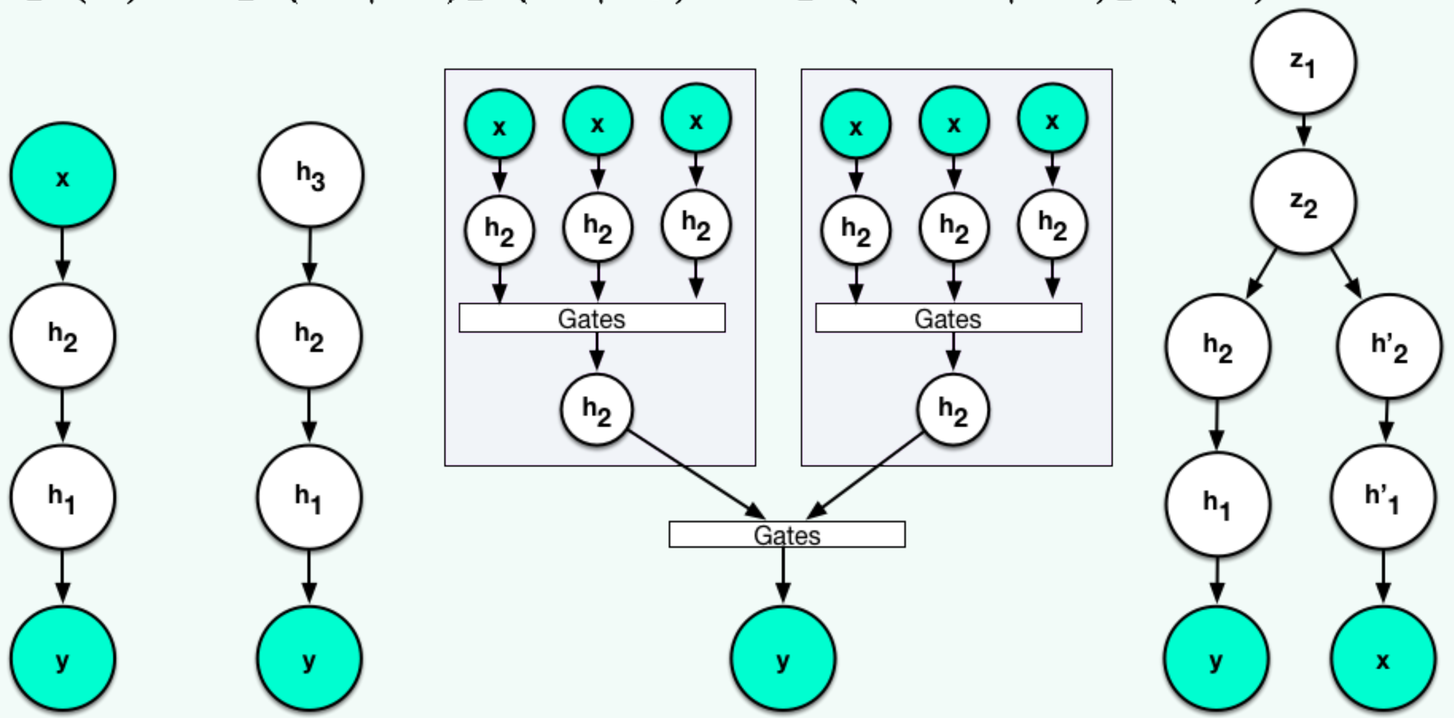




Deep and Hierarchical

Hierarchical Model: models where the (prior) probability distributions can be decomposed into a sequence of conditional distributions

$$p(z) = p(z_1|z_2)p(z_2|z_3)\dots p(z_{L-1}|z_L)p(z_L)$$



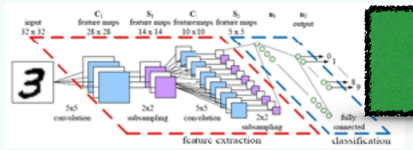
Deep feed-forward regression

Deep directed generative model

Hierarchical Mixture of Experts

Deep Multi-view model/ Information Bottleneck

Two Streams of ML



Deep Learning

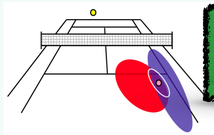
+ Rich non-linear models for classification and sequence prediction.

+ Scalable learning using stochastic approximation and conceptually simple.

+ Easily composable with other gradient-based methods

- Only point estimates

- Hard to score models, do selection and complexity penalisation.



Bayesian Reasoning

- Mainly conjugate and linear models

- Potentially intractable inference, computationally expensive or long simulation time.

+ Unified framework for model building, inference, prediction and decision making

+ Explicit accounting for uncertainty and variability of outcomes

+ Robust to overfitting; tools for model selection and composition.

Natural to consider the combination of these approaches: Bayesian Deep Learning

Take a minute to think of your answer.

What is a Likelihood?

Afterwards, raise hand/unmute/share your answer.
Or write in channel `#Lec_bayesian_inference_mohamed`

Likelihood Functions

Probabilistic Model

$$p(y|\mathbf{x}) = p(y|h(\mathbf{x});\boldsymbol{\theta})$$

Likelihood function

$$\mathcal{L}(\boldsymbol{\theta}) = \sum_n \log p(y_n|\mathbf{x}_n;\boldsymbol{\theta})$$

Likelihood of parameters

Prescribed Likelihoods

Efficient Estimators
- Statistically efficient (Cramer-Rao lower bound)
- Asymptotically unbiased, consistent
- Maximum entropy (principle of indifference)

Widely-applicable
- Handle data that is incompletely observed, distorted, samples with bias
- Can offset or correct these issues.

Tests with Good Power
- Likelihood ratio tests
- Can construct small confidence regions

Pool Information
- Combine different data sources
- Knowledge outside the data can be used, like constraints on domain or prior probabilities.

Misspecification: Inefficient estimates; or confidence intervals/tests can fail completely.

Estimation Theory

Probabilistic Model

$$p(y|\mathbf{x}) = p(y|h(\mathbf{x});\boldsymbol{\theta})$$

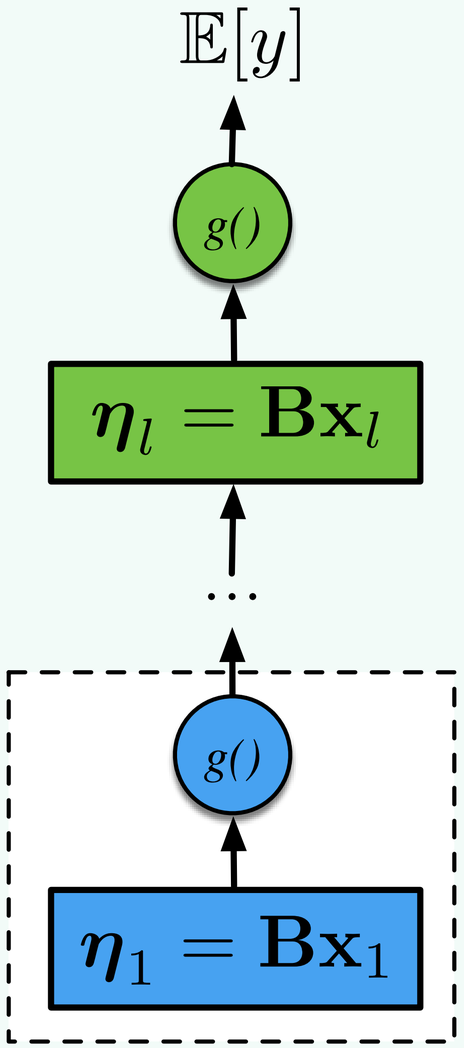
Likelihood function

$$\mathcal{L}(\boldsymbol{\theta}) = \sum_n \log p(y_n|\mathbf{x}_n;\boldsymbol{\theta})$$

Maximum Likelihood

Optimisation Objective

$$\arg\max_{\theta} \mathcal{L}(\boldsymbol{\theta})$$



- Straightforward and natural way to learn parameters
- Can be biased in finite sample size, e.g., Gaussian variances with N and N-1.
- Easy to observe **overfitting** of parameters.

Estimation Theory

Probabilistic Model

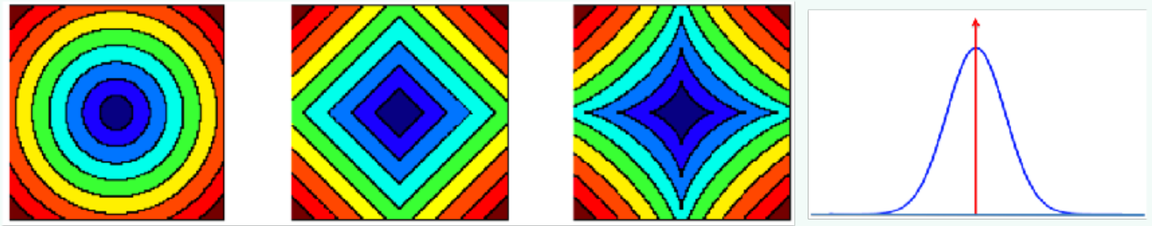
$$p(\boldsymbol{\theta}|y,\mathbf{x}) \propto p(y|h(\mathbf{x});\boldsymbol{\theta})p(\boldsymbol{\theta})$$

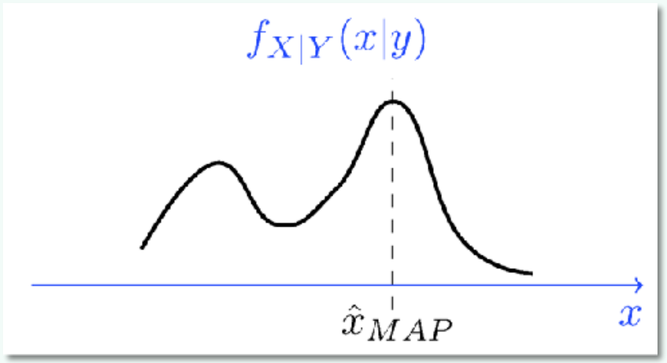
Likelihood function

$$\mathcal{L}(\boldsymbol{\theta}) = \sum_n \log p(y_n|\mathbf{x}_n;\boldsymbol{\theta}) + \tfrac{1}{\lambda}\mathcal{R}(\boldsymbol{\theta}))$$

✦ **Regularisation** is essential to overcome the limitations of maximum likelihood estimation.

✦ **Other names:** Regularisation, penalised regression, shrinkage.





Maximum a Posteriori (MAP)

Optimisation Objective

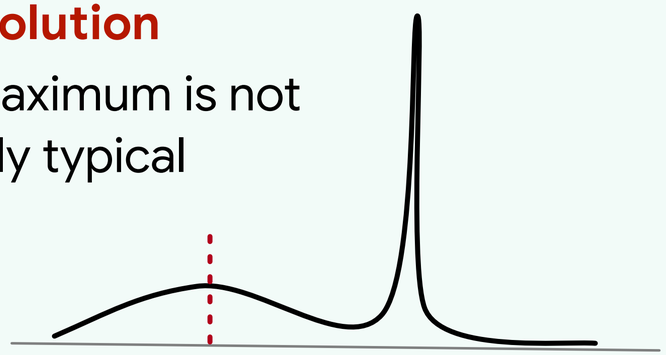
$$\arg\max_{\theta} \mathcal{L}(\boldsymbol{\theta})$$

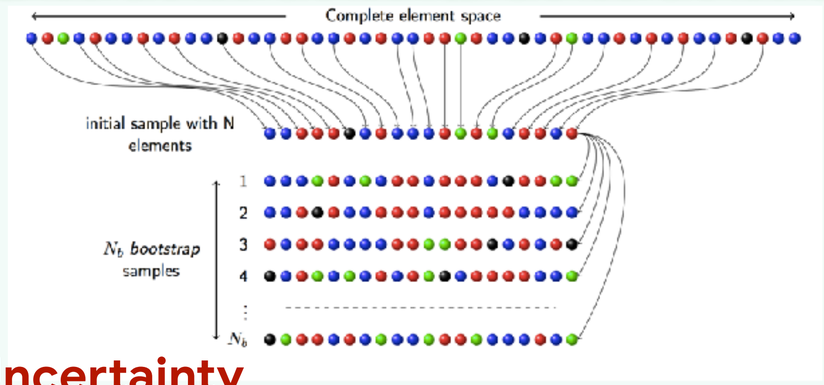
- Generalises the MLE (uniform prior)
- **Shrinkage**: shrink parameters back to initial beliefs.
- Not every regulariser corresponds a valid probability distribution.

MAP Estimation

Type of Solution

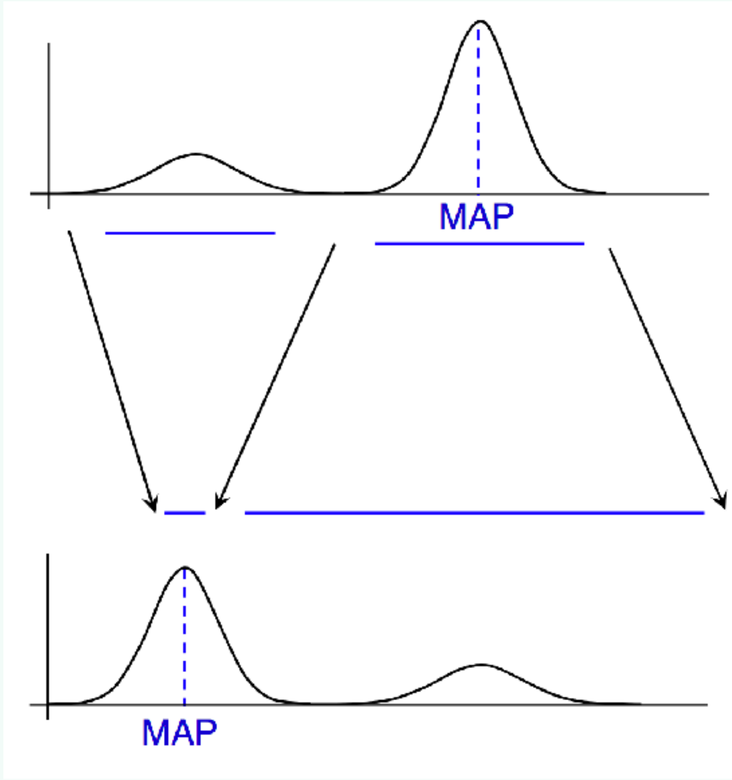
What is maximum is not necessarily typical





Uncertainty

Can be reported using confidence intervals or bootstrap estimates.



Parameterisation sensitive

Location of max will change depending on parameterisation

Invariant MAP

Popular Example

$$y \in 0,1; \quad 0 \leq \mu \leq 1$$

Change of variables

$$p(\phi) = p(\mu) \left| \frac{d\mu}{d\phi} \right|$$

Bernoulli

$$p(y = 1|\mu) = \mu$$

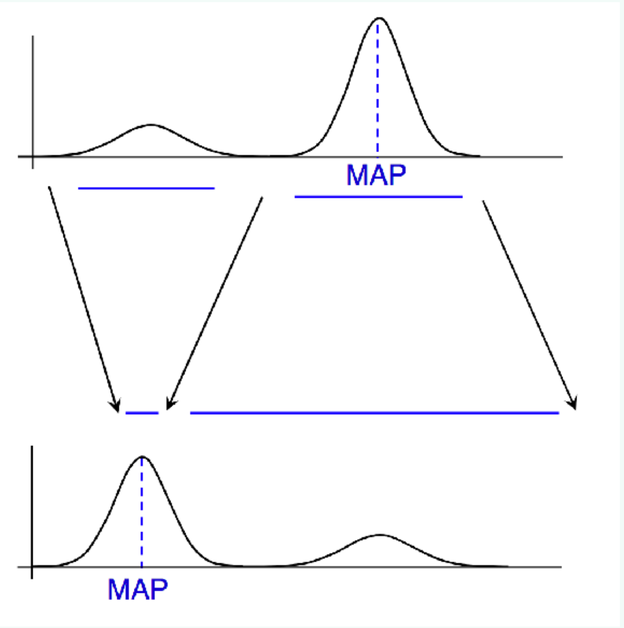
Uniform

$$p(\mu) = 1$$

Mode of the prior

$$\hat{\phi}_{MAP} = \arg \max_{\phi \in [0,1]} p(\phi)$$

	Parameterisation 1	**Parameterisation 2**
Transform	$\mu = \phi^2$	$\mu = 1 - (1 - \phi)^2$
New prior	$p(\phi) = 2\phi$	$p(\phi) = 2(1 - \phi)$
MAP Est.	$\hat{\phi}_{MAP} = 1$	$\hat{\phi}_{MAP} = 0$



Clear sensitivity: Sensitive to units, affects interpretability, affects gradients, learning stability, design of models.

Invariant MAP

Use a modified probabilistic model that removes sensitivity

Invariant MAP $p(y|h(\mathbf{x});\boldsymbol{\theta})p(\boldsymbol{\theta})|\mathcal{I}(\boldsymbol{\theta})|^{\frac{1}{2}}$

- Use the Fisher information
- Connection to the natural gradients and trust-region optimisation.
- Uninformative priors.

Proposed solutions have not fully dealt with the underlying issues.

Bayesian Inference

Evidence $p(\mathbf{x}) = \int p(\mathbf{x}|\boldsymbol{\theta})p(\boldsymbol{\theta})\mathrm{d}\boldsymbol{\theta}$

Posterior $p(\boldsymbol{\theta}|\mathbf{x})$

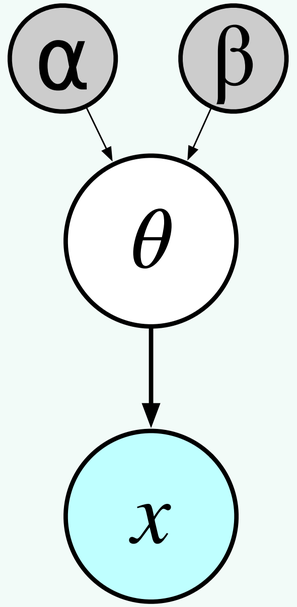
How and when can this integral to be computed tractably?

What pairs of likelihood and prior make this easy?

Approach 1: Using Conjugacy

- Choose priors by mimicking the form of the likelihood.
- In a large set of cases, the posterior is in the same form of the observation distribution, e.g., Beta prior -> Beta posterior
- Priors chosen this way are called conjugate priors.

Beta-Bernoulli Model



Bernoulli Distribution

$$p(x|\theta) = \theta^x(1-\theta)^{(1-x)}$$

Conjugate Prior

$$p(\theta|\alpha,\beta) \propto \theta^{\alpha}(1-\theta)^{(1-\beta)}$$

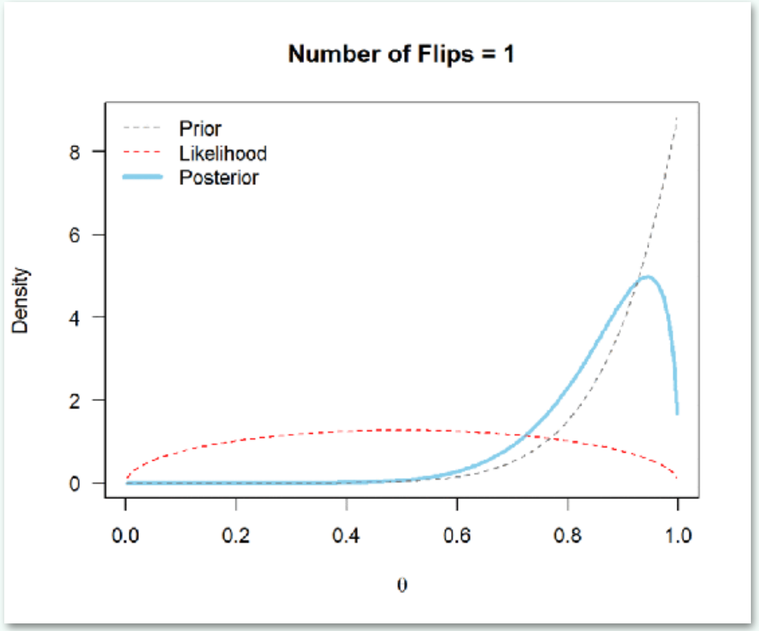
Beta Distribution

$$p(\theta|\alpha,\beta) \propto \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}\theta^{\alpha}(1-\theta)^{(1-\beta)}$$

Posterior Beta Distribution

$$p(\theta|x) = \mathcal{B}(\theta|\alpha^*,\beta^*)$$

$$\alpha^* = \sum_{n=1}^{N} x_n + \alpha - 1 \qquad \beta^* = N - \sum_{n=1}^{N} x_n + \beta - 1$$

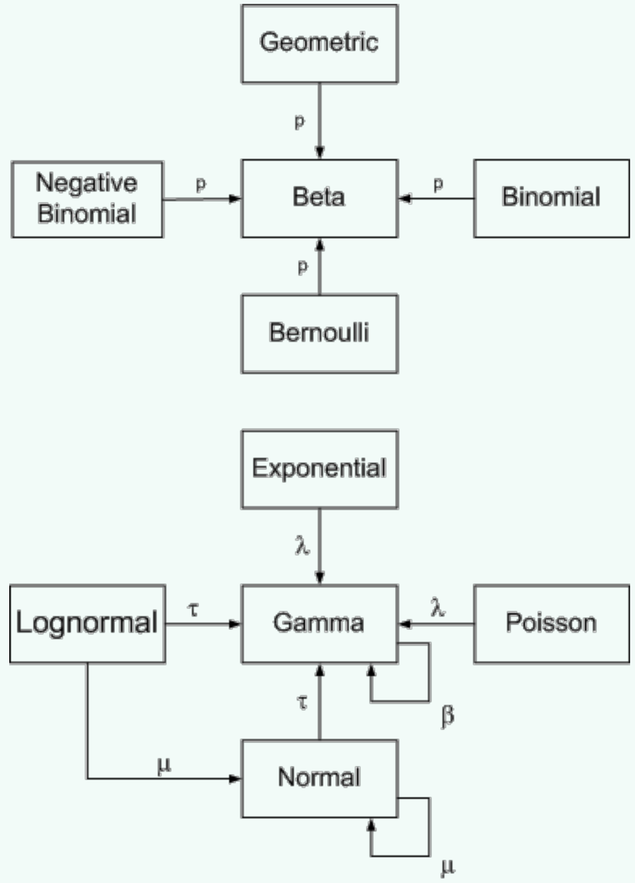


https://jimgrange.wordpress.com/2016/01/18/pesky-priors/

Conjugacy

Conjugate priors are those in the **exponential family of distributions**.

