**1. Intro to Bayesian Inference**

1.1 Comp Neuroscience

* Cox’s axioms –> “degrees of belief must satisfy rules of probability”. Sum & product rules. (1).
* Example: Likelihood density function non-zero in certain range with parameter λ. Introduce Bayes rule by getting form of posterior for this likelihood (1). Plots of density function and likelihood to illustrate difference (i.e. the independent variable is x for prob density and λ for likelihood) (1). Plots showing how posterior changes as more data are added. MAP estimate and error bar also indicated. Posterior is then approximated and samples are taken + displayed.

1.2 David Mackay: An Introduction To Bayesian Inference (II): Inference Of Parameters And Models

* Example: Exponential decay example. Find parameter λ that describes exponential that fits observed data. (25 mins):
  + Students asked for different ways (pre-Bayesian inference) to find λ. Different ways of finding λ were considered e.g. binning points and taking logs, considering cumulative distribution, using fact that mean of exponential distribution is λ. All used as preamble to better Bayesian inference that is a much more general method.
* Bayesian inference examples (using gnuplot code) shown similar to those of Comp Neuroscience above. (17 mins).
* Introduces model comparison by considering mixture of exponentials and performing inference (12 mins)
* Likelihood and posterior plots for mixture of exponentials and considers model comparison (15 mins).

**2. Regression**

*2.1 3F3 notes*

* Goal of regression (1)
* Presentation of 2-D linear regression problem (1)
* Considering problem probabilistically (1)
* Finding point estimate of parameters through ML (3) (One slide on derivative of vectors).
* Show how the simple linear case with 2 variables is easily extended to D-dimensional, polynomial and non-linear regression problems (3)
* Consider other noise models, specifically non-Gaussian. **Example**: Laplacian noise (models outliers). (1)
* Bayes’ rule for model parameters and models (1).
* Implicit comparison of ML & MAP and relating the latter to regularization (was set as task in lecture). Bayes’ rule shows relation between these estimators. (1)

**3. Classification**

*3.1 3F3 Notes*

* Presentation of problem to solve: labelled data where the label is a categorical variable. (1)
* Example: Fisher Iris Dataset (1)
* Linear classification: deterministic (Heaviside function) (1), probabilistic (sigmoid) (1). Comparison to linear regression.
* Logistic classification:
  + Form likelihood equation for labelled dataset(1).
  + Derives ML estimate by differentiating likelihood w.r.t. parameters.(1)
  + Learning rules: batch/online (1)
  + Matlab figure showing classification contour plot for linear features (1)
* MAP: Shows equivalence of maximising MAP to maximising likelihood and minimising L2 norm of parameters. Contrasted with Bayesian learning of parameters.(1)
* Nonlinear + multinomial classification (2).
* Exercises (1).

*3.2 4F13*

**4. Clustering/EM**

*4.1 Mackay*

* Starts with example on binary information channel with Gaussian noise and task is to find the posterior for one of the values given the observed output. Introduces mixture of Gaussians for output can be represented as a mixture of two Gaussians with means determined by the binary input values. (9 mins).
* Real world example of speech spectrograms to show the different clusters for different words.(1 min)
* Introduces k-means algorithm. (33 min):
  + Illustrates algorithm by going through the steps on a normal dataset as an example. (15 mins)
  + Shows abnormal example of two clusters separated horizontally where the greatest variance is along the y-axis. Illustrates failure case of k-means.(2 mins)
  + Explains failure case through assumptions made by k-means algorithm. Namely, assuming equal priors and spherical Gaussians of same variance. Highlights k-means as roughly a MAP algorithm. (Intuition from earlier information channel example). (2 mins)
  + Moves on to the first version of soft k-means algorithm where all the clusters share different responsibilities for a data point. Shows examples of how it works well even if more clusters than necessary are assumed. Still fails on ‘lozenges’ type of dataset (10 mins).
* Soft k-means version 2. (8 mins).
  + Introduce k-means version where the variances are updated at each step.
  + Examples with the ‘lozenges’ dataset and other datasets of varying sizes.

*4.2 4F13 – Rasmussen*

* Gives notation (1).
* Reformulation of Bayes rule for latent variables so as to get the log-likelihood of the observations given the parameters which is expressed as the sum of the parameters and a lower bound functional of an arbitrary distribution of the latent variables (1).
* E and M steps. Short proof that log-likelihood always decreases.(1)
* Mixture of Gaussians (2) and Factor analysis (1) examples.
* Appendix on properties of KL-divergence i.e. the condition for the minimum is derived and it is shown that KL is always non-negative (1).

