10-708 – Probabilistic Graphical Models.

A pedagogical tool to track the key aspects of the lecture that the instructor emphasises, and as a checklist of things you have judged are important.

Some content is given little weight in the lectures for time-keeping purposes. They are listed in italics, and you must make a judgement on whether it is necessary to allocate time to understanding them.

Key areas to understand.

Week 1.

Lecture 1 – Introduction.

- Understand that probabilistic graphical models as a language for representing complex probability distributions, reasoning under uncertainty and noise.
- Understand the intractability of working with full joint probability tables.
- Understand the benefits of probabilistic graphical models in this context.
- Understand the formal definition of a probabilistic graphical model.
- Understand that graph structure, topology, and conditional independence can be used to encode domain knowledge and yield representational cost-savings.

Lecture 2 – Directed Graphical Models.

- Understand that Bayesian Networks/Directed Graphical Models have directed edges - causality relationships.
- Understand that Markov Random Fields/Undirected Graphical Models have undirected edges- correlation relationships.
- Understand the notational conventions of the course.
- Understand the HMM dishonest casino model in context of the knowledge engineering process.
- Understand the definition and properties of a Bayesian Network, the factorisation theorem, and the role of conditional independence.
- Understand the local structures and independencies.
- Understand the I-maps.
- Understand "explaining away"
- Understand the "d-separation criterion" on moralised ancestral graphs.
- Understand equivalence theorem.
- Understand "soundness" and "completeness" of d-separation with respect to a Bayesian Network factorisation law.
- Understanding of the concepts mathematically, but also be able to rehearse the semantics.

- *Understand the identifiability of local and global independence of Bayesian Networks via d-separation and the Bayes ball.*

Outstanding:

- *Jordan (2003): Understanding of Bayes ball algorithm.*
- *Jordan (2003): Understanding of explaining away.*
- *Complete this.*

Lecture 3 – Undirected Graphical Models.

- Understand a motivation for undirected graphical models through P-maps.
- Understand that undirected graphical models have undirected edges – pairwise, non-causal, relationships.
- Understand the formal definition and properties of a Markov Random Field/Markov Network/UGM, in terms of the potential and partition function; and the Gibbs distribution.
- Understand the properties and interpretations of the clique potential and partition functions.
- Understand the formal definition of cliques, max-cliques, and sub-cliques and their relation to potential functions.
- Understand that the choice of clique size can yield representations of varying granularity in terms of the Gibbs distribution.
- Understand global Markov independencies – separation.
- Understand local Markov independencies – Markov blankets.
- Understand soundness, completeness of global Markov properties.
- Understand the theorems on the relation between local and global Markov properties.
- Understand the Hammersley-Clifford theorem.
- Understand the formal theorems on P-maps for UGMs.
- Be fluent in the semantics of the formalism to express relationships between distributions and graphs.
- Understand the unconstrained for of clique potential using exponentiated energy functions.
- Be familiar with UGM representations of Boltzmann Machines, Ising Models, Restricted Boltzmann Machines, and Conditional Random Fields.

Outstanding:

- *Minimal I-maps and P-maps*

Week 3.

Lecture 4 – Exact inference and variable elimination.

- Understand that queries on the graphical model can be viewed as statistical inference and learning/estimation problems.
- Understand evidence as an assignment of values to a set of random variables.
- Understand that we can query the likelihood, posterior, and maximum a posterior (MAP) or most probable assignment.
- Understand the associated examples, and how PGMs illustrate the explicit role of the context of an MAP query.
- Understand that computational complexity results of computing posteriors on arbitrary PGMs as NP-hard.
- Understand that inference approaches can be viewed as exact, or approximate.
- Understand the elimination algorithms for PGMs with a chain structure.
- Understand its application to examples such as the forward-backward algorithm for HMMs, undirected chains, and CRFs.
- Understand the role of the sum-product operation as providing a means of quantifying the complexity of the inference precisely.
- Understand the variable elimination algorithm for general PGMs at a high-level.
- Understand the specifics of variable elimination, such as the introduction of evidence, outcome of elimination, sum-product variable elimination.
- Understand how to evaluate the computational complexity of variable elimination and its notable determinants.