- Because of the ‘closure’ property, they make recursive updating very easy to do.
- Get closed-form local computations allowing automated inference.
- Allows us to study relationship to maximum likelihood, and relationship to convexity, information geometry, Bregman divergences.
- General and interesting in more complex scenarios: multivariate systems, priors over functions, infinite dimensional systems,



Measure

Normalising Constant

$$p(x|\theta) = h(x) \exp\left(\eta(\theta) T(x) - A(\eta)\right)$$

Natural Parameters

Sufficient Statistics

Table 1.2: Several well known exponential families listing their log-partition functions $A(\eta)$, conjugate dual functions $A^*(\theta)$, corresponding Bregman divergence $B(x\|\theta)$ and canonical link functions $\eta(\theta)$.

Family	$A(\eta)$	$A^*(\theta)$	$B_{A^*}(x\|\theta)$	$\eta(\theta)$
Bernoulli	$-\log(1+\exp(\eta))$	$\theta\ln\theta$–$(1-\theta)\ln(1-\theta)$	$\ln(1+\exp(x^*\theta))$ $x^*=2x-1$	$\ln\left(\frac{\theta}{1-\theta}\right)$
Exponential	$-\log(-\eta)$	$\theta\ln\theta-\theta$	$x\ln\left(\frac{x}{\theta}\right)-(x-\theta)$	θ
Poisson	$\exp(\eta)$	$-(1+\ln\theta)$	$\frac{x}{\theta}-\ln\left(\frac{x}{\theta}\right)-1$	$\ln(\theta)$
Multinomial	$\ln\left(1+\sum_{i=1}^{k-1}\exp(\eta_i)\right)$	$\sum_{j=1}^{k}\theta_j\ln\left(\frac{\theta_j}{N}\right)$	$\sum_{j=1}^{k}x_j\ln\left(\frac{x_j}{\theta_j}\right)$	$\ln\left(\frac{\theta_j}{1-\sum_{i=1}^{k-1}\theta_i}\right)$
Gaussian (location family)	$\frac{1}{2\sigma^2}\eta^2$	$\frac{1}{2\sigma^2}\theta^2$	$\frac{(x-\theta)^2}{2\sigma^2}$	θ

https://www.johndcook.com/blog/conjugate_prior_diagram/

Integral Approximations

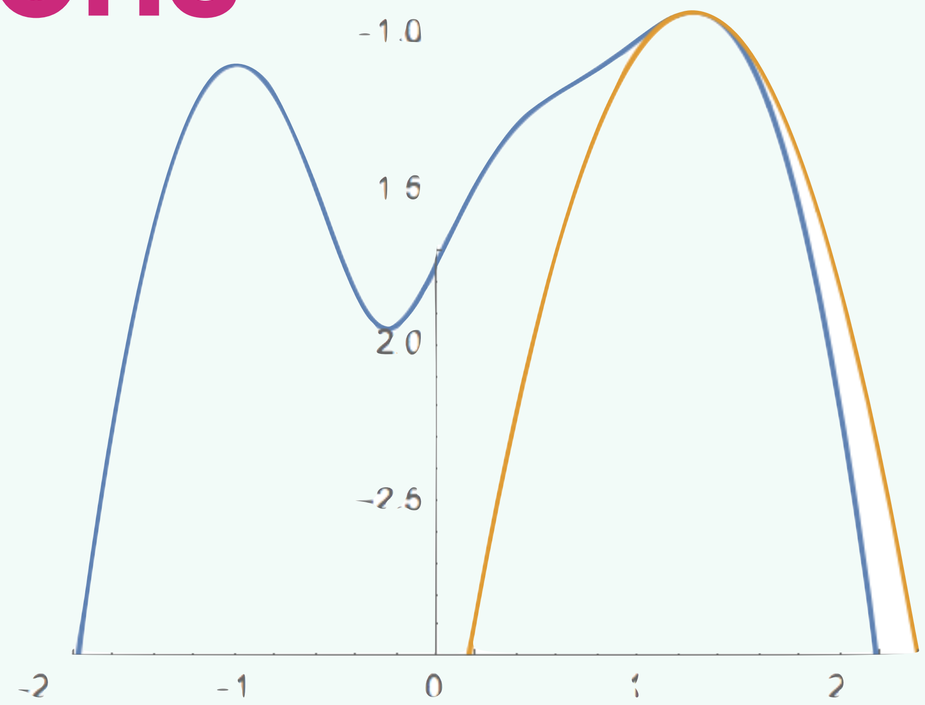
Integral Problem

$$\log p(\mathbf{x}) = \int p(\mathbf{x}|\boldsymbol{\theta})p(\boldsymbol{\theta})\mathrm{d}\boldsymbol{\theta}$$

Energy

$$u(\mathbf{x}, \boldsymbol{\theta}) = -\log\left(p(\mathbf{x}|\boldsymbol{\theta})p(\boldsymbol{\theta})\right)$$

$$\log \int \exp\left\{-u(\mathbf{x}, \boldsymbol{\theta})\right\}$$



Taylor Expansion

$$u(\mathbf{x}, \boldsymbol{\theta}) \approx u(\mathbf{x}, \boldsymbol{\mu}) + \frac{1}{2}(\boldsymbol{\theta} - \boldsymbol{\mu})^\top \mathbf{H}(\boldsymbol{\theta} - \boldsymbol{\mu})$$

$\mathbf{H}(\boldsymbol{\mu}) = \nabla^2_{\boldsymbol{\theta}} u(\mathbf{x}, \boldsymbol{\theta})|_{\theta=\mu}$

Use MAP estimate for $\boldsymbol{\mu}$

$$\log \int \exp\left\{-u(\mathbf{x}, \boldsymbol{\mu}) - \frac{1}{2}(\boldsymbol{\theta} - \boldsymbol{\mu})^\top \mathbf{H}(\boldsymbol{\mu})(\boldsymbol{\theta} - \boldsymbol{\mu})\right\} \mathrm{d}\boldsymbol{\theta}$$

Related names:
Delta Method or
Saddle-point approx.

Laplace Approximation

$$-u(\mathbf{x}, \boldsymbol{\mu}) - \frac{1}{2}\log\det(2\pi\mathbf{H}^{-1})$$

Take a minute to think of your answer.

What is Learning? What is Inference?

Afterwards, raise hand/unmute/share your answer.
Or write in channel `#Lec_bayesian_inference_mohamed`

Learning and Inference

Statistics, no distinction between learning and inference - only inference (or **estimation**).

Machine learning makes a distinction between **inference and learning**:

- **Inference**: reason about (and compute) unknown probability distributions.
- (Parameter) **Learning** is finding point estimates of quantities in the model.

Bayesian statistics, all quantities are probability distributions, so there is only the problem of **inference**.

Software engineering, **inference** is the forward evaluation of a trained model (to get predictions).

Decision making and AI, refer to **learning** in general as the means of understanding and acting based on past experience (data).

Prediction

Posterior predictive distribution is the prediction of new data given the model.

Likelihood of the test data averaged over the posterior distribution.

Posterior Predictive

$$p(\mathbf{x}^*|\mathbf{x}) = \int p(\mathbf{x}^*|\boldsymbol{\theta})p(\boldsymbol{\theta}|\mathbf{x})\mathrm{d}\boldsymbol{\theta} = \mathbb{E}_{p(\boldsymbol{\theta}|\mathbf{x})}\left[p(\mathbf{x}^*|\boldsymbol{\theta})\right]$$

- In conjugate models, this predictive distribution can be known in closed form.
- For the Beta-Bernoulli model this is a Beta distribution.
- In other cases, we will have to approximate the integral or use a Monte Carlo method to evaluate the integral.

Bayes Factors

Bayes Factor is a way of comparing two models. It is a Bayesian approach to hypothesis testing.

Bayes Factor

$$B = \frac{p(\mathbf{x}|\mathcal{M}_1)}{p(\mathbf{x}|\mathcal{M}_2)} = \frac{\int p(\mathbf{x}|\boldsymbol{\theta_1}, \mathcal{M}_1) p(\boldsymbol{\theta}_1|\mathcal{M}_1) \mathrm{d}\boldsymbol{\theta}_1}{\int p(\mathbf{x}|\boldsymbol{\theta_2}, \mathcal{M}_2) p(\boldsymbol{\theta}_2|\mathcal{M}_2) \mathrm{d}\boldsymbol{\theta}_2}$$

Posterior Odds

$$\frac{p(\mathcal{M}_1|\mathbf{x})}{p(\mathcal{M}_2|\mathbf{x})} = \frac{p(\mathcal{M}_1)}{p(\mathcal{M}_2)} \frac{p(\mathbf{x}|\mathcal{M}_1)}{p(\mathbf{x}|\mathcal{M}_2)}$$

- Compares to competing models rather than against a null hypothesis.
- Accounts for uncertainty
- Can compare nested models.
- Large-sample approximations of Bayes factors: BIC,AIC, DIC, WAIC.

Central problem is computing the marginal likelihood.

Evidence

Marginal likelihoods have the following properties.

- **Consistency**: As the number of data points becomes large, it will favour the true model.
- **Ockham's razor:** We will prefer simpler models to more complex ones if they have the same performance.
- **Comparison**: Models and parameters that are compared need not be nested or equivalent in any way.
- **Reference**: We can store the marginal likelihood as a property of the model-data instance, and use it for any future model selection or comparison.
- **Weight-of-evidence**: We can compute the evidence for data points individually and use this as a score or measure of surprise, allowing us to characterise each data point we observe.

Learning Principles

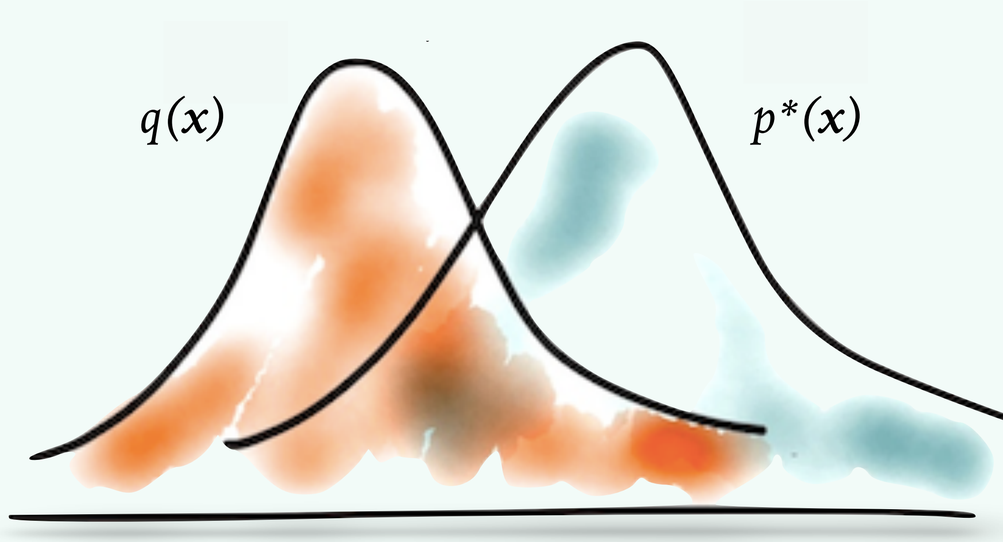
Bayesian computations show that there are different types of learning principles that are available to us.

Learning principle: Model Evidence

$$p(\mathbf{x}) = \int p(\mathbf{x}, \mathbf{z}) d\mathbf{z}$$

Learning principle: Two-sample tests

$$\frac{p^*(\mathbf{x})}{q(\mathbf{x})} = 1 \qquad p^*(\mathbf{x}) = q(\mathbf{x})$$



Inferential Questions

Evidence Estimation

$$p(\mathbf{x}) = \int p(\mathbf{x}, \mathbf{z}) d\mathbf{z}$$

Moment Computation

$$\mathbb{E}[f(\mathbf{z})|\mathbf{x}] = \int f(\mathbf{z}) p(\mathbf{z}|\mathbf{x}) d\mathbf{z}$$

Parameter Estimation

$$p(\boldsymbol{\theta}|\mathbf{x}_{0:N})$$

Prediction

$$p(\mathbf{x}_{t+1}|\mathbf{x}_{0:t})$$

Planning

$$\mathcal{J} = \mathbb{E}_p \left[\int_0^\infty C(\mathbf{x}_t) dt | \mathbf{x}_0, \mathbf{u} \right]$$

Hypothesis Testing

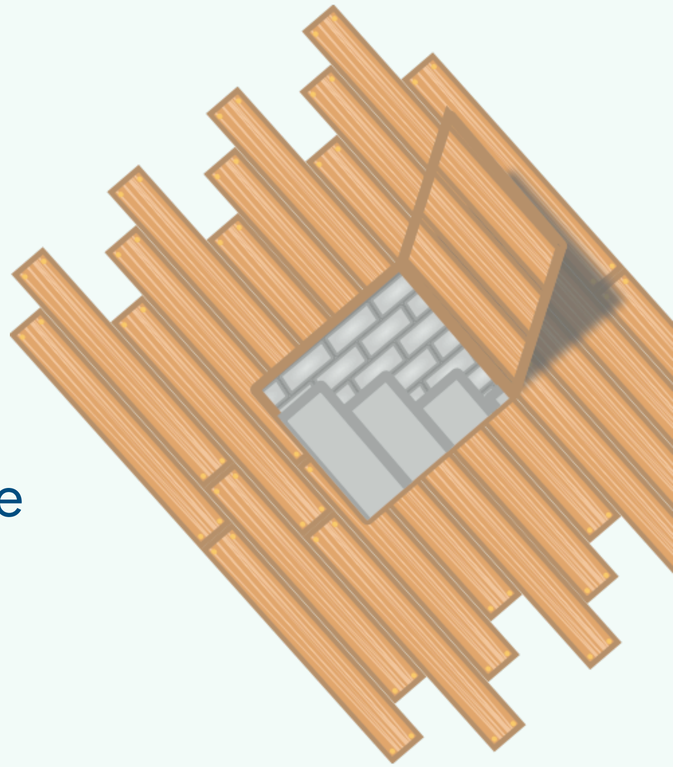
$$\mathcal{B} = \log p(\mathbf{x}|H_1) - \log p(\mathbf{x}|H_2)$$

Experimental Design

$$\mathcal{IG} = D[p(\mathbf{x}_{t:T}|u) \| p(\mathbf{x}_{0:t})]$$

Neutrality Traps

- **The Solutionism Trap**: Failure to recognise the possibility that the best solution to a problem may not involve technology.
- **The Formalism Trap:** Failure to account for the full meaning of social concepts such as fairness, which be resolved through mathematical formalisms.
- **The Portability Trap:** Failure to understand how repurposing algorithmic solutions designed for one social context may be inaccurate / do harm when applied to a different context.
- **The Ripple Effect Trap**: Failure to understand how the insertion of technology into an existing social system changes the behaviours and embedded values of the pre-existing system .



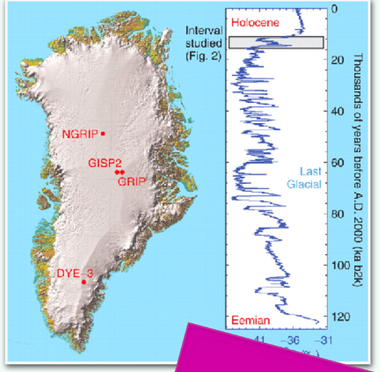
"Technology is neither good nor bad; nor is it neutral."

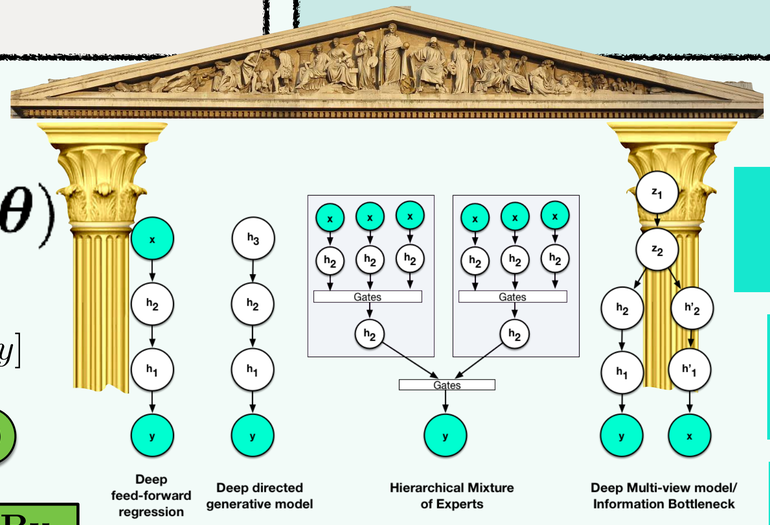
Statistical Probability

Frequency ratio of items

Subjective Probability

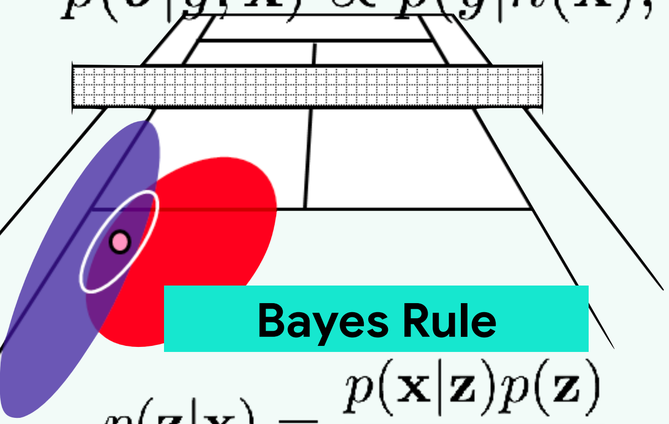
Probability as a degree of belief





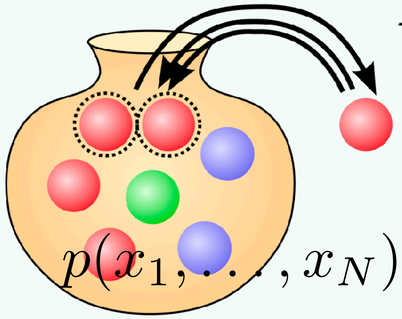
Posterior

$$p(\boldsymbol{\theta}|y,\mathbf{x}) \propto p(y|h(\mathbf{x});\boldsymbol{\theta})p(\boldsymbol{\theta})$$

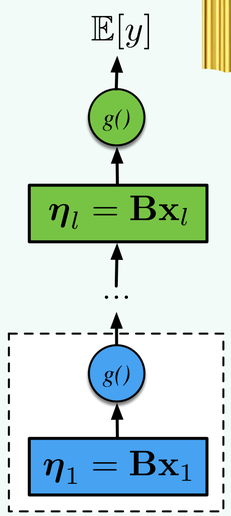


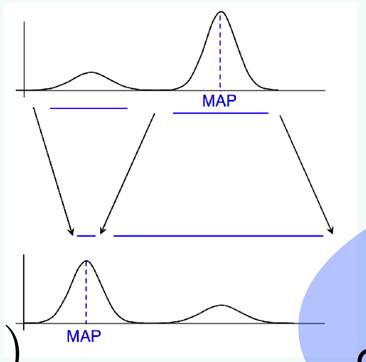
Bayes Rule

$$p(\mathbf{z}|\mathbf{x}) = \frac{p(\mathbf{x}|\mathbf{z})p(\mathbf{z})}{p(\mathbf{x})}$$



$$p(x_1,\ldots,x_N) = \int \prod_{n=1}^{N} p(x_n|\theta)P(\mathrm{d}\theta)$$



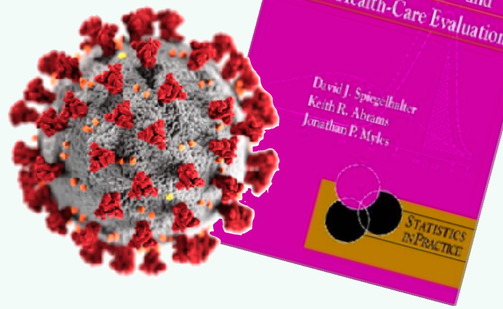


Exponential Family and Conjugacy

Laplace Approximation

Posterior Predictive

Bayes Factor



Epistemic values

Contextual Values

Bayesian statistics, all quantities are probability distributions, so there is only the problem of **inference**.

Some papers and books

Nelder, John Ashworth, and Robert WM Wedderburn. "Generalized linear models." Journal of the Royal Statistical Society: Series A (General) 135.3 (1972): 370-384.

Goodfellow, Ian, Yoshua Bengio, and Aaron Courville. Deep learning. MIT press, 2016.

Gelman, Andrew, et al. Bayesian data analysis. CRC press, 2013.

Murphy, K.P., 2012. Machine Learning: A Probabilistic Perspective.

Friedman, Jerome, Trevor Hastie, and Robert Tibshirani. The elements of statistical learning. Vol. 1. No. 10. New York: Springer series in statistics, 2001.

Owen, Art B. Empirical likelihood. CRC press, 2001.

Jermyn, Ian H. "Invariant Bayesian estimation on manifolds." The Annals of Statistics 33.2 (2005): 583-605.

Blei, David M, The Exponential Family, notes.

Wainwright, Martin J., and Michael I. Jordan. "Graphical models, exponential families, and variational inference." Foundations and Trends in Machine Learning 1, no. 1-2 (2008): 1-305.

Kass, Robert E., and Adrian E. Raftery. "Bayes factors." Journal of the american statistical association 90.430 (1995): 773-795.

Selbst, Andrew D., et al. "Fairness and abstraction in sociotechnical systems." Proceedings of the Conference on Fairness, Accountability, and Transparency. 2019.

Shakir Mohamed

@shakir_za

End of Part 1

Bayesian Learning

Basics | Computation | Approximation | Futures

Bayesian Learning

Basics | Computation | Approximation | Futures

Shakir Mohamed

@shakir_za

DeepMind

3

Bayesian | Approximation

Shakir Mohamed

@shakir_za

Outcomes

1. Direct and Indirect Inferences
2. Monte Carlo methods
3. Variational Methods

Inferential Questions

Evidence Estimation

$$p(\mathbf{x}) = \int p(\mathbf{x}, \mathbf{z}) d\mathbf{z}$$

Moment Computation

$$\mathbb{E}[f(\mathbf{z})|\mathbf{x}] = \int f(\mathbf{z}) p(\mathbf{z}|\mathbf{x}) d\mathbf{z}$$

Parameter Estimation

$$p(\boldsymbol{\theta}|\mathbf{x}_{0:N})$$

Prediction

$$p(\mathbf{x}_{t+1}|\mathbf{x}_{0:t})$$

Planning

$$\mathcal{J} = \mathbb{E}_p \left[\int_0^\infty C(\mathbf{x}_t) dt | \mathbf{x}_0, \mathbf{u} \right]$$

Hypothesis Testing

$$\mathcal{B} = \log p(\mathbf{x}|H_1) - \log p(\mathbf{x}|H_2)$$

Experimental Design

$$\mathcal{IG} = D[p(\mathbf{x}_{t:T}|u) \| p(\mathbf{x}_{0:t})]$$

Learning Principles

- **Statistical Inference**
  - **Direct**
    - Laplace approximation
    - Maximum Likelihood
    - Maximum a posteriori
    - Variational Inference
    - Cavity Methods
    - Integr. Nested Laplace Approx
    - Expectation Maximisation
    - Markov chain Monte Carlo
    - Noise Contrastive
    - Sequential Monte Carlo
  - **Indirect**
    - Two Sample Comparison
    - Method of Moments
    - Approx Bayesian Computation
    - Transportation methods
    - Max Mean Discrepency

Take a minute to think of your answer.