*4.3 4F13 – 07/08*

Similar to Rasmussen’s notes above but more detailed.

* Derives lower bound functional from Jensen’s inequality (1).
* EM algorithm (1):
  + E step – maximise lower bound (LB) functional w.r.t. distr over hidden variables (q(x)). Easier to understand than Rasmussen which says that minimising KL maximises LB functional as likelihood is independent of likelihood is zero (i.e. derivative w.r.t. q(x) is zero).
  + M step – maximise LB functional w.r.t. parameters. Same as Rasmussen.
* Example: Contour plot of parameters vs. q(x) showing trace of gradient ascent.(1)
* Derive KL from diff of likelihood and lower bound. Show monotonic increase of likelihood like Rasmussen above. (1)
* Example – Factor Analysis: FA introduced earlier in lecture. Derivation of E and M steps. (3)
* Limitations of MOG when used to approximate arbitrary distributions.
* E and M steps for MOG and some issues (local maxima).

**5. Sequence Modelling**

*5.1 Rich Turner Notes*

* Pictorial examples of sequence data and goals of sequence modelling (2).
* Markov models:
  + Markov N-gram models and how some properties may be calculated. Questions that are presumably answered by audience (3).
  + Example: Dasher Project (1)
  + Markov models for continuous data: AR models. Examples of waveforms generated by model and some properties e.g. stationary distribution (3).
  + Example: pendulum swing up control (1).
* HMMs:
  + Example uses (1)
  + Discrete hidden states + some questions (1)
  + Continuous hidden states with example calculations + plots. (1)
* Varieties of inference: Distributional and point estimates.(1)
* Kalman filter – derivation of prediction equation with graphs for illustration. (1)
* Forward algorithm – derivation + graphs for illustration. (1)
* Computation of likelihood used in the Kalman filter and Forward algorithms efficiently. (1)
* Dynamic programming and Forward algorithm - the algorithm runs in linear time as opposed to the exponential time needed to calculate the likelihood (1).
* ML learning of HMMs. Log-likelihood gradient depends on simple posterior moments (1).

*5.2 Sam Roweis Notes*

* Generative models (1).
* Markov Models intro (1).
* Learning Markov models (1).
* HMMs , links to Markov chains and mixture models, using a trick to calculate probability of observed sequence. (4)
* Example: Bugs on Grid (transition between discrete states over time). Compares a naïve approach whereby the large sum of the likelihood has to be calculated and a recursion technique utilising dynamic programming.(1)
* Forward-Backward algorithm to infer hidden states. Refers to bugs example for illustration (4)
* Viterbi decoding. (1)
* EM algorithm. Special case called Baum-Welch considered.(1)
* Parameter estimation – one and two frame estimates. (2)
* Applications: Recognition (1), Character sequences (1), Geyser data (1).
* HMM regularization (2).
* List of advanced topics (1).
* Computational concerns (3).
* List of more applications (1).
* Introduces Linear Dynamical Systems and relates them to HMMs. (1)
* Application: Computational Biology (5).

**6. MCMC**

*6.1 Iain Murray NIPS*

* Review of linear regression. Focus on looking at samples from large range that have high likelihoods of data. (10 mins)
* Monte Carlo computations:
  + Monte Carlo prediction and integration + extension to the assumed linear model e.g. noise, placing priors on parameters and noise, modelling outliers (10 mins).
  + Goal of MC Inference (5 mins).
* Generating samples (standard generators):
  + Goal distribution: un-normalized posterior
  + Sampling from discrete and continuous distributions (5 mins).
  + Rejection sampling (3 mins).
  + Importance sampling + linear regression and high dimensional parameters examples (7 mins).
* Getting samples (Markov chains):
  + Markov chains and some properties in context of random sampling (4 mins).
  + Metropolis-Hastings with example calculation (with audience question to find why an example is wrong) and code sample with example plots (17 mins).
  + Intro to MCMC – 2-D and higher dimensions, creating an MCMC scheme. (10 mins)
  + Gibbs sampling (8 mins).
  + Auxiliary variable methods: slice sampling (7 mins).
* Practical issues: diagnostics, testing code for consistency, which method to use for different problems (7 mins).
* Reading recommendations.
* Kaggle MCMC example (6 mins).
* Audience questions (8 mins).