- Understand the variable elimination algorithm for more complex non-chain structured PGMs.
- Understand the graph-theoretic formulation of variable elimination.
- Understand that elimination can be viewed as message passing over an ordering over clique trees.
- Understand the nuances regarding complexity.

Outstanding:

Lecture slides from 10-708 Spring 2020 go into more comprehensive detail concerning the steps to convert DGMs and UGMs into junction trees in the supplementary slides, which is not present in these Spring 2019 slides.

Lecture 5 – Parameter estimation in fully observed Bayesian Networks.

This was skipped in the Spring 2019 series; but is covered in Spring 2020 in the first half of Lecture 5. Only a proportion of the material in the slides is covered as the Spring 2020 concatenates the lecture content in Lecture 5 and 6 in Spring 2019 into Spring 2020 Lecture 5.

- Have knowledge of the main scenarios of learning/parameter estimation of PGMs, i.e. completely observed/partially observed/directed/undirected and estimation principles.
- Understand the distinction between PGM structure learning and parameter estimation.
- Understand the role of the generalised exponential family of distributions for parameter estimation in the context outlined.
- Understand that single and two-node graphical models form generalised building blocks for more complex graphical models.
- Understand the terms comprising the functional form of the exponential family distribution, such as the canonical parameter, sufficient statistic, and log normaliser.
- Understand the exponential family representation of the multivariate Gaussian, Multinomial.
- Understand the benefits of the exponential family representation.
- Understand the relation between the exponential family distributions and moments.
- Understand the relation between canonical and moment parameters, and the process of moment matching.
- Understand the definition and role of a sufficient statistic.
- Understand the relation between data, sufficient statistics and parameters under the frequentist and Bayesian paradigms; and the Neyman factorisation theorem.
- Understand how these various methods can give a unified approach to density estimation for single random variables/single node graphical models.
- Understand the definition of a Generalised Linear Model (GLM), and the terms comprising it, such as the response function, inverse transform function, and the role of the exponential family distribution.
- Understand the relation between maximum likelihood estimation (MLE) for canonical GLMs, online learning, batch learning.
- Understand the trade-offs between stochastic gradient ascent (SGA) and iteratively reweighted least squares (IRLS).
- Understand that MLE for general BNs, under certain conditions, can be decomposed analytically.
- Appreciate that the exponential family and GLMs, as one and two-node graphical models, can be used as generalised building blocks for more complex graphical models.

Outstanding:

Week 4.

Lecture 6 – Parameter estimation in partially observed GMs (POGMs).

- Understand how parameter estimation is conducted for fully observed DGMs/BNs.
- Understand that in this setting, the log-likelihood function decouples into a set of local terms.
- Understand plate notation.
- Understand the process of ML parameter estimation for BNs with tabular CPDs.
- Understand ML estimation from HMMs in a fully observed, supervised learning setting.
- Understand ML in this setting estimates empirical probabilities/relative frequencies.
- Understand ML overfitting may be addressed via the inclusion of pseudo-counts; and amounts to the inclusion of a prior from a Bayesian perspective.
- Understand that parameter estimation in partially observed GMs involves iterating between inference tasks.
- Understand that inference can be viewed as a subroutine for parameter estimation.
- Understand the various senses in which variables can be unobserved, or latent.
- Understand the role latent variables play in mixture models, and Gaussian mixture models (GMMs).
- Understand that parameter estimation in POGMs is due to a coupling of likelihoods with latent variables.
- Understand the usefulness of the EM algorithm for ML estimation in the presence of unobserved latent variables.
- Understand a high-level description of the EM algorithm (inference of unobserved variables, given observed data and current parameter estimates, followed by updating parameter estimates).
- Understand the relation between EM for GMMs and soft-clustering K-means.
- Understand the isomorphy of MLE and EM via sufficient statistics.
- Understand the mathematical details of the components of the EM algorithm, such as the complete log-likelihood, incomplete log-likelihood, expected complete log-likelihood, and the posterior distribution over latent variables.
- Understand that the EM algorithm maximises the expected complete log-likelihood, a lower bound on the incomplete log-likelihood, via Jensen's inequality.
- Understand the EM algorithm in terms of free energy, entropy, and co-ordinate ascent.
- Understand applications of the EM algorithm for HMMs (Baum-Welch), BNs, and conditional mixture models (CMMs).
- Be aware of EM variants, and of its advantages and disadvantages.