How do you represent a distribution?

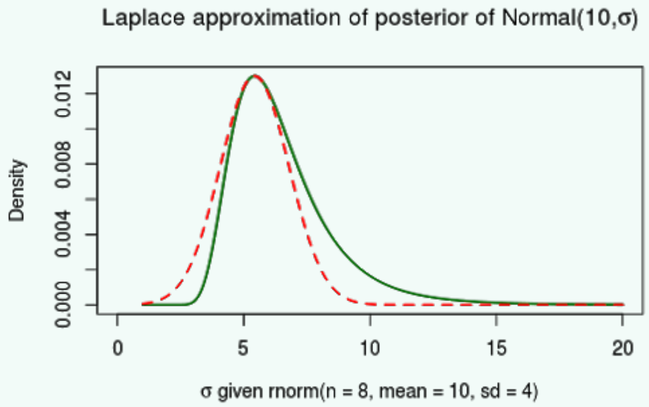
Afterwards, raise hand/unmute/share your answer.
Or write in channel `#Lec_bayesian_inference_mohamed`

Representing Distributions

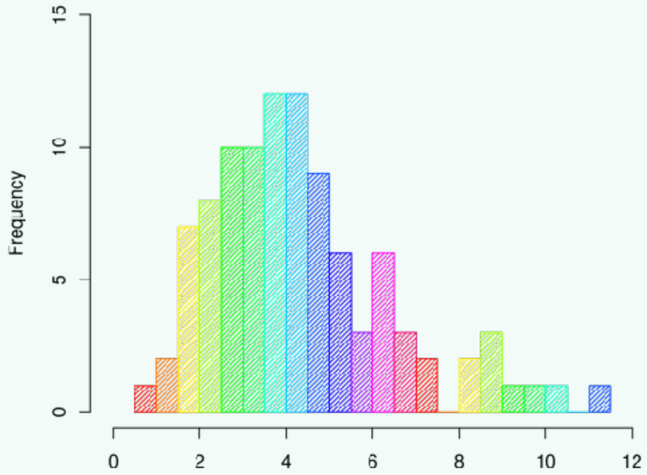
Closed-form/Analytic

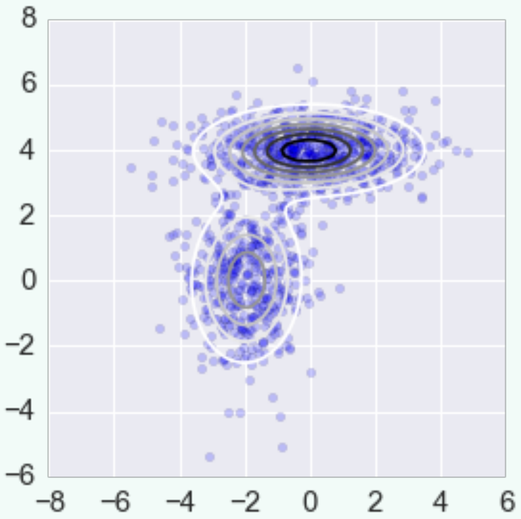
$$\mathcal{B}(\theta \mid \alpha, \beta)$$

Approximations

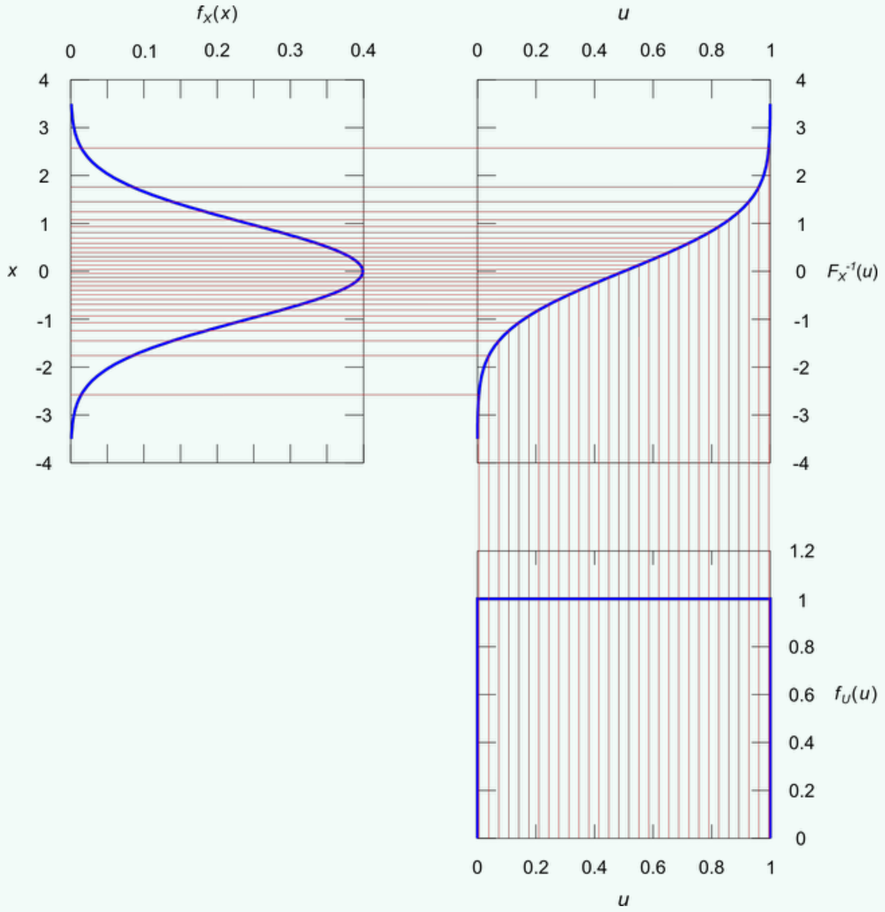


Samples





Sampling procedure



Monte Carlo methods

One of the most general methods we have for computing integrals and probabilities.

Integral

$$\mathcal{F}(\boldsymbol{\theta}) = \int f(\mathbf{x}) p(\mathbf{x}|\boldsymbol{\theta}) d\mathbf{x} = \mathbb{E}_{p(x|\theta)}[f(\mathbf{x})]$$

Monte Carlo method is simple:

1. Draw independent samples from p.
2. Compute the average of the function at these points.

Estimator

$$\hat{\mathcal{F}}(\boldsymbol{\theta}) = \frac{1}{N} \sum_{n=1}^{N} f(\hat{\mathbf{x}}^n); \quad \hat{\mathbf{x}}^n \sim p(\mathbf{x}|\boldsymbol{\theta}), \text{for } n = 1, \ldots, N$$

Monte Carlo methods

Estimator

$$\hat{\mathcal{F}}(\boldsymbol{\theta}) = \frac{1}{N}\sum_{n=1}^{N} f(\hat{\mathbf{x}}^n); \quad \hat{\mathbf{x}}^n \sim p(\mathbf{x}|\boldsymbol{\theta}), \text{for } n = 1, \ldots, N$$

- **Consistency**: The estimate should converge to the true value of the integral.
- **Unbiased**: When we repeat this process using different sets of samples, the estimate should be centred on the true value.
- **Low variance**: The estimator is a random variable because many sets of samples can be drawn to compute the average.
- **Computation**: Want estimators that are computationally efficient to compute, require few samples, easy parallelisation, and easy generation of variates x.

Evaluating Integrals

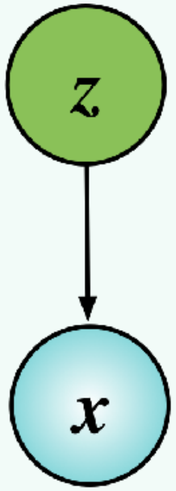
$$\int p(\mathbf{x}) f(\mathbf{x}) d\mathbf{x} = \mathbb{E}_{p(\mathbf{x})}[f(\mathbf{x})]$$

$$\updownarrow$$

$$\mathbb{E}_{q(\mathbf{x})}[g(\mathbf{x}; f)] = \int q(\mathbf{x}) g(\mathbf{x}, f) d\mathbf{x}$$

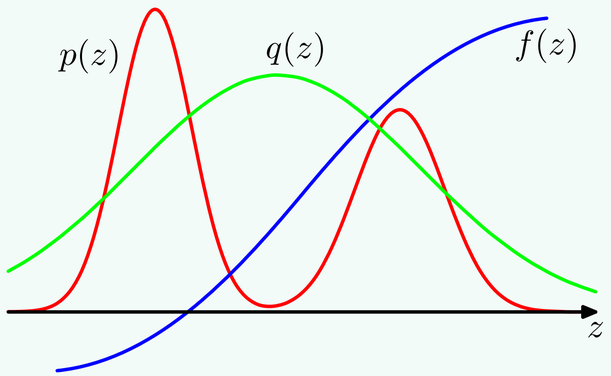
Do this by introducing a **probabilistic one** $\frac{p(\mathbf{x})}{p(\mathbf{x})}$

Importance Sampling



Conditions

- $q(z)>0$, when $f(z)p(z) \neq 0$.
- *$q(z)$ is known/easy to handle.*



Integral problem $$m = \int f(\mathbf{z})p(\mathbf{z})d\mathbf{z}$$

Probabilistic one $$m = \int f(\mathbf{z})p(\mathbf{z})\frac{q(\mathbf{z})}{q(\mathbf{z})}d\mathbf{z}$$

Re-group/re-weight $$m = \int f(\mathbf{z})\frac{p(\mathbf{z})}{q(\mathbf{z})}q(\mathbf{z})d\mathbf{z}$$

$$m = \mathbb{E}_{q(\mathbf{z})}\left[f(\mathbf{z})\frac{p(\mathbf{z})}{q(\mathbf{z})}\right]$$

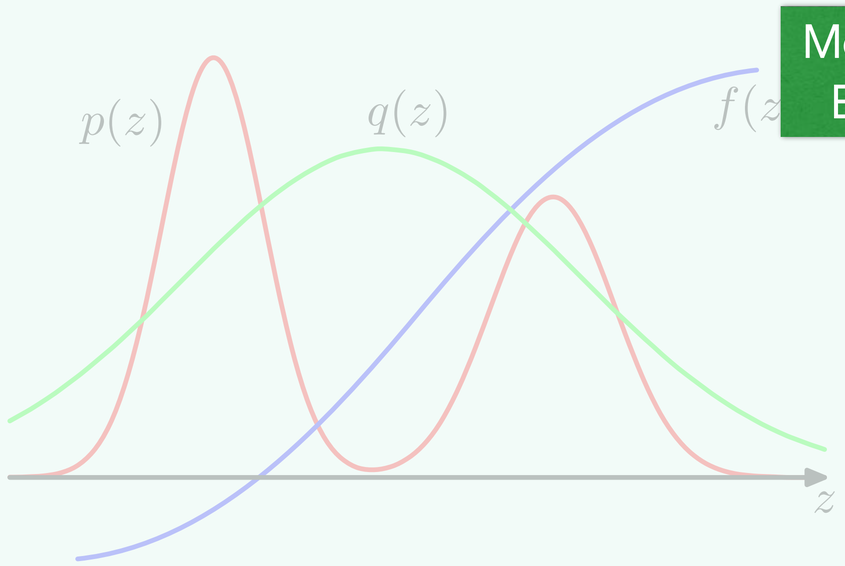
Importance Sampling

$$m = \mathbb{E}_{q(\mathbf{z})}\left[f(\mathbf{z})\frac{p(\mathbf{z})}{q(\mathbf{z})}\right]$$

Identity Trick Elsewhere

- Manipulate stochastic gradients
- Derive probability bounds
- RL for policy corrections

$$w^{(s)} = \frac{p(z)}{q(z)} \quad z^{(s)} \sim q(z)$$



Monte Carlo Estimator

$$m = \frac{1}{S}\sum_s w^{(s)} f(\mathbf{z})$$

$$\log m = \log \sum_s w^{(s)} f(\mathbf{z}) - \log S$$

Markov Chain Monte Carlo

Can we generate samples using a search process and waiting till we reach a stationary distribution.

1st-order Markov chain $p(\mathbf{x}_t|\mathbf{x}_{1:t-1}) = p(\mathbf{x}_t|\mathbf{x}_{t-1})$

Reversibility $Q(\mathbf{x}|\mathbf{y})\pi(\mathbf{y}) = Q(\mathbf{y}|\mathbf{x})\pi(\mathbf{x})$

- **Irreducible**. We can get from any state to any other state. We don't get stuck in a (set of) states.
- **Aperiodic.** We don't loop between states.
- These two conditions mean we have an **ergodic** chain, i.e. it has a stationary distribution.

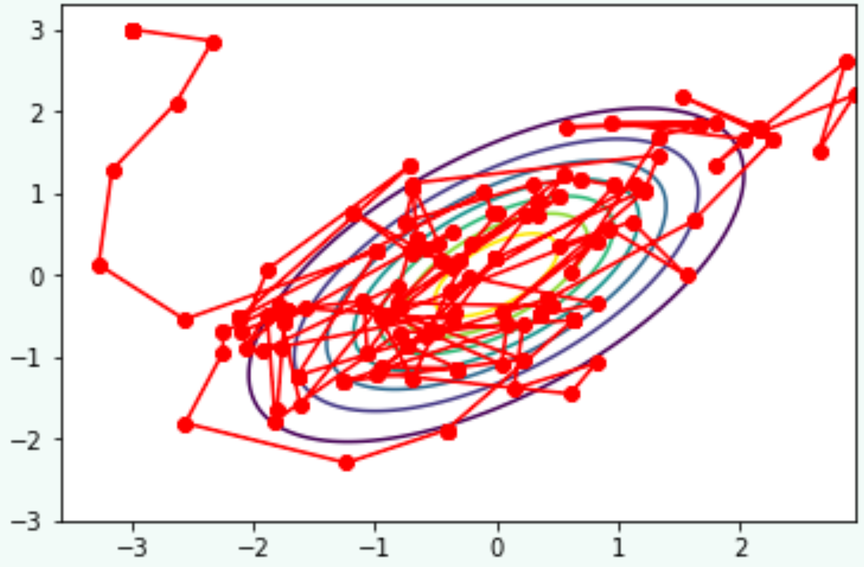
Metropolis Hastings Algorithm

Target $p(\mathbf{x}) = \frac{\tilde{p}(\mathbf{x})}{Z_p}$

Proposal $\mathbf{y} \sim Q(\cdot|\mathbf{x})$

MH Criterion $\alpha(\mathbf{y}|\mathbf{x}) := \min\left\{\frac{\tilde{p}(\mathbf{x})}{\tilde{p}(\mathbf{y})}\frac{Q(\mathbf{x}|\mathbf{y})}{Q(\mathbf{y}|\mathbf{x})}, 1\right\}$

Accept/Reject $\mathbf{x}_{t+1} = \mathbf{y}$ $\mathbf{x}_{t+1} = \mathbf{x_t}$



Other MCMC Methods

- Rejection sampling
- Metropolis-Hastings
- Gibbs sampling
- Slice sampling
- Metropolis within Gibbs
- Hamiltonian MC
- Sequential MC
- Reversible Jump MC
- Non-Markovian Methods (Stein, nested methods)

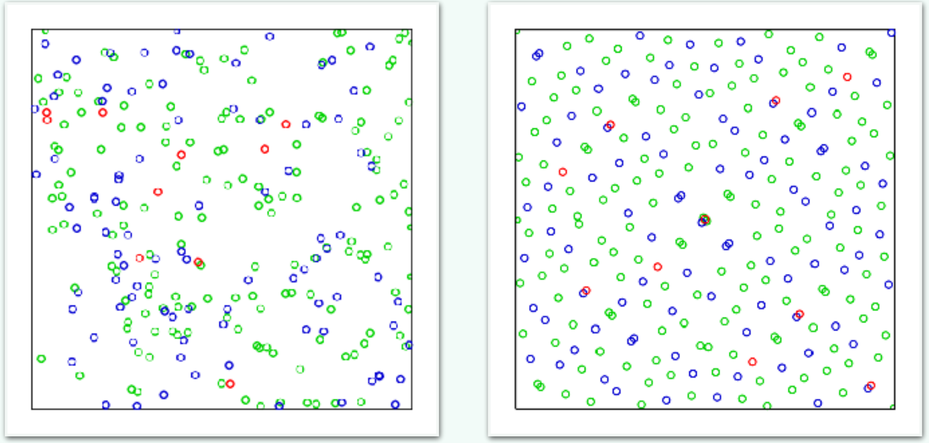
Limitations and Considerations

- MCMC methods can be computationally intensive and slow because of the need to simulate.
- Can be difficult to evaluate when you have a good set of samples and to know when your chain has reached the stationary distribution (has mixed).
- They Markov property doesn't make use of where in the state space past samples have been taken, so they don't make efficient use of the information they have.
- Take a frequentist approach of large-sample behaviour to address the needs of Bayesian averaging.

https://chi-feng.github.io/mcmc-demo/app.html?algorithm=HamiltonianMC&target=banana

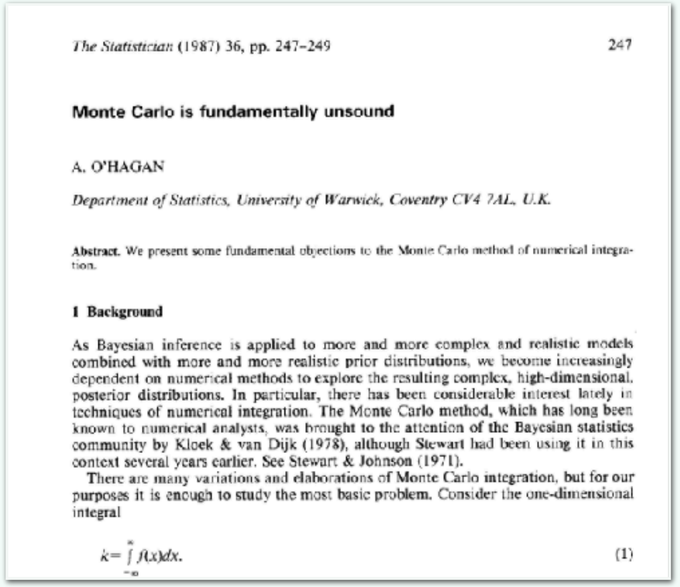
Beyond Monte Carlo

Quasi-MC methods



- Instead of using random draws in the evaluation of an MC estimator, can we instead use a more structured set of samples to allow faster convergence.
- These are low-discrepancy sequences and known as quasiMC methods.

Bayesian MC



The Statistician (1987) 36, pp. 247–249 247

Monte Carlo is fundamentally unsound

A. O'HAGAN

Department of Statistics, University of Warwick, Coventry CV4 7AL, U.K.

Abstract. We present some fundamental objections to the Monte Carlo method of numerical integration.

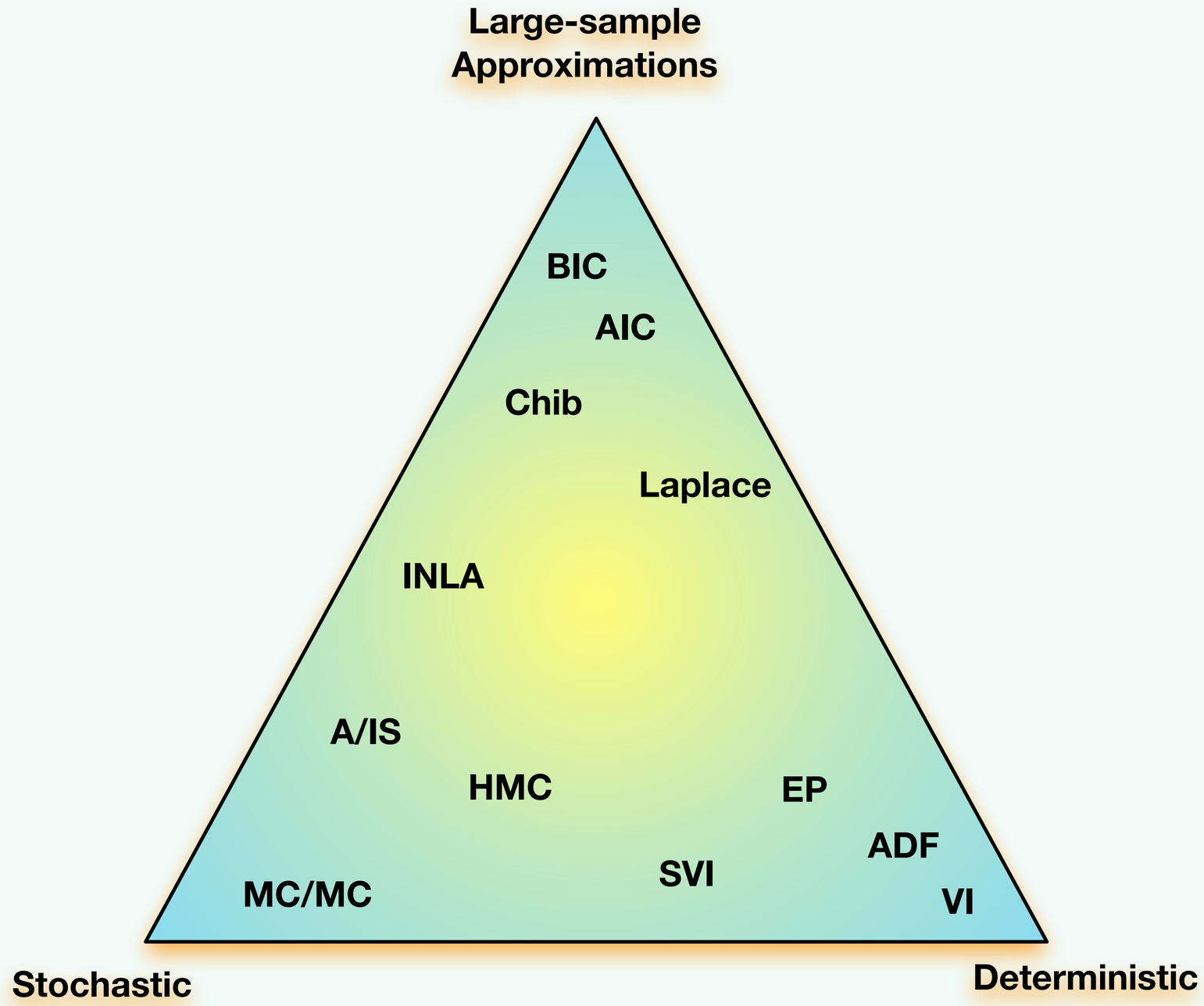
1 Background

As Bayesian inference is applied to more and more complex and realistic models combined with more and more realistic prior distributions, we become increasingly dependent on numerical methods to explore the resulting complex, high-dimensional, posterior distributions. In particular, there has been considerable interest lately in techniques of numerical integration. The Monte Carlo method, which has long been known to numerical analysts, was brought to the attention of the Bayesian statistics community by Kloek & van Dijk (1978), although Stewart had been using it in this context several years earlier. See Stewart & Johnson (1971).