*6.2 Mackay*

Lecture 1:

* Introduces MC methods by considering summations and integrals that have to be calculated for inference (10 mins).
* MC methods – Goals introduced as seeking to estimate expectations and drawing samples from posterior. Explained in context of clustering and nanotechnology example (17 mins).
* Importance sampling – Introduced by considering a sampler density to weight the samples. Shows problems of choosing wrong sample distribution (13 mins).
* Rejection sampling with live code examples (7 mins)
* MCMC – Metropolis algorithm with live code examples (12 mins).
* Gibbs sampling – applied to Gaussian inference and clustering with live code examples (10 mins)

Lecture 2:

* Review of previous lecture and introduction of toy problem to illustrate problems with Metropolis algorithm. Idea is that it takes a long time to get a good sample from the target distribution. Code examples of location of sample with number of iterations and plots of how estimated distribution changes with time. (15 mins)
* Illustration of random walk problems in high dimensions: step size must be smaller than smallest length scale of distribution. (8 mins)
* Efficient MC methods:
  + Hamiltonian Monte Carlo with live code examples (16 mins)
  + Adler’s Overrelaxation – compared to Gibbs sampling to show its advantages over the latter. (6 mins).
  + Ordered Overrelaxation – short intro (3 mins)
* Robust MC methods:
  + Slice sampling (11 mins).
* Self-terminating MC - Exact sampling (20 mins):
  + Uses example of random walk to illustrate procedure.
  + Example: adding and removing tiles of 3 colours from a hexagon (can also be thought of as adding/removing bricks to a brickyard if viewed in 3D) to find out when they coalesce ().
  + Example: Ising model – array of spins that are coupled to neighbour. Use exact sampling to find perfect samples at critical temperatures.

*6.3 4F13*

* Introduces objective as seeking to approximate expectations of a certain function (1).
* Motivations: predictions and approximating marginal likelihoods.(1)
* Example: numerical integration (1).
* MC introduced by considering estimation of expectation as the average of a number of samples. (1)
* Introduces random number generation: transformation methods or discretisation of desired distribution to sample from. (1)
* Rejection sampling – simply states algorithm (graph for illustration) (1)
* Importance sampling – example of limits on choice of sampling density. (1)
* Caution that is taken when selecting sampling density.(1)
* Transition from Independent sampling (Rejection & Importance) to Markov Chains (1).
* MCMC – Intro + relevant properties of Markov chains (detailed balance and requirement of ergodicity for stationary distributions) (3).
* Metropolis–Hastings:
  + Algorithm description + properties and constraints on choice of proposal distribution (3).
  + M-H example: A plot of the samples and the traces they formed for a sharp Gaussian. (1)
  + Things to consider when choosing proposal distribution (1).
  + Variants of M-H: considering each dimension separately (1).
* Gibbs sampling:
  + Introduced as special case of M-H (1)
  + Example: same type as M-H above. (1)
* Hybrid MC (Hamiltonian MC) + example (3)
* Summary + notes on MC in practice (2).

**7. Model Comparison**

*7.1 Mackay*

* Introduces Occam’s razor by asking question: ‘how many boxes are behind a tree when we only have an orthographic projection of the scene?’ and explains some motivations e.g. aesthetic + empirical successes (1 page).
* Shows Bayesian inference embodies Occam’s razor by reasoning that the posterior probability for a simpler model is greater than that of a complex one (under some assumptions). Gives an example to illustrate this whereby an integer sequence can be generated by a simple arithmetic progression or by using a complex polynomial (2 pages).
* Example: Roman Inquisition epicyclical model (more parameters) vs. Copernican heliocentric model of solar system.
* Explains the two levels of inference and mentions why Occam’s razor may be needed when inferring models (complex models can always fit data better) (2 pages).
* Talks about evaluating the evidence using an Occam factor that captures the complexity of a model. (2 pages)
* Relates initial example of boxes behind a tree to Occam factor. (1 page)
* MDL (Minimum Description Length) – Introduction and relation to Bayesian model comparison + other topics (on-line learning + cross-validation, ‘bits-back’ encoding method).

*7.2 4F13*

* Introduces model complexity through fitting polynomials to data and then asks questions that are considered in model comparison (2).
* Occam’s razor (1)
* Terminology (1)
* Occam’s razor example. (code demo) (1)
* Practical Bayesian approaches:
  + Laplace approximation (1)
  + Bayesian Information Criterion (1)
  + Sampling Approximations: Example: Importance Sampling (1).
  + Variational Bayesian Learning (2).
* Example: Mixture of Factor Analysers (code demo) (1)
* Example: HMM (code demo) (1)
* Linear Dynamical Systems (1).