Lecture 7 – Parameter estimation for UGMs.

- Understand that in general, log-likelihoods for UGMs do not easily decompose into local terms, due to the presence of a normalisation constant over potential functions.
- Understand that inference/marginalisation will be required for parameter estimation of UGMs, even when all variables are fully observed.
- Reinforce understanding on the role of counts, empirical probabilities in MLE of PGMs.
- Reinforce understand that the Hammersley-Clifford theorem allows the specification of an UGM in terms of a Gibbs distribution and partition function.
- Understand how key sufficient statistics are computed – total counts and clique counts.
- Understand the issues the normalisation constant poses for parameter estimation.

- Understand the process of MLE for UGMs only produces a condition on clique marginals, and no means of attaining the ML parameters.
- Understand how the above yields a motivation for two workhorse algorithms: iterative proportional fitting (IPF); and generalised iterative scaling (GIS).
- Understand that IPF can be used for tabular potentials.
- Understand the derivation and properties of IPF as a fixed-point equation.
- Understand how IPF can be viewed as a co-ordinate ascent algorithm, and from the perspective of information theory as minimising Kullback-Leibler (KL) divergence.

- Appreciate that the spirit of the course is not only to be theoreticians, but to ultimately be concerned with the "messiness" of practical implementation, and its requisite considerations.
- Understand the infeasibility of using tabular potential functions in practice with a heuristic example – exponential representational cost.
- Understand that the granularities of tabular potential functions can be compressed using features encoding domain knowledge.
- Understand the relation between features, weights, micro-potentials and clique potentials, in context of the exponential family distribution and GLMs.
- Understand how linearisation and lower-bounds are invoked mathematically to set up a scaled-log-likelihood optimisation problem.
- Understand how GIS iteratively improves the lower-bound on scaled-log-likelihood, and its similarities with the IPF algorithm.

- Reinforce understanding of the functional form of the exponential family distribution.
- Understand the relation between the exponential family, maximum likelihood, and expectations of sufficient statistics.
- Understand that the exponential family distribution is the solution to a constrained, variational optimisation problem over either an entropy or KL-divergence objective – maximum entropy methods.
- Understand how the maximum-entropy principle (MEP) to parametrisation offers a dual perspective to MLE.
- Understand the label bias problem in HMMs, and the motivation for conditional random fields (CRFs).

- *Review CRF tutorial papers, inference and learning in CRFs*

Week 5.

Lecture 8 – Causal discovery and inference. Guest lecture by Kun Zhang.

- Understand a distinction between causality and dependence through conditional probability.
- Understand the notation of conditioning on an intervention – Judea Pearl's do-calculus.
- Understand the distinction between causal connections and associational information.
- Understand how causal thinking can shed light on examples such as Simpson's paradox, the Monty Hall problem, and other strange dependence patterns (V-structures).
- Understand the role of intervention in causal models.
- Understand how interventions and their effects may be represented in DAGs.
- Understand three archetypal questions in computer science/AI, and how they are formulated:

    - Prediction
    - Intervention
    - Counterfactual thinking
- Understand the distinction between DAGs, and causal DAGs, i.e. DAGs endowed with a causal interpretation.
- Understand the formal conditions that must be satisfied in order for DAGS to admit a causal interpretation.
- Understand the "gold-standard" epistemological status of RCTs within causal inference.
- Understand how manipulating and conditioning are formalised using marginal and conditional probabilities.
- Understand how causal information and observational data may be combined in situations where RCTs cannot be conducted.
- Understand the formal definitions of a causal effect and causal effect identifiability through do-calculus.
- Understand the effect of confounders on identifiability.
- Understand graphical criteria that can establish whether causal effects are identifiable e.g. front/back-door criterion.
- Understand the principles of constraint-based causal discovery.
- Understand the role and interplay of conditional independence constraints, causal Markov conditions, and "faithfulness" assumptions in the above.
- Appreciate that appeals to graphical criteria on causal DAG structure alone can under certain conditions yield powerful insights on the presence of confounders, or whether one variable directly causes another.
- Be aware that this forms the basis for PC ad FCI algorithms (insert authors?)