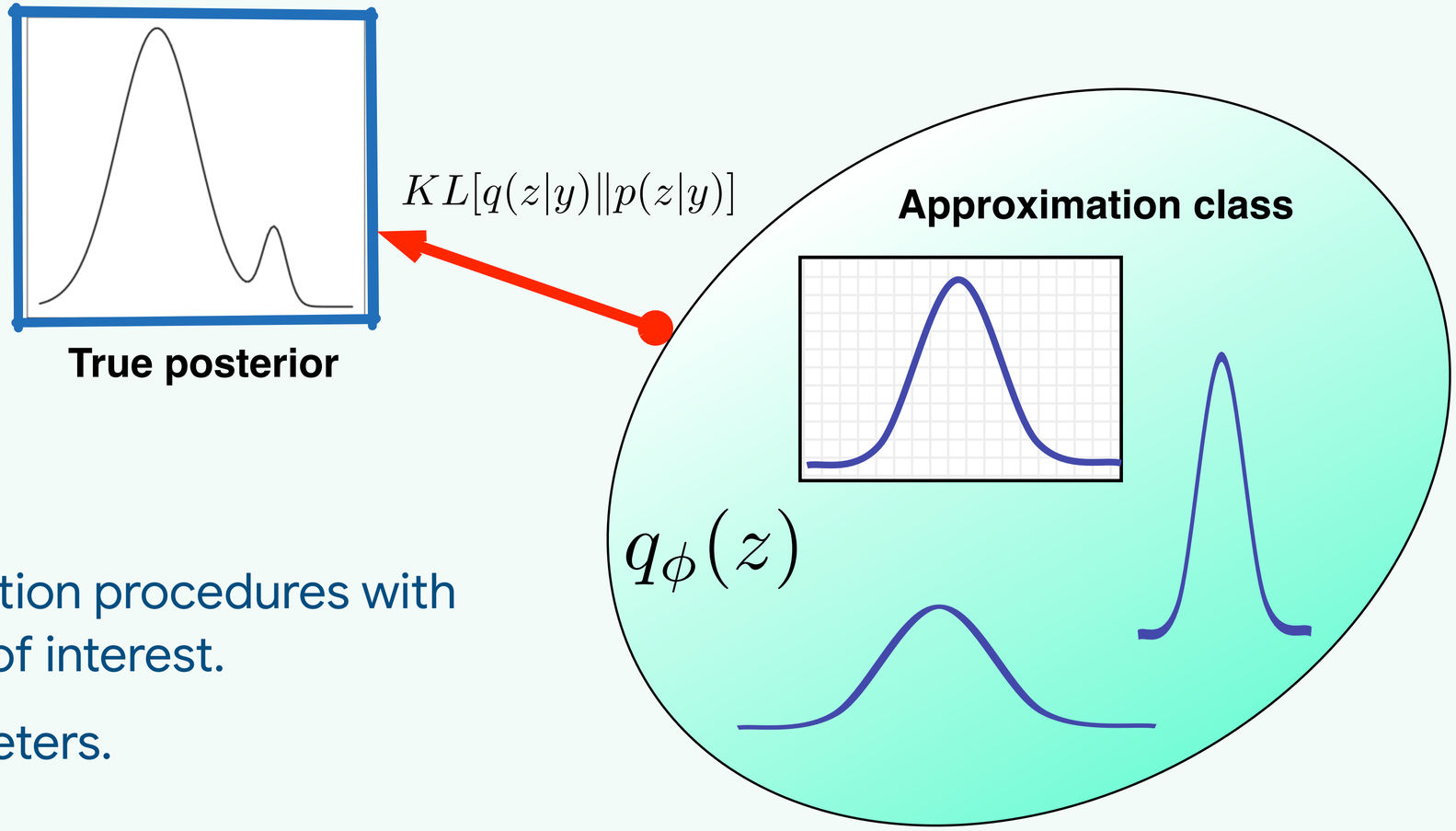
There are many variations and elaborations of Monte Carlo integration, but for our purposes it is enough to study the most basic problem. Consider the one-dimensional integral

$$k = \int_{-\infty}^{\infty} f(x)dx. \quad (1)$$

- Integrals are unknown quantities, so can put prior over this unknown value.
- Compute the posterior distribution over the unknown integral.


Large-sample
Approximations
BIC
AIC
Chib
Laplace
INLA
A/IS
HMC
EP
ADF
SVI
MC/MC
VI
Stochastic
Deterministic

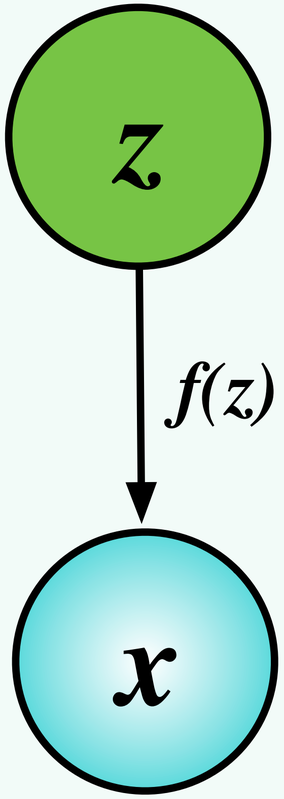
Variational Methods

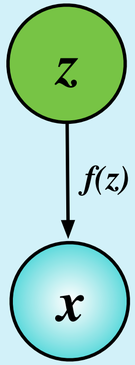


Deterministic approximation procedures with bounds on probabilities of interest.

Fit the variational parameters.

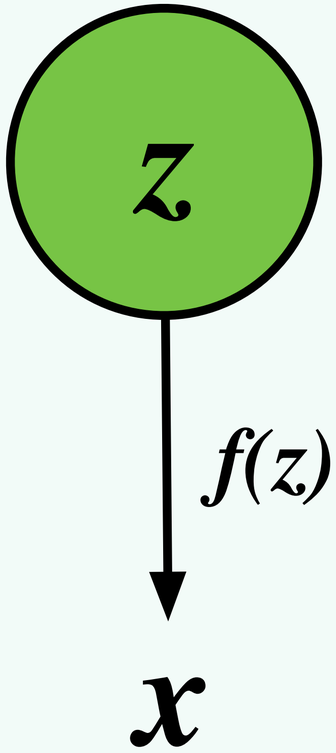
Latent Variable Models





Latent variable models

Introduce an unobserved **local random variables** that represents hidden causes.



Prescribed models

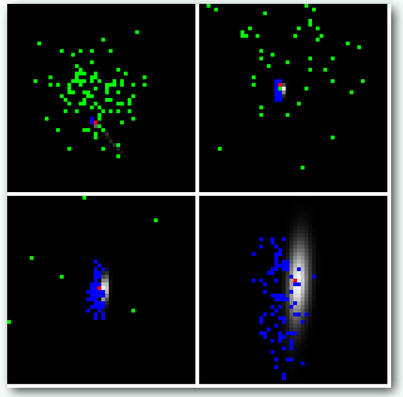
Use observer likelihoods and assume observation noise.

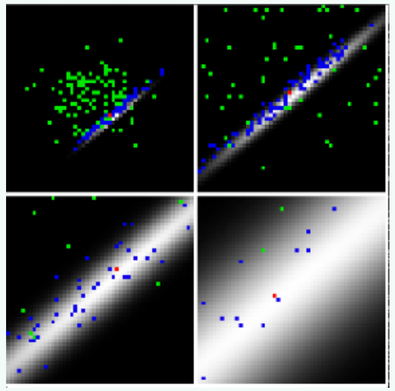
Implicit models

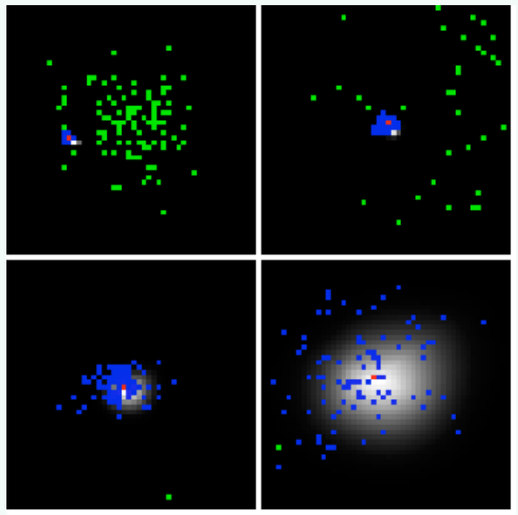
Likelihood-free or simulation-based models.

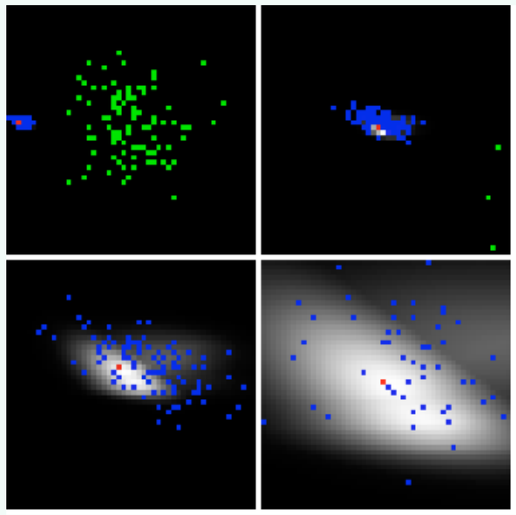
Real Posteriors

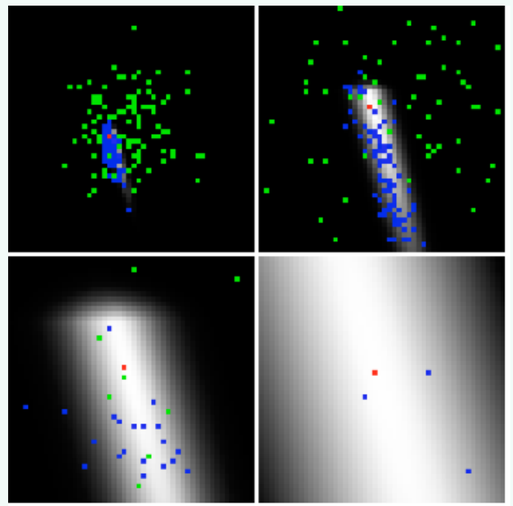
Require flexible approximations for the types of posteriors we are likely to see.



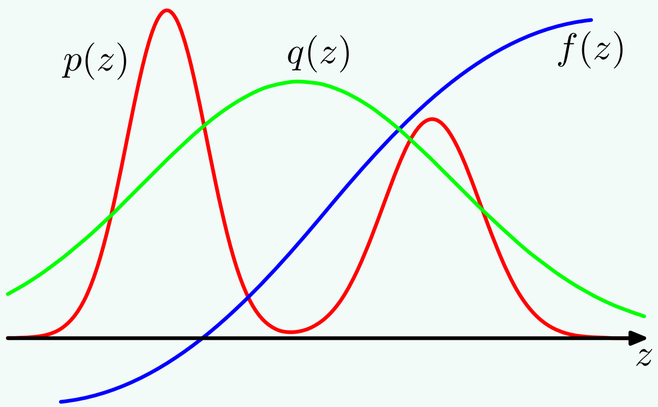








Importance Sampling



Notation

Always think of *q(**z|x**)* but often will write *q(**z**)* for simplicity.

Conditions

- *q(z|x)*>0, when *f(z)p(z)* ≠ 0.
- Easy to sample from *q(z)*.

Integral problem

$$p(\mathbf{x}) = \int p(\mathbf{x}|\mathbf{z})p(\mathbf{z})d\mathbf{z}$$

Proposal

$$p(\mathbf{x}) = \int p(\mathbf{x}|\mathbf{z})p(\mathbf{z})\frac{q(\mathbf{z})}{q(\mathbf{z})}d\mathbf{z}$$

Importance Weight

$$p(\mathbf{x}) = \int p(\mathbf{x}|\mathbf{z})\frac{p(\mathbf{z})}{q(\mathbf{z})}q(\mathbf{z})d\mathbf{z}$$

$$w^{(s)} = \frac{p(z)}{q(z)} \quad z^{(s)} \sim q(z)$$

Monte Carlo

$$p(\mathbf{x}) = \frac{1}{S}\sum_s w^{(s)}p(\mathbf{x}|\mathbf{z}^{(s)})$$

$$\log p(\mathbf{x}) = \log \sum_s w^{(s)}p(\mathbf{x}|\mathbf{z}^{(s)}) - \log S$$

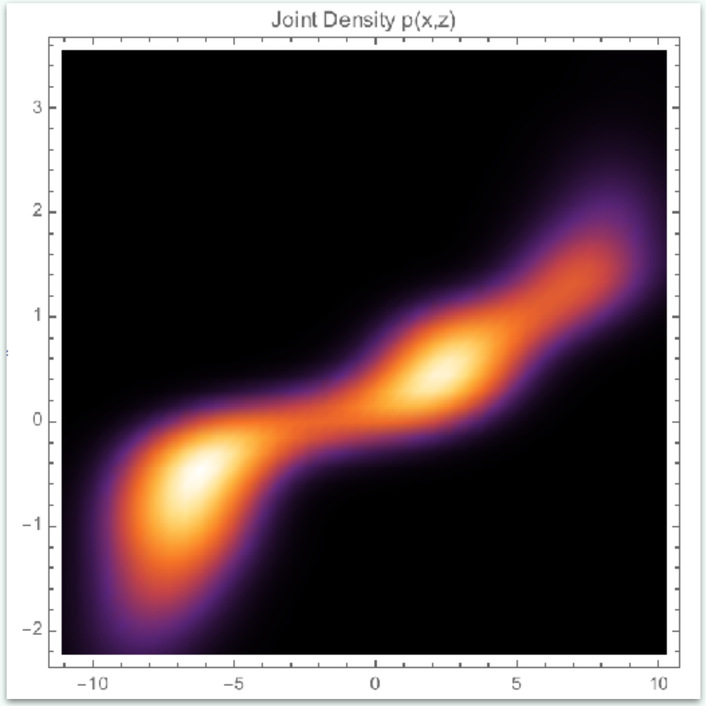
Bounds in Expectation

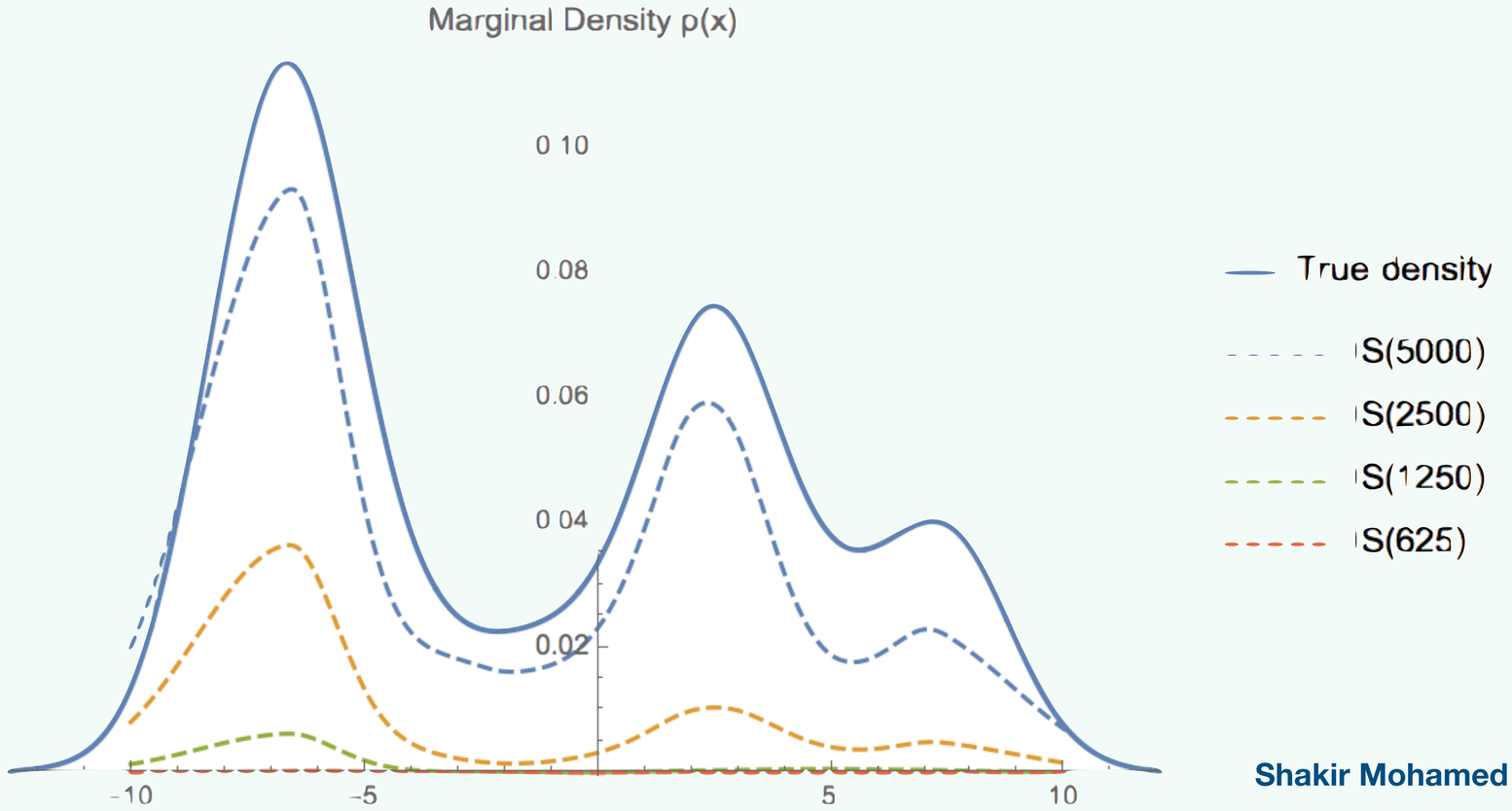
Jensen's Inequality

$$\log \int p(x)g(x)dx \geq \int p(x) \log g(x)dx$$

IS Expectation

$$\log p(\mathbf{x}) \geq \mathbb{E}_{q(\mathbf{z})} \left[\log \sum_s w^{(s)} p(\mathbf{x}|\mathbf{z}^{(s)}) \right] - \log S$$





IS to Variational Inference

Integral problem

$$p(\mathbf{x}) = \int p(\mathbf{x}|\mathbf{z})p(\mathbf{z})d\mathbf{z}$$

Proposal

$$p(\mathbf{x}) = \int p(\mathbf{x}|\mathbf{z})p(\mathbf{z})\frac{q(\mathbf{z})}{q(\mathbf{z})}d\mathbf{z}$$

Importance Weight

$$p(\mathbf{x}) = \int p(\mathbf{x}|\mathbf{z})\frac{p(\mathbf{z})}{q(\mathbf{z})}q(\mathbf{z})d\mathbf{z}$$

Jensen's inequality

$$\log \int p(x)g(x)dx \geq \int p(x)\log g(x)dx$$

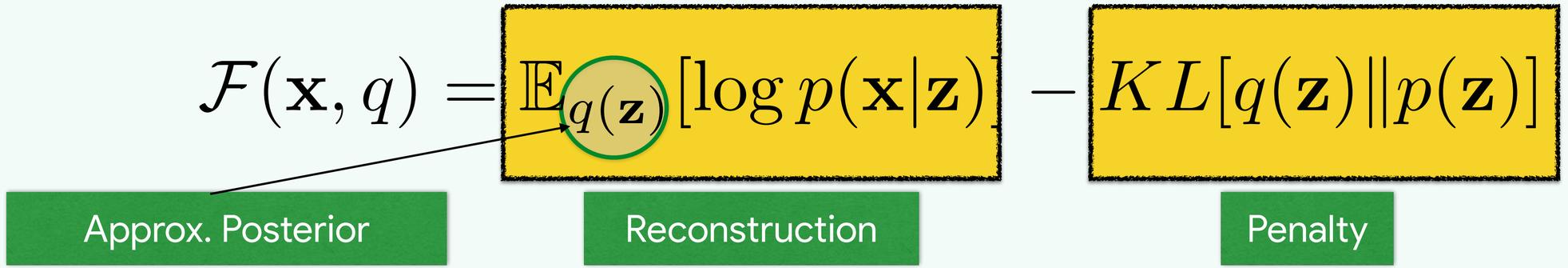
$$\log p(\mathbf{x}) \geq \int q(\mathbf{z}) \log \left(p(\mathbf{x}|\mathbf{z})\frac{p(\mathbf{z})}{q(\mathbf{z})} \right) d\mathbf{z}$$

$$= \int q(\mathbf{z}) \log p(\mathbf{x}|\mathbf{z}) - \int q(\mathbf{z}) \log \frac{q(\mathbf{z})}{p(\mathbf{z})}$$

Variational lower bound

$$\mathbb{E}_{q(\mathbf{z})}[\log p(\mathbf{x}|\mathbf{z})] - KL[q(\mathbf{z})||p(\mathbf{z})]$$

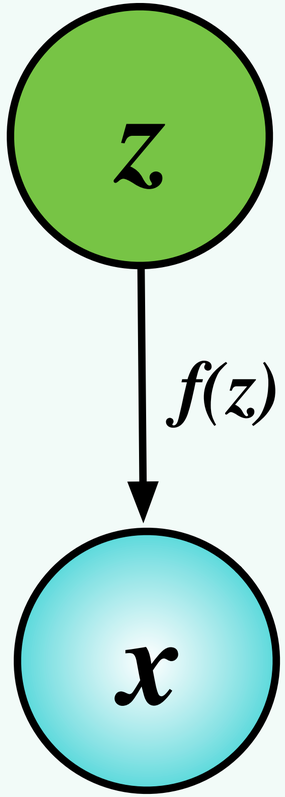
Variational Bound



Some comments on *q*:

- **Integration is now optimisation**: optimise for *q(z)* directly.
  - I write *q(z)* to simplify the notation, but it depends on the data, *q(z|x)*.
  - *Easy convergence assessment* since we wait until the free energy (loss) reaches convergence.
- **Variational parameters:** parameters of *q(z)*
  - E.g., if a Gaussian, variational parameters are mean and variance.