<u>Lecture 9 – Modelling networks: Gaussian graphical models and Ising models.</u>

- Be sensitive to the ontological status of a (relational) network.
- Be aware that graphs, networks, relational networks etc, as objects of study, are synthetic artefacts, in the sense of lacking physicality.
- Understand that the lecture is concerned with algorithms for inferring the *structure*, or *topology* of a completely observed graphical model.
- Understand that the focus of the lecture is on two classes of algorithms that enjoy certain optimality guarantees:
    - Chow-Liu algorithm for trees – cover in own time.
    - Pairwise MRFs: covariance selection and neighbourhood selection with the graphical LASSO.
- Understand the key idea of the lecture – inference of network structure may be recast as a parameter estimation problem!
- Understand the mathematical formulation of the pairwise MRF.
- Understand that discrete and continuous nodal states yield the Ising/Potts model and Gaussian graphical models (GGMs) respectively.
- Understand how GGMs arise from the pairwise MRF formulation of MVG distribution.
- Understand that the parameter matrix encodes graph structure.
- Understand how parameter estimation of weights over potential functions can be viewed as the estimation of a precision matrix, in this context.
- Understand the distinction between a correlation network and a Markov network.
- Appreciate why and how the latter may employ a more meaningful dependence measure.
- Appreciate the role of sparsity assumption.

- Understand issues such as computational tractability with covariance selection motivate neighbourhood selection with LASSO/L1 regularisation for inferring the structure of a sparse GGM.
- Understand the significance of results on theoretical guarantees and consistent structure recovery, in the recommended readings.
- Reinforce knowledge of mathematical results concerning marginals, and conditionals of Gaussian distributions; and the matrix inversion lemma.
- Understand these results, covered in permit parameter estimation of GGMs via single-node conditional distributions, conditional auto-regressions, and conditional independence relations.
- Be aware of the history of academic trends in the development of GGMs (covariance selection vs L1-regularisation based methods)
- Understand how the above methods can be adapted for the Ising/Potts model by employing L1-regularised logistic regressions.
- Understand recent results adapt these principles to non-IID settings, and time-varying evolving networks e.g. kernel weighted L1-regularised logistic regression (KELLER) and temporally smoothed L1-regularised logistic regression (TESLA).
- Appreciate a particular way of conducting research in ML, as the systematic employment of "tricks", which are tractable, explicit, and yield proof opportunities.

- *This particular lecture is an excellent instructor-motivated presentation that covers the results of involved readings in an accessible manner.*
- *Formal results on the graphical LASSO in the main part of the lecture are stated in Meinshausen and Buhlmann (2006). Covariance selection is covered in Dempster (1972). Other methods related to sparse inverse covariance estimation with the graphical LASSO are contained in Friedman, Hastie, Tibshirani (2007). Results on estimating time-varying network structure e.g. KELLER, TESLA are given in Kolar, Le Song, Ahmed, Xing (2009).*

Week 6.

Lecture 10 – Sequential models.

- Understand the graphical model representation of factor analysis (FA) models, and that they consist of both continuous latent states and observations, modelled by MVGs.
- Understand that FAs form the building blocks for state-space models (SSMs)/linear dynamical systems (LDSs), in an analogous manner to MMs as building blocks of HMMs.
- Appreciate that the art of modelling can be viewed as invoking mathematics to express aspects of a phenomenon of interest in a language that is systematic, rather than seeing the mathematics as the phenomenon of primary interest.
- Reinforce knowledge of the properties of MVGs (marginal, conditional, joint), and the matrix inversion lemma.
- Know useful matrix algebra properties e.g. trace properties, derivatives, determinants etc.
- Understand the generative specification of the components of FA.
- Understand how MVG properties may be invoked to perform inference for FA – by specifying FA joint distributions, and then deriving posterior distributions of the latent variable.
- Understand how the matrix inversion lemma can be used together with the viewpoint of FA as a constrained Gaussian model to yield computational efficiencies.
- Understand that EM is used for ML parameter estimation of FA models, due to nonlinear coupling of parameters in log-likelihood.