Latent Gaussian Models



Probabilistic Model $z \sim \mathcal{N}(z|0,1) \qquad y \sim p(y|f_\theta(z))$

Mean-field approx $q(z) = \prod_i \mathcal{N}(z_i|\mu_i, \sigma_i^2)$

Variational bound $\mathcal{F}(y,q) = \mathbb{E}_{q(z)}[\log p(y|z)] - KL[q(z)\|p(z)]$

$$\mathcal{F}(y,q) = \mathbb{E}_{q(z)}[\log p(y|z)] - \sum_i KL[q(z_i)\|p(z_i)]$$

$$\mathcal{F}(y,q) = \mathbb{E}_{q(z)}[\log p(y|z)] - \sum_i KL[\mathcal{N}(z_i|\mu_i, \sigma_i^2)\|\mathcal{N}(z_i|0,1)]$$

$$\mathcal{F}(y,q) = \mathbb{E}_{q(z)}[\log p(y|f_\theta(z))] - \frac{1}{2}\sum_i (\sigma_i^2 + \mu_i^2 - 1 - \ln \sigma_i^2)$$

Variational Optimisation

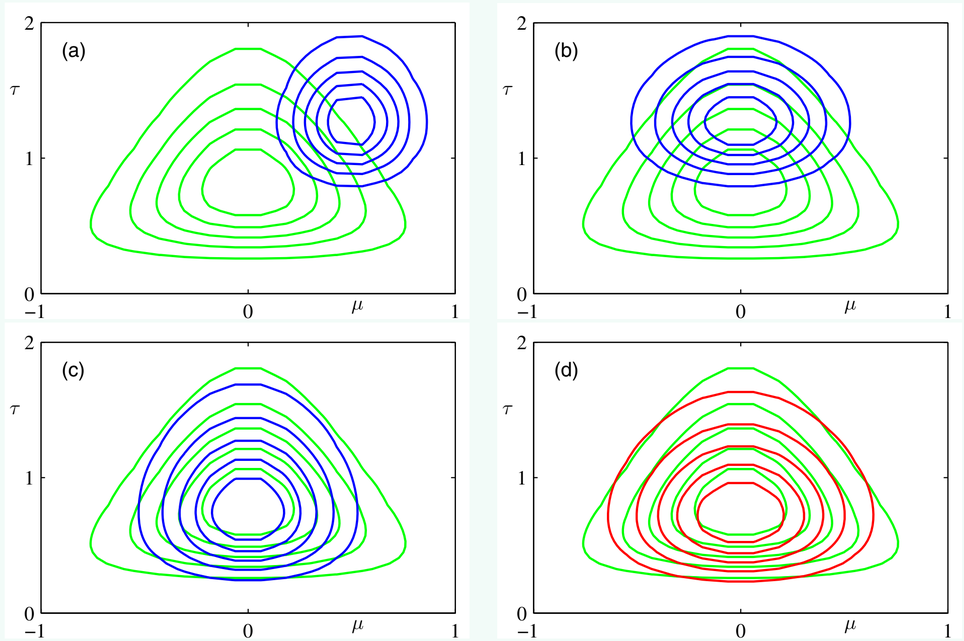
$$\mathcal{F}(\mathbf{x}, q) = \mathbb{E}_{q(\mathbf{z})}[\log p(\mathbf{x}|\mathbf{z})] - KL[q(\mathbf{z})||p(\mathbf{z})]$$

Approx. Posterior

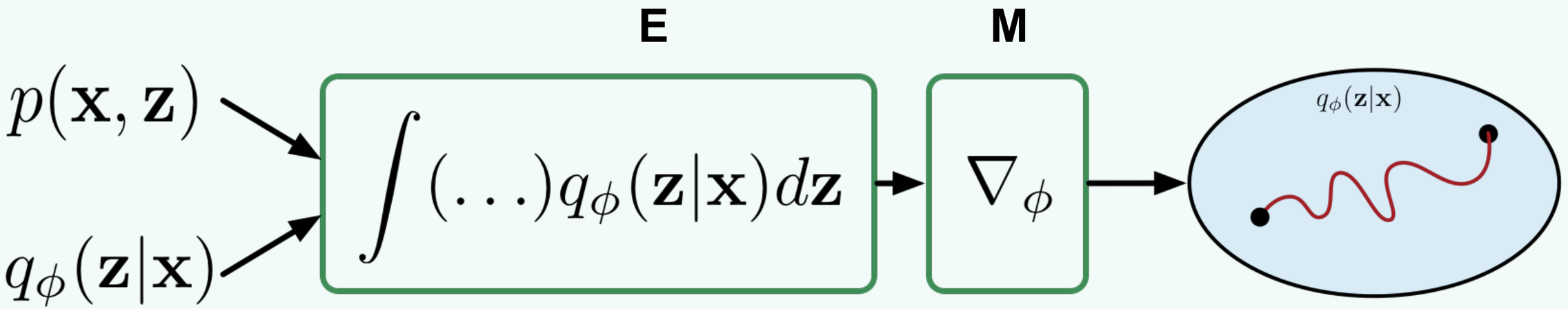
Reconstruction

Penalty

- Variational EM
- Stochastic Variational Inference
- Doubly Stochastic Variational Inference
- Amortised Inference



Classical Inference Approach



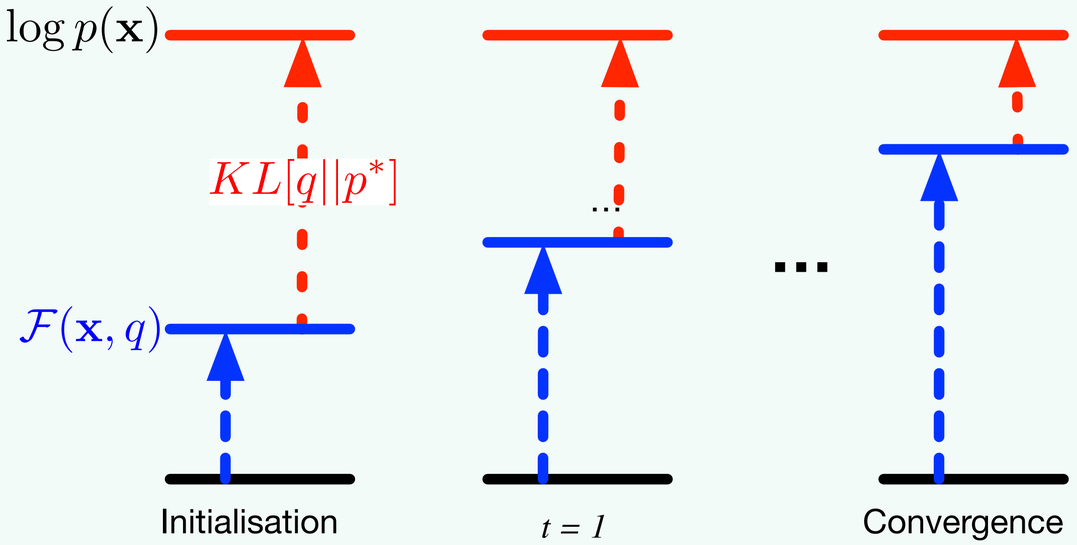
Compute Expectations then M-step gradients

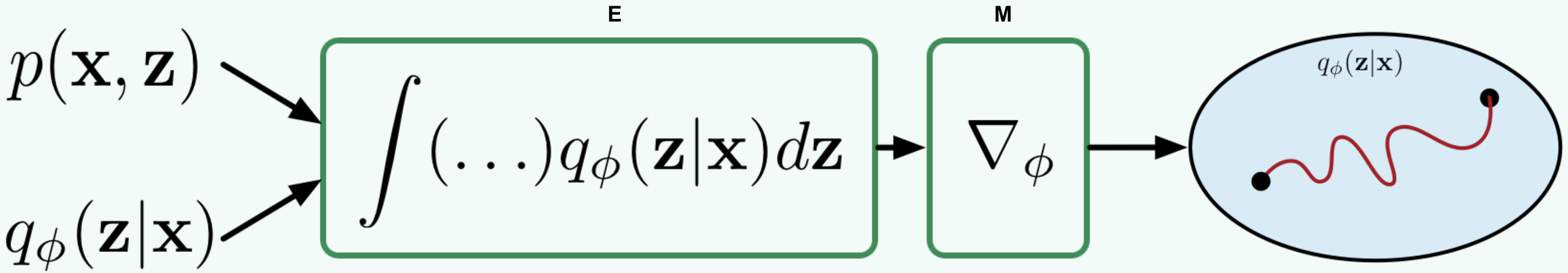
Variational EM

$$\mathcal{F}(\mathbf{x}, q) = \mathbb{E}_{q(\mathbf{z})}[\log p(\mathbf{x}|\mathbf{z})] - KL[q(\mathbf{z})||p(\mathbf{z})]$$

Repeat:

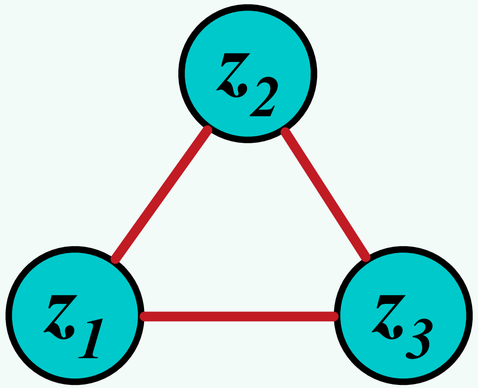
E-step	$\phi \propto \nabla_\phi \mathcal{F}(\mathbf{x}, q)$	*Var. params*
M-step	$\theta \propto \nabla_\theta \mathcal{F}(\mathbf{x}, q)$	*Model params*



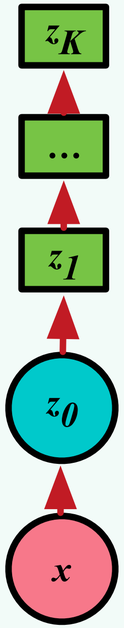


Families of Posterior Approximations

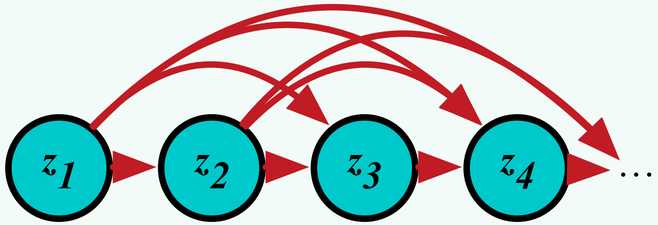
True Posterior



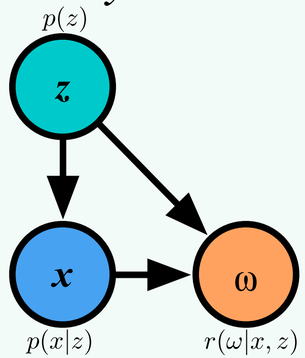
Normalising flows



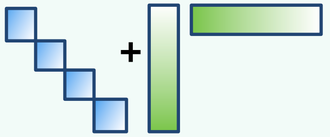
Structured mean-field



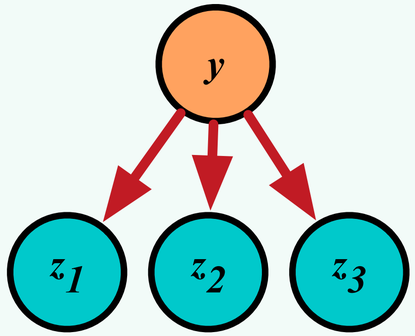
Auxiliary variables



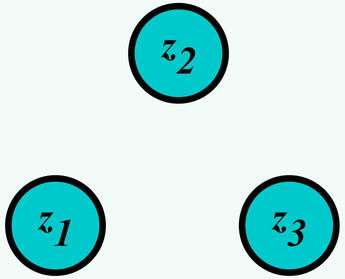
Covariance models



Mixtures



Fully-factorised

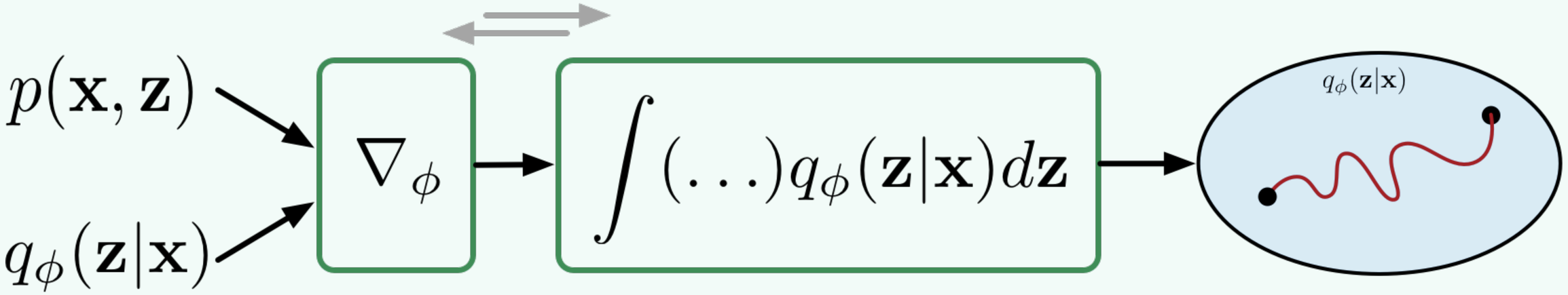


Most Expressive ⟵ ⟶ ***Least Expressive***

$$q^*(z|x) \propto p(x|z)p(z)$$

$$q_{MF}(z|x) = \prod_k q(z_k)$$

Stochastic Inference Approach

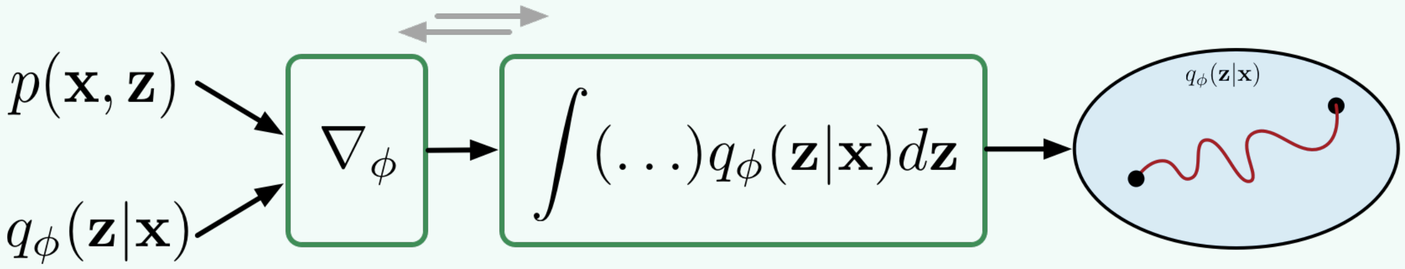


In general, we won't know the expectations.

Gradient is of the parameters of the distribution w.r.t. which the expectation is taken.

Stochastic Gradients

$$\nabla_\phi \mathbb{E}_{q_\phi(\mathbf{z})}[f_\theta(\mathbf{z})] = \nabla \int q_\phi(\mathbf{z}) f_\theta(\mathbf{z}) d\mathbf{z}$$

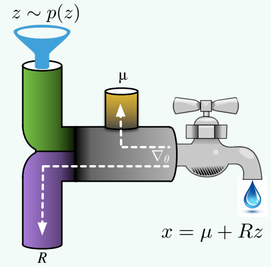


Doubly stochastic estimators

Pathwise Estimator

When easy to use transformation is available and differentiable function *f*.

$$\mathbb{E}_{p(\epsilon)}[\nabla_\phi f_\theta(g(\epsilon, \phi))]$$
$$z \sim q_\phi(\mathbf{z})$$
$$\mathbf{z} = g(\epsilon, \phi) \quad \epsilon \sim p(\epsilon)$$



Score-function estimator

When function *f* non-differentiable and *q(z)* is easy to sample from.

$$\mathbb{E}_{q(z)}[f_\theta(\mathbf{z}) \nabla_\phi \log q_\phi(\mathbf{z}))]$$

Log-derivative Trick

Score function is the derivative of a log-likelihood function.

$$\nabla_\phi \log q_\phi(\mathbf{z}) = \frac{\nabla_\phi q_\phi(\mathbf{z})}{q_\phi(\mathbf{z})}$$

Several useful properties

Expected score

$$\mathbb{E}_{q(z)}\left[\nabla_\phi \log q_\phi(\mathbf{z})\right] = 0$$

«Show this

$$\mathbb{E}_{q(z)}[\nabla_\phi \log q_\phi(\mathbf{z})] = \int q(z) \frac{\nabla_\phi q_\phi(\mathbf{z})}{q_\phi(\mathbf{z})} = \nabla \int q_\phi(\mathbf{z}) = \nabla 1 = 0$$

Fisher Information

$$\mathbb{V}[\nabla_\theta \log p(\mathbf{x};\theta)] = \mathcal{I}(\theta) = \mathbb{E}_{p(x;\theta)}[\nabla_\theta \log p(\mathbf{x};\theta) \nabla_\theta \log p(\mathbf{x};\theta)^\top]$$

Score Function Gradient

$$\nabla_\phi \mathbb{E}_{q_\phi(\mathbf{z})}[f_\theta(\mathbf{z})] = \nabla \int q_\phi(\mathbf{z}) f_\theta(\mathbf{z}) d\mathbf{z}$$

Leibnitz integral rule

$$= \int \frac{q_\phi(\mathbf{z})}{q_\phi(\mathbf{z})} \nabla_\phi q_\phi(\mathbf{z}) f(\mathbf{z}) d\mathbf{z}$$

Identity

$$= \int q_\phi(\mathbf{z}) \nabla_\phi \log q_\phi(\mathbf{z}) f(\mathbf{z}) d\mathbf{z}$$

Log-deriv

$$= \mathbb{E}_{q_\phi(\mathbf{z})} \left[f(\mathbf{z}) \nabla_\phi \log q_\phi(\mathbf{z}) \right]$$

Gradient

$$\nabla_\phi \mathbb{E}_{q_\phi(\mathbf{z})}[f_\theta(\mathbf{z})] = \mathbb{E}_{q_\phi(\mathbf{z})} \left[(f(\mathbf{z}) - c) \nabla_\phi \log q_\phi(\mathbf{z}) \right]$$

Control Variate

Other names

- Likelihood ratio method
- REINFORCE and policy gradients
- Automated & Black-box inference

When to use

- Function is not differentiable, not analytical.
- Distribution *q* is easy to sample from.
- Density *q* is known and differentiable.

Amortised Inference

Repeat:

E-step (compute q)

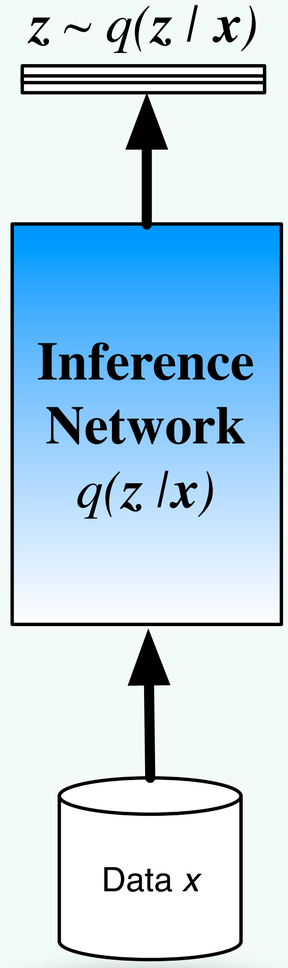
For i = 1, ... N

$$\phi_n \propto \nabla_\phi \mathbb{E}_{q_\phi(z)}[\log p_\theta(\mathbf{x}_n|z_n)] - \nabla_\phi KL[q(z_n)||p(z)]$$

Instead of solving for every observation, amortise using a model.

M-step

$$\theta \propto \frac{1}{N} \sum_n \mathbb{E}_{q_\phi(z)}[\nabla_\theta \log p_\theta(\mathbf{x}_n|z_n)]$$



- **Inference network**: q is an ***encoder,*** an ***inverse*** model, ***recognition model***.
- Parameters of q are now a set of *global parameters* used for inference of all data points - test and train.
- **Amortise (spread) the cost of inference over all data.**
- Joint optimisation of variational and model parameters.

Inference networks provide an efficient mechanism for **posterior inference with memory**

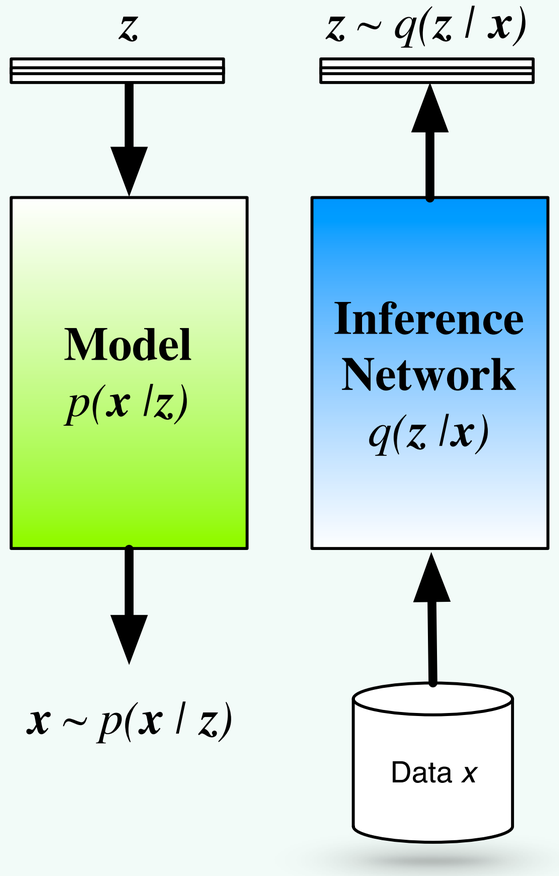
Variational Autoencoder

$$\mathcal{F}(\mathbf{x}, q) = \mathbb{E}_{q(\mathbf{z})}[\log p(\mathbf{x}|\mathbf{z})] - KL[q(\mathbf{z})\|p(\mathbf{z})]$$

Approx. Posterior

Reconstruction

Penalty

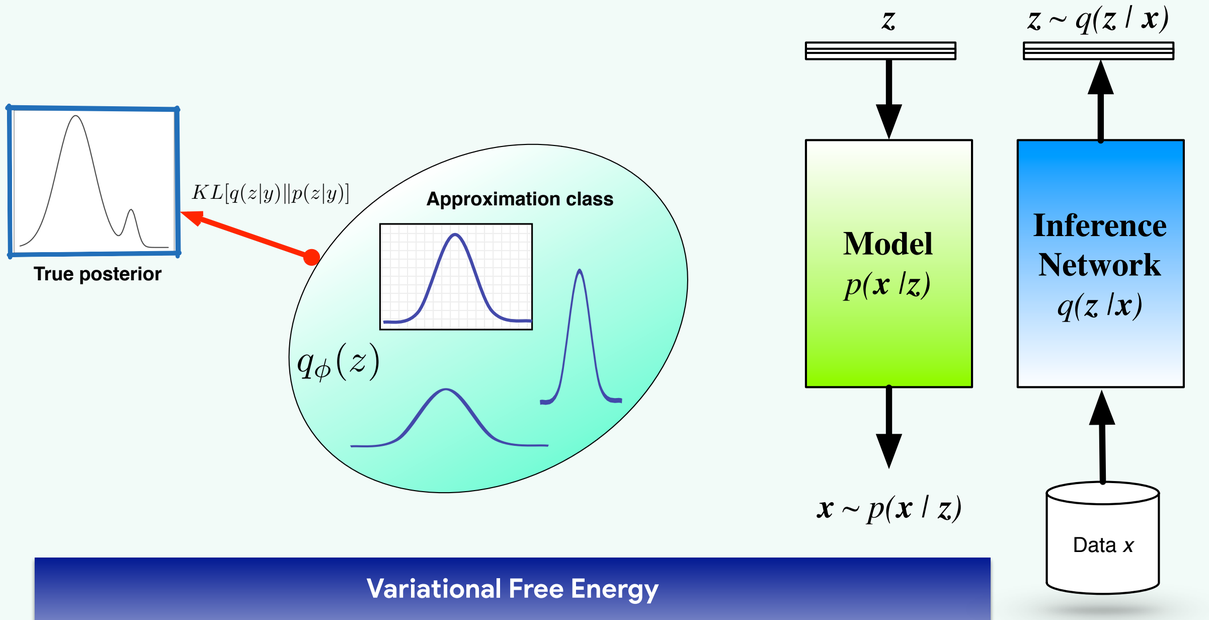


Stochastic encoder-decoder system to implement variational inference.

Specific combination of **variational inference** in **latent variable models** using **inference networks**
Variational Auto-encoder

But don't forget what your model is, and what inference you use.

Variational Methods



Variational Free Energy

$$\mathcal{F}(\mathbf{x}, q) = \mathbb{E}_{q(\mathbf{z})}[\log p(\mathbf{x}|\mathbf{z})] - KL[q(\mathbf{z})||p(\mathbf{z})]$$

Multi-sample Variational Objective

$$\mathcal{F}(\mathbf{x}, q) = \mathbb{E}_{q(z)}\left[\log \frac{1}{S}\sum_s \frac{p(\mathbf{z})}{q(\mathbf{z})} p(\mathbf{x}|\mathbf{z})\right]$$

Renyi Variational Objective

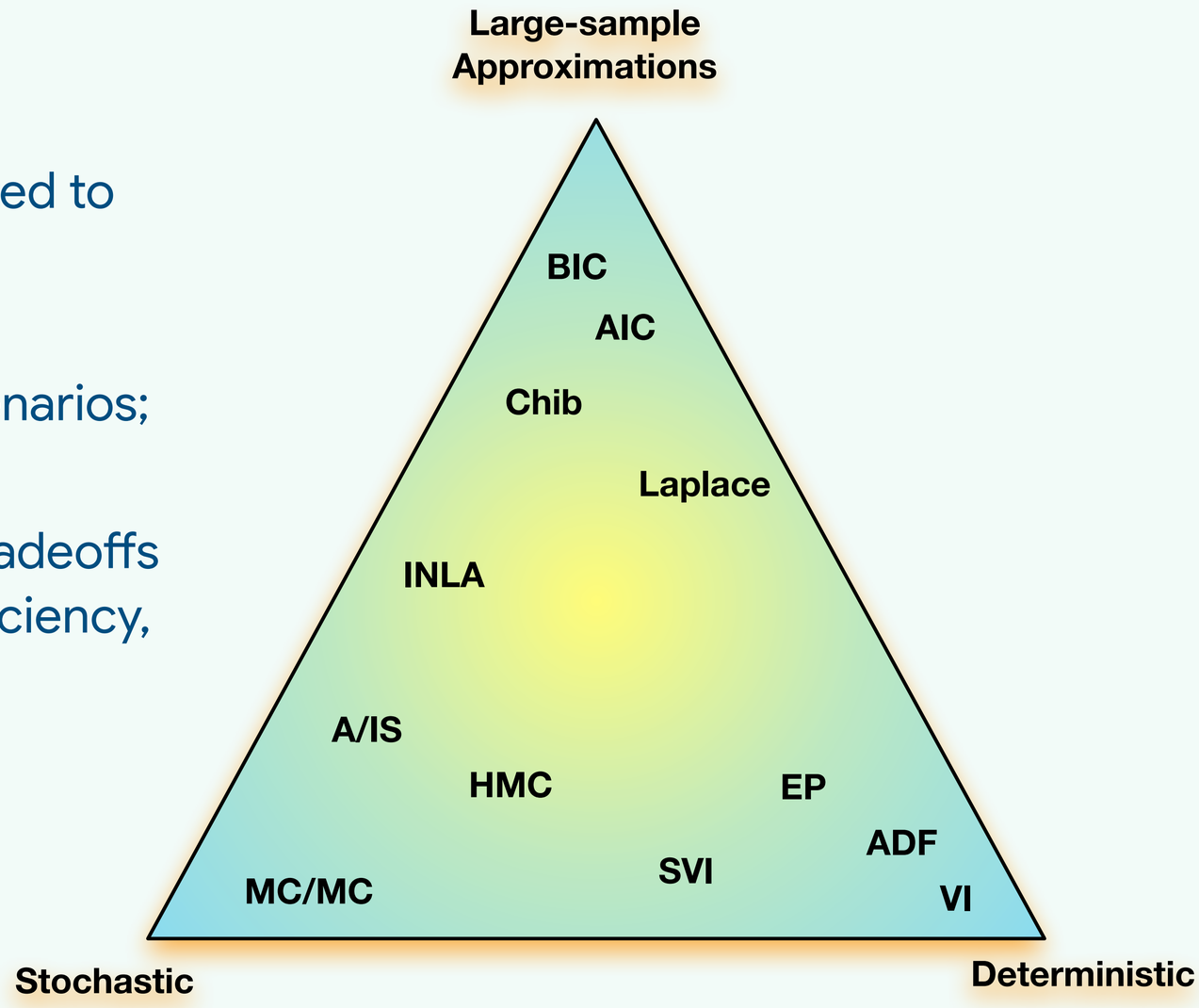
$$\mathcal{F}(\mathbf{x}, q) = \frac{1}{1-\alpha}\mathbb{E}_{q(z)}\left[\left(\log \frac{1}{S}\sum_s \frac{p(\mathbf{z})}{q(\mathbf{z})} p(\mathbf{x}|\mathbf{z})\right)^{1-\alpha}\right]$$

Limitations and Considerations

- 'Biased' - you can never have the true posterior, only an approximation.
- It is hard to design rich families of approximate posteriors, although we have other methods like Normalising Flows, Mixtures, Auxiliary Variable methods.
- Optimisation can be challenging because of high-dimensionality (if being fully Bayesian).
- Tied to methods with known likelihood functions.
- Evaluation (especially in unsupervised cases is hard).

Bayesian Approximation

- Much of Bayesian analysis is dedicated to the computation of integrals.
- This is because we are interested in averaging over multiple possible scenarios; reasoning with uncertainty.
- Many other approaches that have tradeoffs in approximation, computational efficiency, bias, etc.



Colonialism

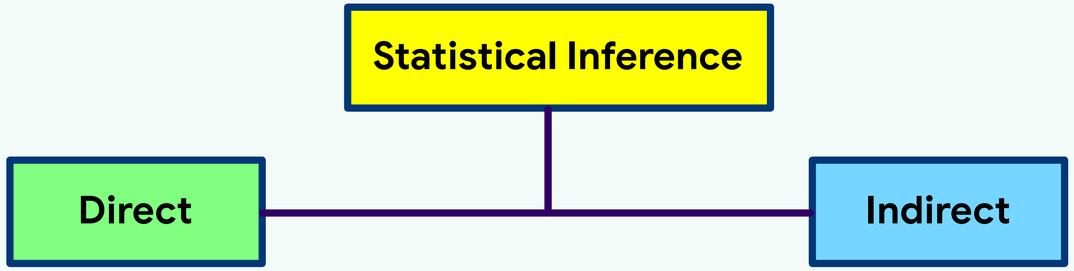
We are shaped by a colonial past: what/how science is done, in the language we speak in, in what we consider valid knowledge, and how we treat and relate to each other.

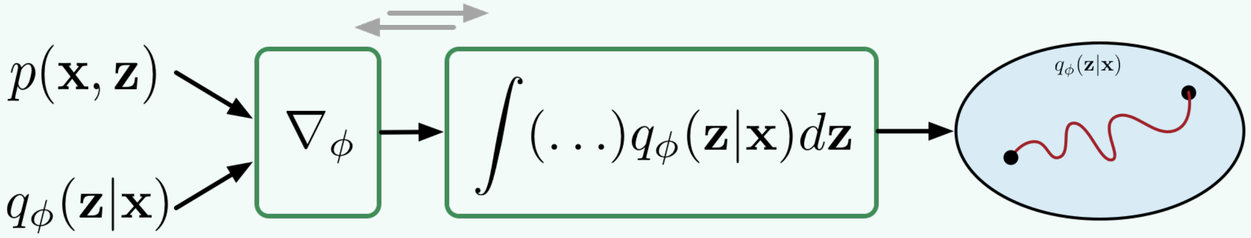
Statistics has been seen as a ***'moral science'***.

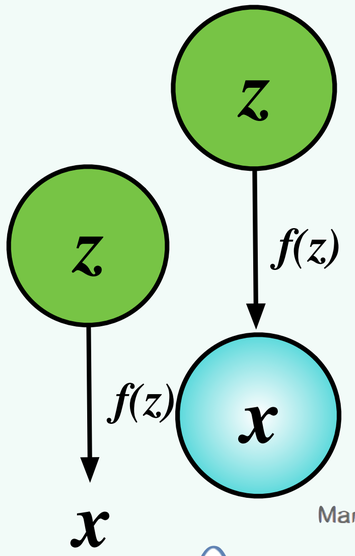
- Estimating the number of tanks in a war using Poisson conjugate models.
- Explain 'intelligence' using Factor analysis.
- Medical statistics and genetics were turned to create racialised differences between people.

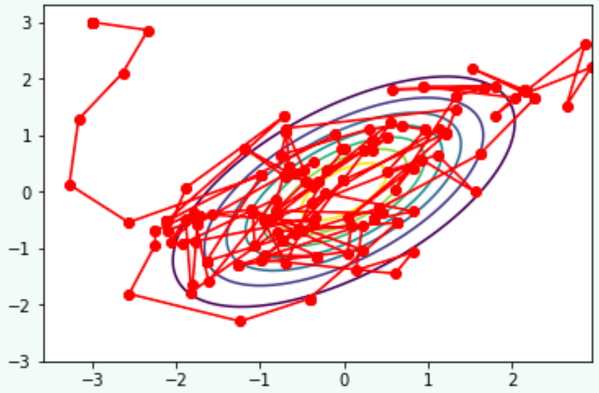
Consider the contextual and historical basis that is shaping our work. Leads to harms, and obscured views of what and who is successful.

Be more critical and skeptical.

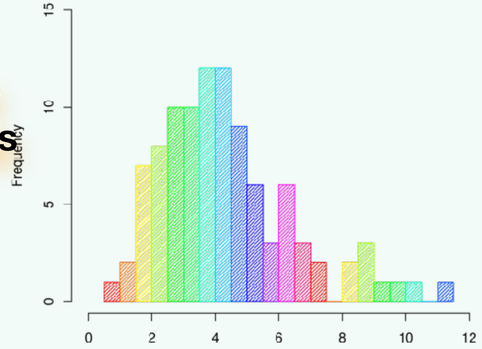


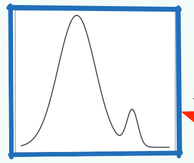




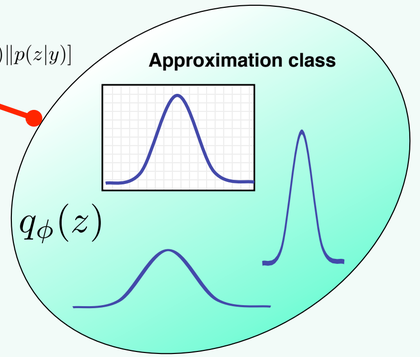


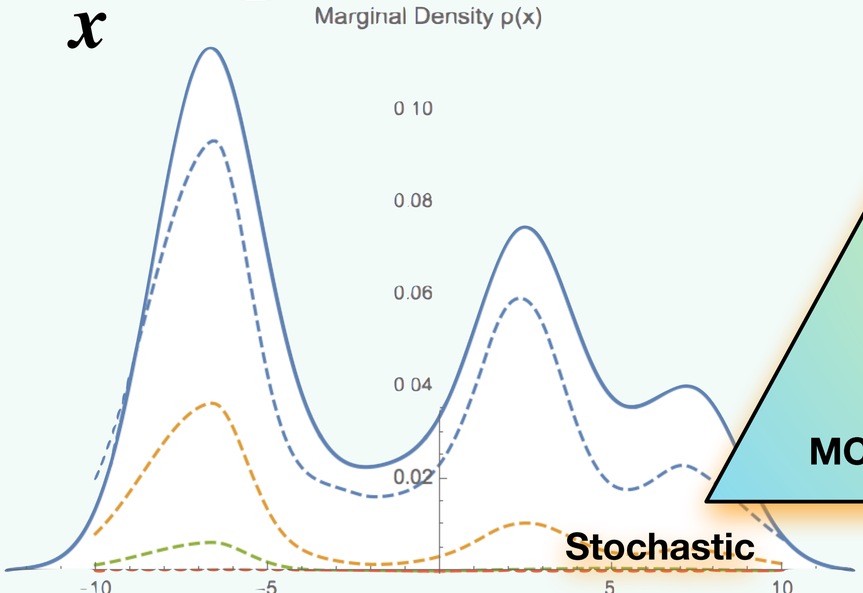
Large-sample Approximations





$KL[q(z|y)||p(z|y)]$

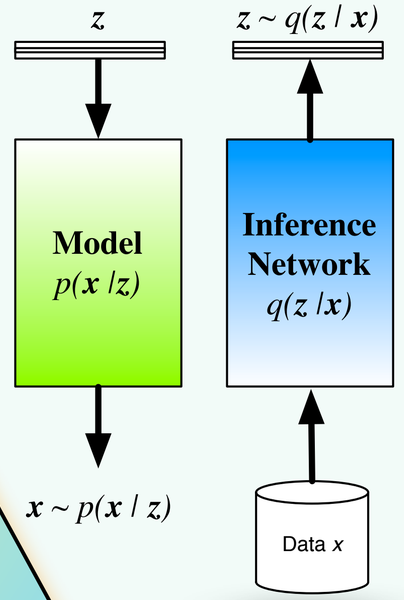


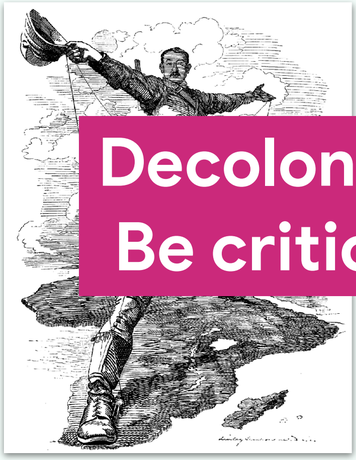


Stochastic

BIC
AIC
Chib
Laplace
INLA
A/IS
HMC
EP
ADF
SVI
MC/MC
VI

Deterministic





Decolonise
Be critical

Some papers and books

MacKay, David JC, and David JC Mac Kay. Information theory, inference and learning algorithms. Cambridge university press, 2003.

Robert, Christian, and George Casella. Monte Carlo statistical methods. Springer Science & Business Media, 2013.

Luc Devroye, Random variate generation in one line of code, Proceedings of the 28th conference on Winter simulation, 1996

Morokoff, William J., and Russel E. Caflisch. "Quasi-monte carlo integration." Journal of computational physics 122.2 (1995): 218-230.

O'Hagan, Anthony. "Monte Carlo is fundamentally unsound." The Statistician (1987): 247-249.

Blei, David M., Alp Kucukelbir, and Jon D. McAuliffe. "Variational inference: A review for statisticians." Journal of the American statistical Association 112.518 (2017): 859-877.

Beal, Matthew J. Variational algorithms for approximate Bayesian inference. Diss. UCL (University College London), 2003.

Kingma, Diederik P., and Max Welling. "Auto-encoding variational bayes." ICLR 2014

Lázaro-Gredilla, Miguel. "Doubly stochastic variational Bayes for non-conjugate inference." (2014).

Rezende, Danilo Jimenez, Shakir Mohamed, and Daan Wierstra. "Stochastic backpropagation and approximate inference in deep generative models." ICML 2014

Omiros Papaspiliopoulos, Gareth O Roberts, Martin Skold, A general framework for the parametrization of hierarchical models, Statistical Science, 2007

Mohamed, Shakir, et al. "Monte carlo gradient estimation in machine learning." JMLR, 2020.

Gershman, Samuel J., and Noah D. Goodman. "Amortized inference in probabilistic reasoning." In Proceedings of the 36th Annual Conference of the Cognitive Science Society. 2014.

Mohamed, Shakir, Png MT, Isaac, W, Decolonial AI: Decolonial Theory and Sociotechnical foresight in Artificial intelligence, Philosophy and technology, 2020.

10mins Break

Bayesian | Futures

Shakir Mohamed

@shakir_za

Outcomes

1. Inference and Decision-making
2. Optimisation and Reinforcement
3. Nonparametrics and Numerics

Inference and Decision-making

Inference

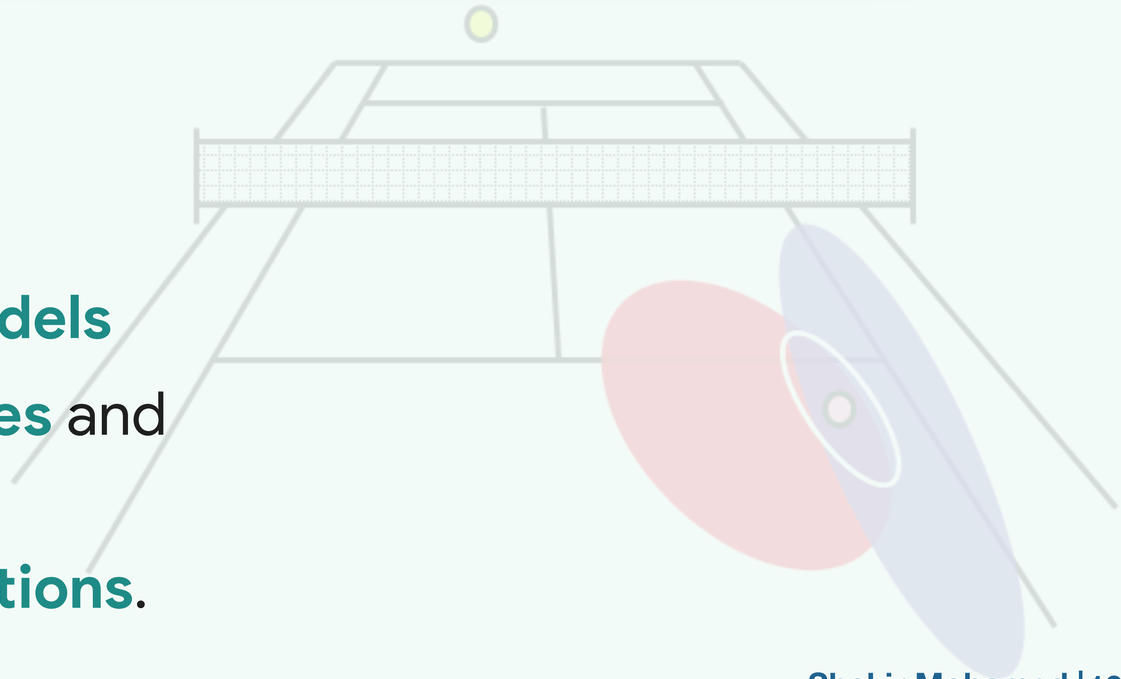
What we can *know* about our data

Decision-making

What we can *do* with our data.

Have many of the tools needed to build plausible reasoning systems:

1. Flexible ways of building rich **probabilistic models**
2. Ability to learn and **make consistent inferences** and maintain beliefs
3. Reason about potential outcomes and **take actions**.



Bayesian Reinforcement Learning

Environment as a generative process:

- An unknown likelihood;
- Not known analytically;
- Only able to observe its outcomes.

$$a \sim p(a)$$

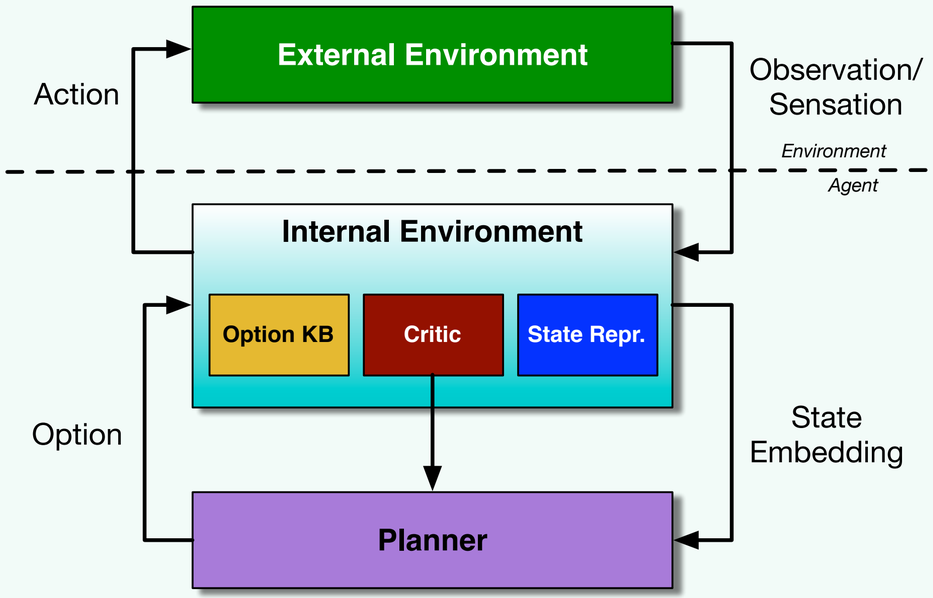
Prior over actions

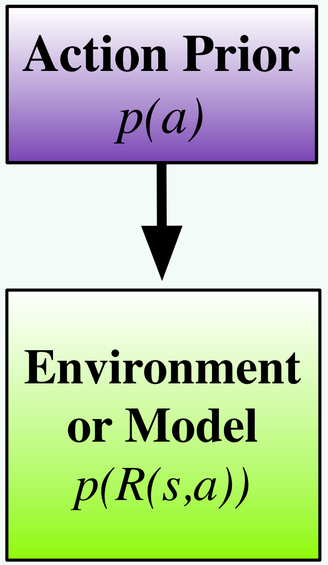
$$u(s,a) \sim \text{Environment}(a)$$

Interaction only

$$p(R(s)|a) \propto \exp(u(s,a))$$

Long-term reward





All the key inferential questions can now be asked in this simple framework.

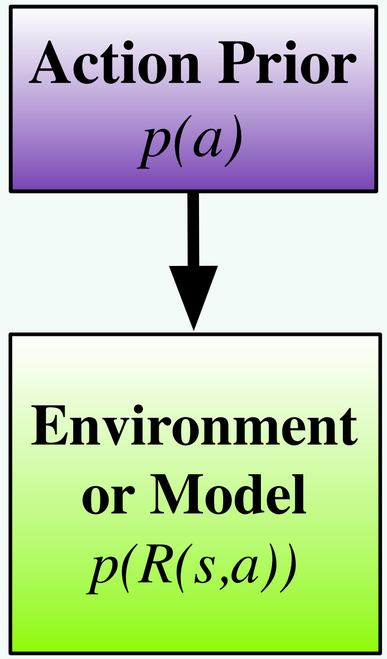
Planning-as-Inference

Simplest question

What is the posterior distribution over actions?
Maximising the probability of the return *log p(R)*.

Variational inference in the hierarchical model

$$\mathcal{F}(\theta) = \mathbb{E}_{\pi(\mathbf{a}|\mathbf{s})}[R(s,a)] - \mathrm{KL}[\pi_\theta(\mathbf{a}|s) \| p(\mathbf{a})]$$



Recover policy search methods:

- Uniform prior over distributions
- Continuous policy parameters
- Can evaluate environment, but not differentiate.

Policy Search

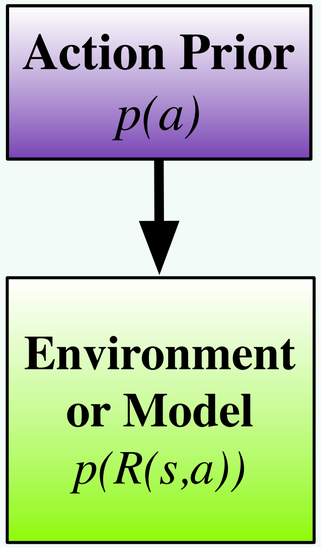
Free Energy

$$\mathcal{F}(\theta) = \mathbb{E}_{\pi(\mathbf{a}|\mathbf{s})}[R(s,a)] - \mathrm{KL}[\pi_\theta(\mathbf{a}|s)\|p(\mathbf{a})]$$

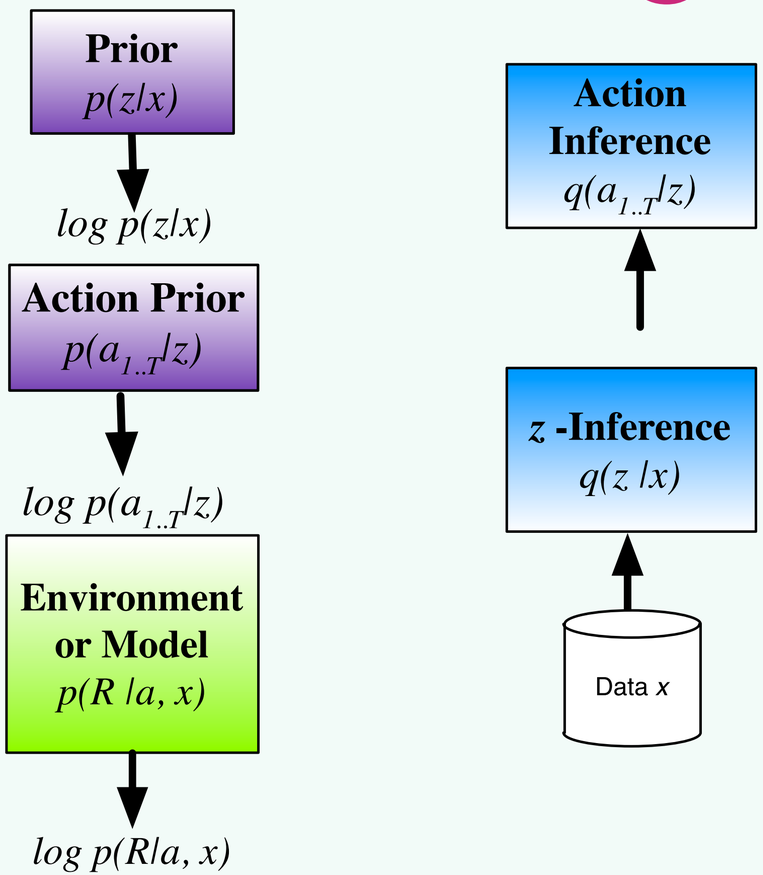
Policy gradient using score-function gradient

$$\nabla_\theta \mathcal{F}(\theta) = \mathbb{E}_{\pi(\mathbf{a}|\mathbf{s})}[(R(s,a) - c)\nabla_\theta \log \pi_\theta(\mathbf{a}|\mathbf{s})] + \nabla_\theta \mathbb{H}[\pi_\theta(\mathbf{a}|s)]$$

- Appearance of the entropy penalty is natural and alternative priors easy to consider and introduces aspect of exploration in a natural way.
- Can easily incorporate prior knowledge of the action space.
- Use any of the tools of probabilistic inference available.
- Easily handle stochastic and deterministic policies.



Hierarchical Planning



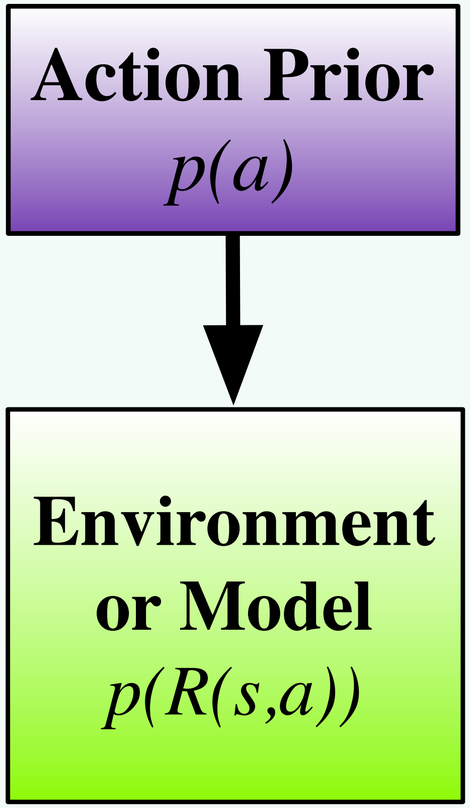
Variational MDP

$$\mathcal{F}^{\pi}(\theta) = \mathbb{E}_{q(a,z|x)}[R(a|x)] - \alpha KL[q_\theta(\mathbf{z}|\mathbf{x})\|p(\mathbf{z}|\mathbf{x})] + \alpha\mathbb{H}[\pi_\theta(\mathbf{a}|\mathbf{z})]$$

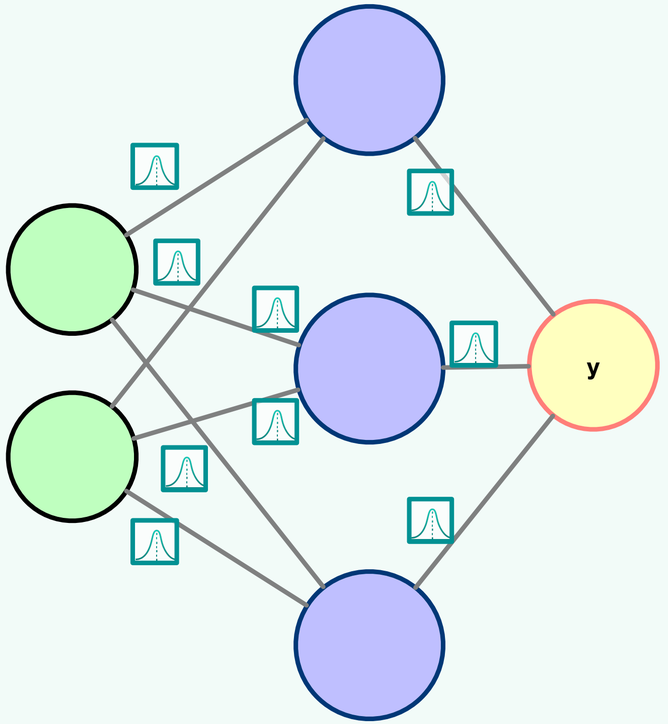
Bayesian RL and Control

With a more realistic expansion as graphical model

- Derive Bellman's equation as a different writing of message passing.
- Application of the EM algorithm for policy search becomes possible.
- Easily consider other variational methods, like EP.
- Both model-free and model-based methods emerge.



Bayesian Deep Learning

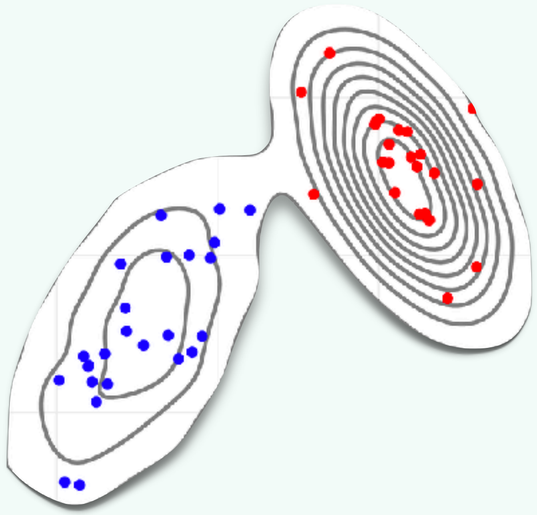


Prior $p(\boldsymbol{\theta}) = \mathcal{N}(\boldsymbol{\theta}|0, \mathbf{I})$

Observation model $p(y|\mathbf{x}, \boldsymbol{\theta}) = Categorical(\pi(\mathbf{x}; \boldsymbol{\theta}))$

Combination of Bayesian methods and Deep Learning

- Pragmatic Bayesian approach: Infer posteriors only for a subset of parameters. E.g., introduce rank-1 components into the session
- Bayesian methods for very deep networks. Difficult.
- Lots of work in variational inference, MC dropout, variational dropout.
- Sampling from multimodal posteriors.
- Analysis of infinite width models using Neural Tangent Kernels.
- New MCMC methods, e.g., cyclical stochastic gradient MCMC

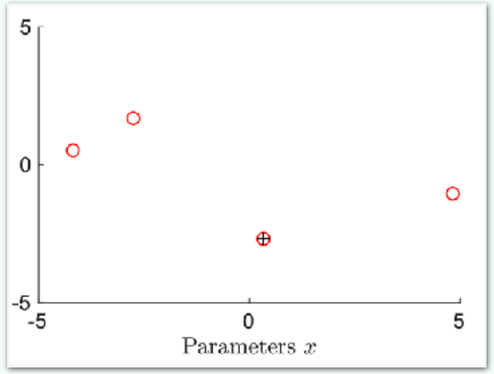


Bayesian Optimisation

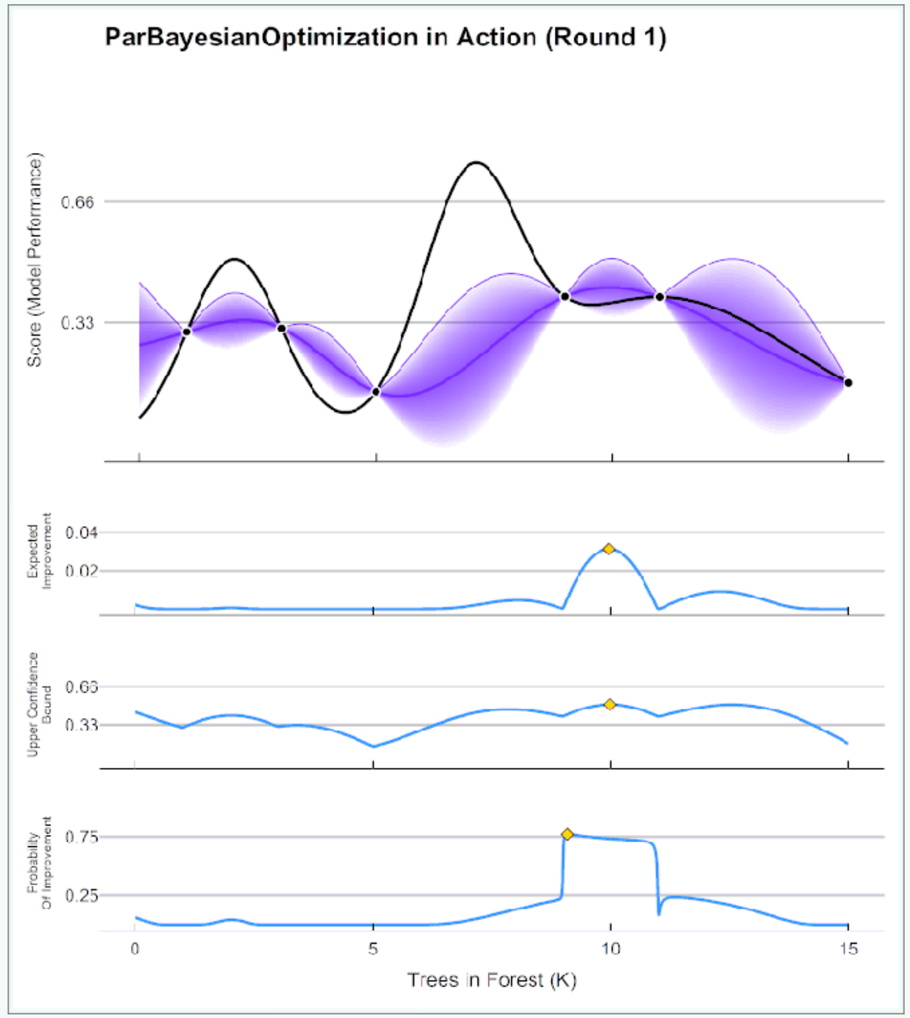
Global optimisation of a function $f(\mathbf{x})$ that is unknown, but that we can evaluate.

Bayesian approach is two fold:

- Place a prior over the unknown function and use evidence to learn the function by building a posterior.
- Use uncertainty from the posterior to decide where to next evaluate the function. Tradeoff exploration and exploitation.
- Recursive updating and optimise the acquisition function.



Prior	$f \sim \mathcal{GP}(0, \mathbf{K})$
Posterior	$p(f\|\mathcal{D}) = \mathcal{GP}(\mu_{f\|\mathcal{D}}, \mathbf{K}_{f\|\mathcal{D}})$
Acquisition fn.	$\alpha(x;\beta) = \mu(x) + \beta\sqrt{K_{xx}}$



Probabilistic Dualities

Basis Function Regression

$$f(\mathbf{x}) = \mathbf{w}^\top \phi(\mathbf{x};\boldsymbol{\theta}); \qquad \{\mathbf{w},\boldsymbol{\theta}\} \sim \mathcal{N}(0,\sigma^2\mathbf{I})$$

$$y = f(\mathbf{x}) + \epsilon; \qquad \epsilon \sim \mathcal{N}(0,\sigma_y^2)$$

Move from primal variables to dual variables

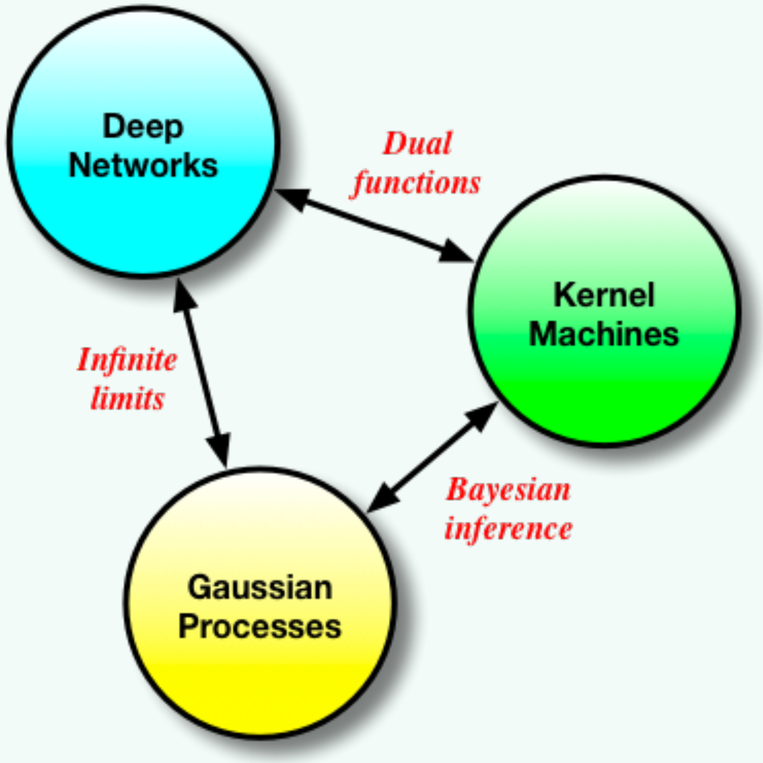
$$\mathcal{L}(f) = \tfrac{1}{2}\sum_n (y_n - f(\mathbf{x}))^2 + \tfrac{\lambda}{2}\|f\|_{\mathcal{H}}^2$$

Kernel trick and methods

Probability distributions over functions

$$p(f) = \mathcal{N}(0,\mathbf{K}) \quad p(y|f) = \mathcal{N}(f,\sigma^2)$$

Gaussian processes



Gaussian processes are priors over functions.

- They adapt to the number of data points, allow easy evaluation of posterior uncertainty,
- Have neat (but computationally challenging) updates.

Bayesian Nonparametrics

Fixed-dimension models

Basis function regression

Finite number of parameters

Adaptive-dimension models

Gaussian process regression

Apriori infinite number of parameters

Bayesian non-parametric models have a complexity that grows with the data
they provide flexible models, robust to overfitting.

De Finetti's theorem gives us some properties of these new types of priors and models.

$$p(x_1, \ldots, x_N) = \int \prod_{n=1}^{N} p(x_n \mid \theta) p(\theta) d\theta$$

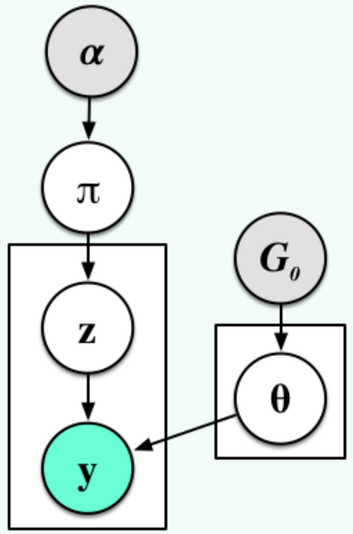
There is a parameter θ

The data $x_1, \ldots x_n$ is then conditionally independent

There is a likelihood $p(x \mid \theta)$

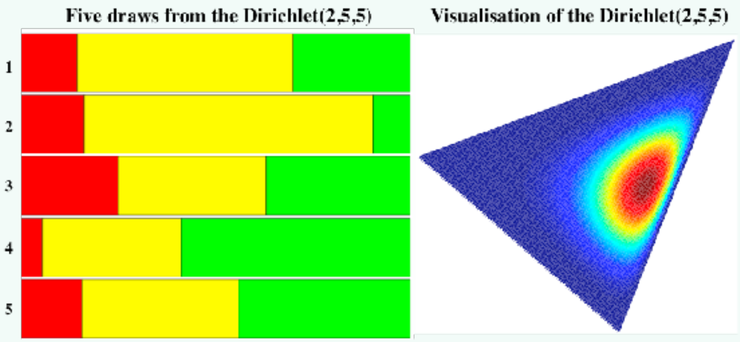
There is a distribution P on θ, if there's a density then a prior $p(\theta)$

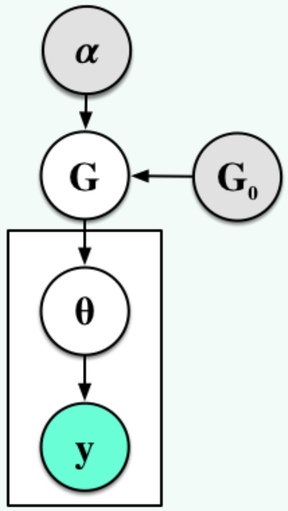
Mixture Models



Standard Generative View

Cluster parameters	$\phi_k \sim G_0 \qquad k = 1, \ldots, K$
Categories π	$\boldsymbol{\pi} \sim \mathcal{D}(\boldsymbol{\alpha}/K)$
Choose index	$z_n \sim Cat(\boldsymbol{\pi})$
Distribution at index	$\theta_n = \phi_{z_n}$
Generate data	$y \sim p(y\|\boldsymbol{\theta}_n)$



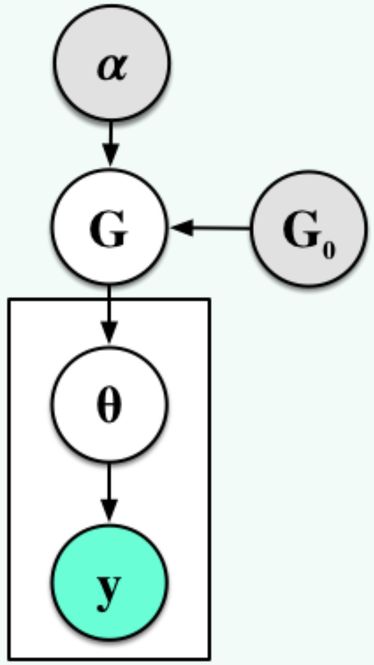


Random Measure View

Cluster parameters	$\phi_k \sim G_0 \qquad k = 1, \ldots, K$
Categories π	$\boldsymbol{\pi} \sim \mathcal{D}(\boldsymbol{\alpha}/K)$
Discrete mixture	$G = \sum_{k=1}^{K} \pi_k \delta_{\phi_k}$
Sample from mixture	$\theta_n \sim G$
Generate data	$y \sim p(y\|\boldsymbol{\theta}_n)$

What happens when K → ∞

Dirichlet Process Mixture



Random Measure View

$$\boldsymbol{\pi} \sim \mathcal{D}(\boldsymbol{\alpha}/K)$$

$$\boldsymbol{\phi}_k \sim G_0 \qquad k = 1, \ldots, K$$

$$G = \sum_{k=1}^{K} \pi_k \delta_{\phi_k}$$

$$\theta_n \sim G$$

$$y \sim p(y|\boldsymbol{\theta}_n)$$

What happens when K → ∞

$$y_n|G \sim \int p(y_n|\theta) dG(\theta)$$

$$G = \sum_{k=1}^{\infty} \pi_k \delta_{\phi_k} \sim DP(\alpha, G_0)$$

Stick-breaking representation of the Dirichlet,
Chinese restaurant process, urn schemes.

Repeat this idea for other models:

- Latent Variable models →Indian Buffet process
- Bayesian Linear regression → Gaussian process
- Time series → Temporal point process

Cascaded Indian Buffet process
Hierarchical Dirichlet process

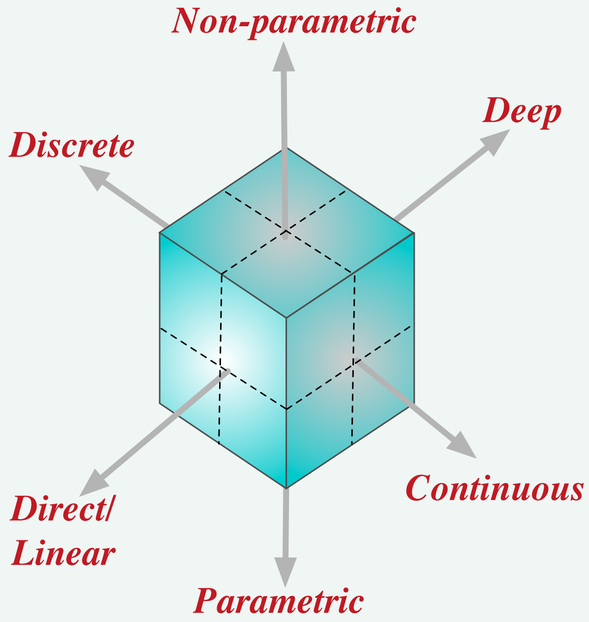
Deep Nonparametric Discrete

Sigmoid Belief Net
Deep auto-regressive networks (DARN)

Deep Parametric Discrete

Indian buffet process
Dirichlet process mixture

Direct Nonparametric Discrete



Deep Gaussian processes
Recurrent Gaussian Process
GP State space model

Deep Nonparametic Continuous

Hidden Markov Model
Discrete LVM
Sparse LVMs

Linear Parametric Discrete

Nonlinear factor analysis
Nonlinear Gaussian belief network
Deep Latent Gaussian (VAE, DRAW)

Deep Parametric Continuous

PCA, factor analysis
Independent components analysis
Gaussian LDS
Latent Gauss Field

Linear Parametric Continuous

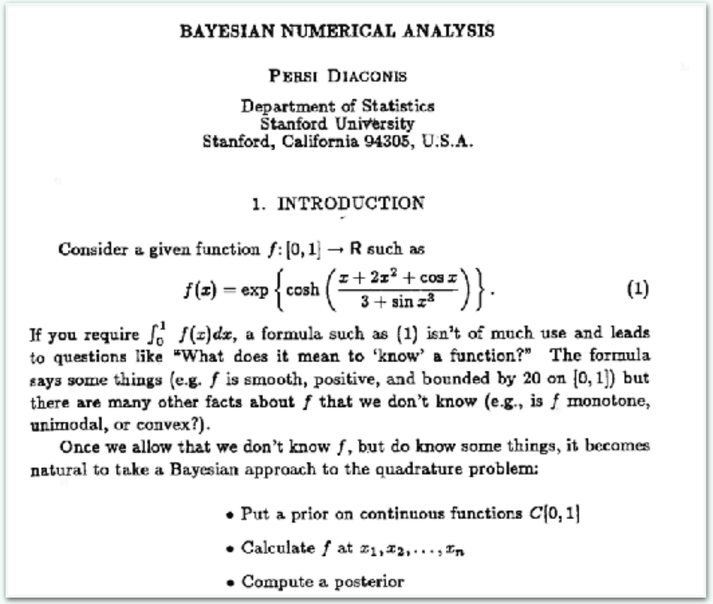
Gaussian process LVM

Direct Nonparametric Continuous

Bayesian Numerics

Think of numerical methods as (Bayesian) learning algorithms.

- Consider uncertainty in computation.
- Computational bottlenecks: optimisation, sampling, numerical integration, ODE solving, linear algebra, discretisation in PDEs, learning rates and line searches.



BAYESIAN NUMERICAL ANALYSIS

PERSI DIACONIS

Department of Statistics
Stanford University
Stanford, California 94305, U.S.A.

1. INTRODUCTION

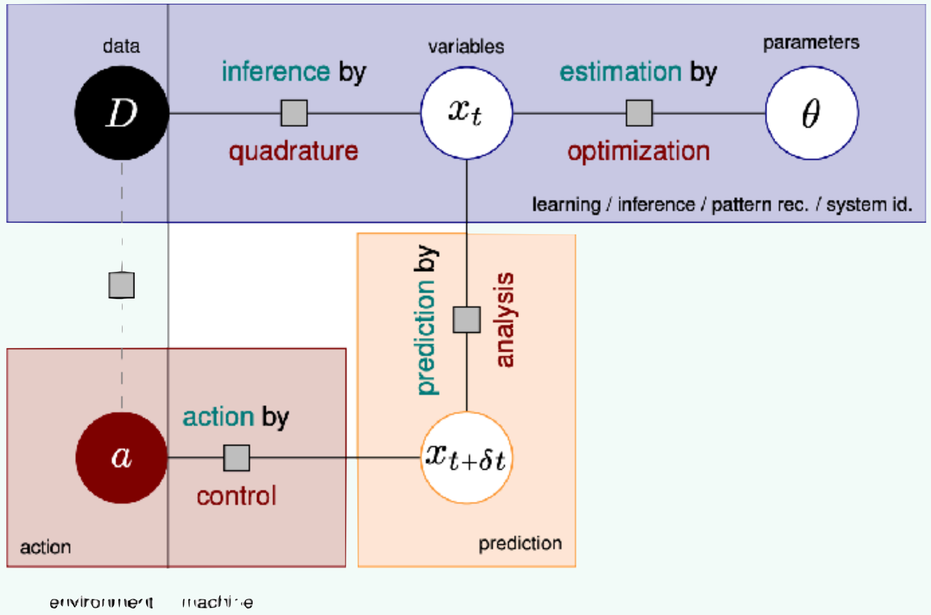
Consider a given function $f:[0,1] \to \mathbb{R}$ such as

$$f(x) = \exp\left\{\cosh\left(\frac{x + 2x^2 + \cos x}{3 + \sin x^3}\right)\right\}. \qquad (1)$$

If you require $\int_0^1 f(x)dx$, a formula such as (1) isn't of much use and leads to questions like "What does it mean to 'know' a function?" The formula says some things (e.g. f is smooth, positive, and bounded by 20 on $[0,1]$) but there are many other facts about f that we don't know (e.g., is f monotone, unimodal, or convex?).

Once we allow that we don't know f, but do know some things, it becomes natural to take a Bayesian approach to the quadrature problem:

- Put a prior on continuous functions $C[0,1]$
- Calculate f at $x_1, x_2, \ldots, x_n$
- Compute a posterior

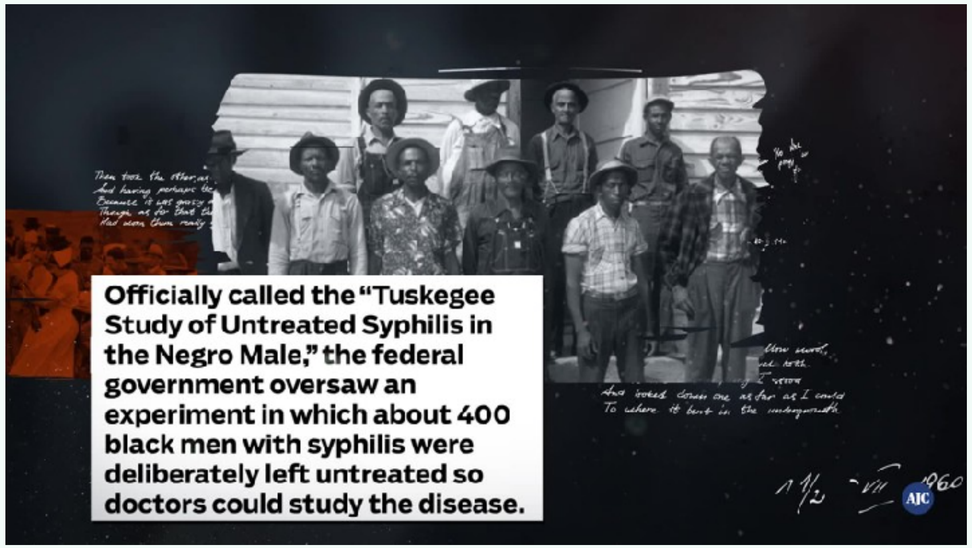


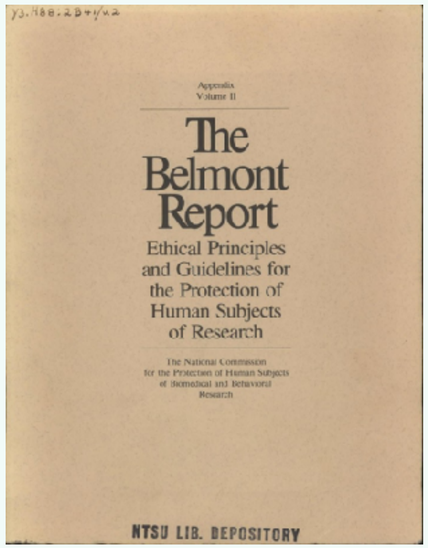
- Bayesian Monte Carlo/ Bayesian Quadrature
- Conjugate Gradients
- Line search
- Runge-Kutta ODE
- PDEs

Contextual Values

Nuremberg Code (1948)

1. The voluntary consent of the human subject is absolutely essential.
2. The experiment should be such as to yield fruitful results for the good of society, unprocurable by other methods or means of study, and not random and unnecessary in nature.
3. The experiment should be so designed and based on the results of animal experimentation and knowledge of the natural history of the disease or other problem under study that the anticipated results will justify the performance of the experiment.
4. The experiment should be so conducted as to avoid all unnecessary physical and mental suffering and injury.
5. No experiment should be conducted where there is an a priori reason to believe that death or disability injury will occur; except, perhaps, in those experiments where the experimental physicians also serve as subjects.
6. The degree of risk to be taken should never exceed that determined by the humanitarian importance of the problem to be solved by the experiment.
7. Proper preparations should be made and adequate facilities provided to protect the experimental subject against even remote possibilities of injury, disability, or death.
8. The experiment should be conducted only by scientifically qualified persons.
9. During the course of the experiment the human subject should be at liberty to bring the experiment to an end if he has reached the physical or mental state where continuation of the experiment seems to him to be impossible.
10. During the course of the experiment the scientist in charge must be prepared to terminate the experiment at any stage, if he has probable cause to believe that a continuation of the experiment is likely to result in injury, disability, or death to the experimental subject.

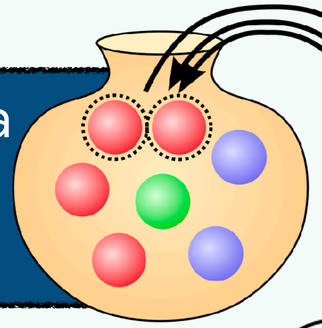




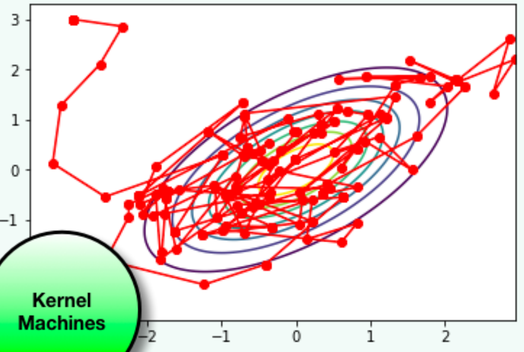
- **Respect-for-persons** - Individual Autonomy
- **Beneficence** - research designed to maximise societal benefit and minimise individual harm.
- **Justice** - research risks must be distributed across society.

- **Non-malfeasance** - Do no harm.
- **Explicability** - Explanation and Transparency.

Probability is a measure of the belief in a proposition **given** evidence.
A description of a state of knowledge.



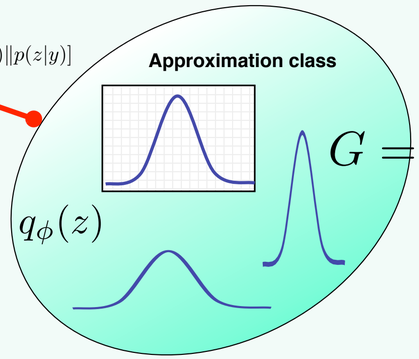
$$p(x_1, \ldots, x_n) = \int \prod_{i=1}^{N} p(x_i \mid \theta) P(d\theta)$$

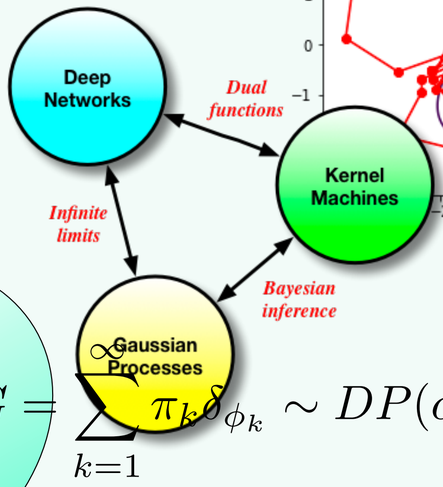


$$p(\mathbf{x}) = \int p(\mathbf{x}, \mathbf{z}) d\mathbf{z}$$

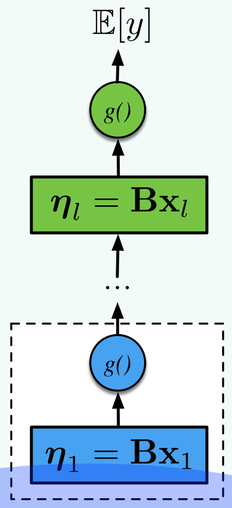
True posterior

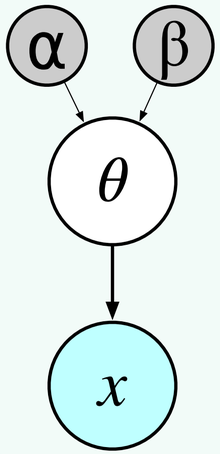
$KL[q(z|y)\|p(z|y)]$

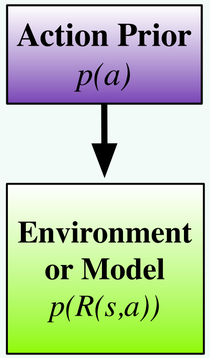


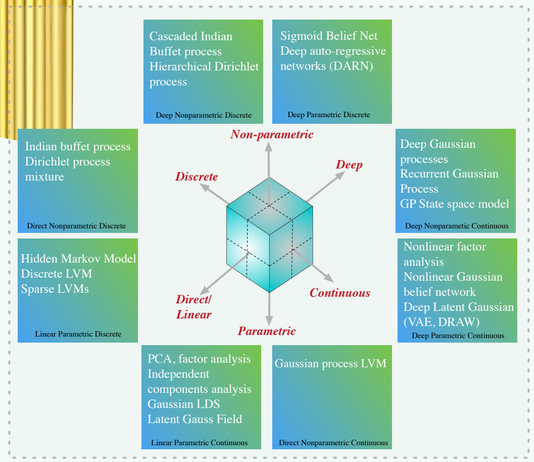


$$G = \sum_{k=1}^{\infty} \pi_k \delta_{\phi_k} \sim DP(\alpha, G_0)$$









Bayesian statistics, all quantities are probability distributions, so there is only the problem of **inference**.

Epistemic values

Contextual Values

Decolonise
Be critical

Neutrality
Decolonisation
Ethical Value Criteria

Some papers and books

Ghavamzadeh, Mohammad, et al. "Bayesian reinforcement learning: A survey." arXiv preprint arXiv:1609.04436 (2016).

Chentanez, Nuttapong, Andrew G. Barto, and Satinder P. Singh. "Intrinsically motivated reinforcement learning." Advances in neural information processing systems. 2005.

Botvinick, Matthew, and Marc Toussaint. "Planning as inference." Trends in cognitive sciences 16.10 (2012): 485-488.

Vezhnevets, Alexander, et al. "Strategic attentive writer for learning macro-actions." Advances in neural information processing systems. 2016.

Brochu, Eric, Vlad M. Cora, and Nando De Freitas. "A tutorial on Bayesian optimization of expensive cost functions, with application to active user modeling and hierarchical reinforcement learning." arXiv preprint arXiv:1012.2599 (2010).

Graves, Alex. "Practical variational inference for neural networks." Advances in neural information processing systems. 2011.

Gal, Yarin, and Zoubin Ghahramani. "Dropout as a bayesian approximation: Representing model uncertainty in deep learning." international conference on machine learning. 2016.

Rasmussen, Carl Edward. "Gaussian processes in machine learning." Summer School on Machine Learning. Springer, Berlin, Heidelberg, 2003.

Orbanz, Peter. "Lecture notes on bayesian nonparametrics." Journal of Mathematical Psychology 56 (2012): 1-12.

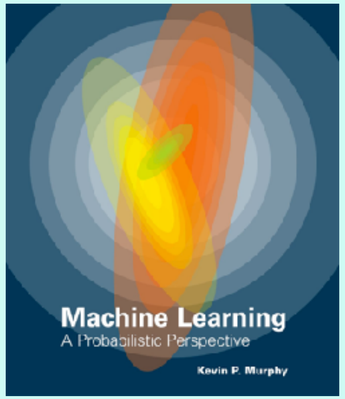
Frigyik et al. (2010), Introduction to the Dirichlet Distribution and Related Processes

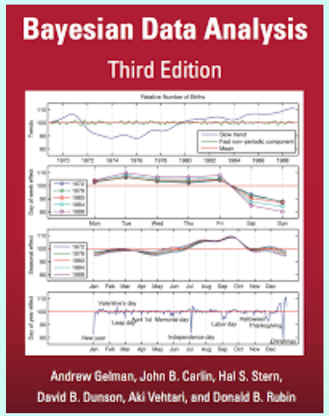
Hennig, Philipp, Michael A. Osborne, and Mark Girolami. "Probabilistic numerics and uncertainty in computations." Proceedings of the Royal Society A: Mathematical, Physical and Engineering Sciences 471.2179 (2015): 20150142.

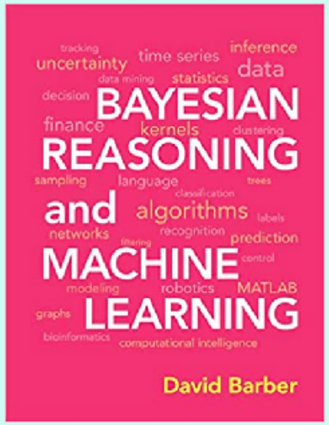
Brandt, Allan M. "Racism and research: the case of the Tuskegee Syphilis Study." Hastings center report (1978): 21-29.

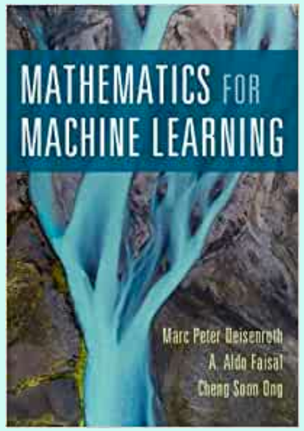
Black, Edwin. IBM and the Holocaust: The strategic alliance between Nazi Germany and America's most powerful corporation. Random House Inc., 2001.

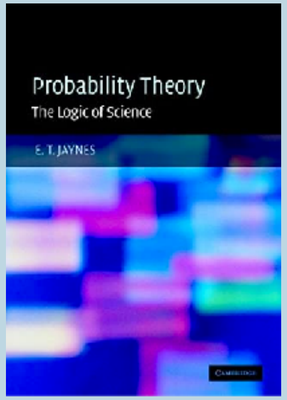
Book Recommendations

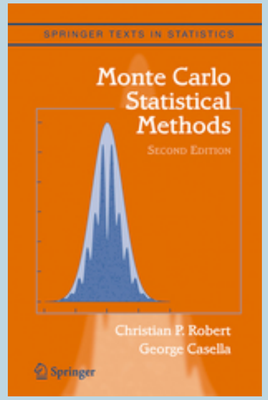


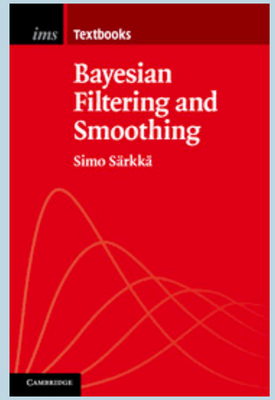


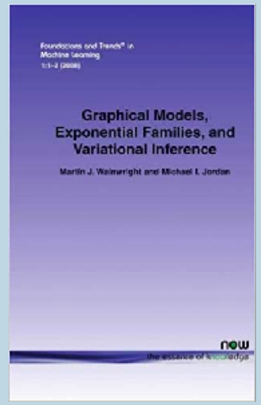


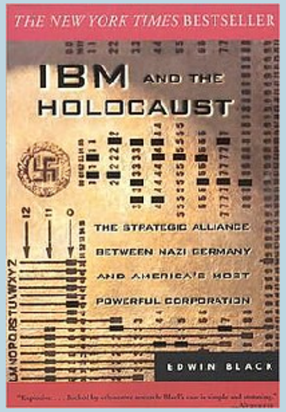


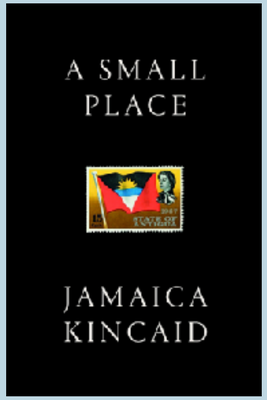












Machines

Dearest, note how these two are alike:

This harpsichord pavane by Purcell
And the racer's twelve-speed bike.
The machinery of grace is always simple.

This chrome trapezoid, one wheel connected
To another of concentric gears,
Which Ptolemy dreamt of and Schwinn perfected,
Is gone. The cyclist, not the cycle, steers.

And in the playing, Purcell's chords are played away.
So this talk, or touch if I were there,
Should work its effortless gadgetry of love,
Like Dante's heaven, and melt into the air.
If it doesn't, of course, I've fallen. So much is chance,
So much agility, desire, and feverish care,
As bicyclists and harpsichordists prove
Who only by moving can balance,
Only by balancing move.

Bayesian Learning

Basics | Computation | Approximation | Futures

Shakir Mohamed

@shakir_za

DeepMind