- Reinforce understanding of EM, as applied to FA models (incomplete, complete, expected complete log-likelihood etc.).
- Understand how posterior density distribution derived above can assist in the computation of sufficient statistics in the E-step.
- Understand how trace and determinant derivative are used in the M-step.
- Understand the degeneracy of the FA model and the implications for identifiability.

- Understand the mathematical specification of continuous state SSMs/LDSs.
- Understand two distinct inference problems – filtering and smoothing.
- Understand that the solution to the filtering problem is given by the Kalman filter, for exact online inference/sequential Bayesian inference
- Understand that the solution to the smoothing problem is given by the Rauch-Tung-Striebel smoother, for exact offline inference in an LDS.
- Understand that the Kalman filter and Rauch-Tung-Striebel smoother are Gaussian analogs of the forward algorithm and the forward-backward (alpha-gamma) algorithm in HMMs, respectively.
- Understand the mathematical derivation of the Kalman filtering equations.

<u>Lecture 11 – Approximate inference: Mean Field (MF) and loopy Belief Propagation (BP) approximations.</u>

In the S.2019 videos, this lecture continues derivation of the Kalman filtering equations from Lecture 10.

For the rest of the lecture, content in the 2nd half of Lecture 11 slides is covered, that is, L11b – Approximate Inference and Topic Models: Mean Field and Loopy Belief Prop.

- Have a high-level understanding of generic topic models.
- Understand that topic models are presented in these lectures to expose a concrete need for approximate inference methods.
- Appreciate and understand the various design choices involved in modelling tasks.
- Understand document embeddings, and the various tasks that topic models could be useful for.
- Understand that the bag-of-words representation ignores word-ordering.
- *More formally, this is equivalent to an exchangeability assumption (de Finetti).*
- Appreciate the trade-offs involved in selecting representations of data.
- Understand how topics and documents can be viewed as points on word and topic simplexes.
- Understand the conceptual similarities and differences between Latent Semantic Indexing (LSI), and topic models.
- Understand how the latter proposes probabilistic inference techniques as a means of bypassing computationally expensive matrix inversions involved in the former.
- Understand the main components and generative process for a generic topic model.
- Understand the implications on efficient inference of electing to use conjugate priors in topic models, such as latent Dirichlet allocation (LDA).
- Understand how this and topic correlation motivate variants such as Correlated Topic Models (CTMs), and Logistic Normal Admixture Models (LonTAM).
- Appreciate the motivation behind modelling choices in the instructor's paper on LonTAMs.
- Have a visual understanding of how the components of topic models relate to the more abstract plate notation.

- Understand concretely (i.e. in terms of mathematics) why inference and learning are intractable.
- Understand the principles underlying variational inference as the reformulation of an inference problem to an optimisation problem.
- Understand how variational inference methods involve iterative maximisation of a variational lower bound, alternating parameter updates similar to co-ordinate ascent.
- Understand an application of the mean-field (MF) approximation in topic models – as a method for approximating the true posterior using a family of distributions (mean-field family) involving a product of marginal factors.
- Understand how the MF approximation yields a co-ordinate ascent algorithm for LDA.

LDA – latent Dirichlet allocation (Blei, Ng, Jordan 2003)
CTMs – correlated topic models (Blei and Lafferty 2007)
LoNTAM – logistic normal admixture models (Ahmed and Xing 2006)

- Be aware of the various inference problems/probabilistic queries that can be posed with a graphical model.
- Reinforce knowledge of methods covered in exact inference on graphical models .e.g brute force, variable elimination, and message passing/belief propagation.
- Reinforce understanding of the relationship between variable elimination and message passing on a clique tree.
- Reinforce understanding of computational complexity of variable elimination in terms of elimination cliques, tree-width, and elimination orderings.
- Reinforce understanding of the sum-product message passing algorithm on generic tree-structured GMs and the generic message passing equation on trees.
- Understand that sum-product message passing on trees is convergent to a unique fixed point after a fixed number of iterations.
- Understand the sum-product message passing applied to HMMs specifically.
- Understand the result concerning the correctness of belief propagation on tree-structured GMs – all marginal inference queries can be obtained.
- Understand the properties of local and global consistency.
- Understand that these properties are equivalent on junction trees; but not in general.
- Understand that approximate inference may provide an appropriate alternative in cases where converting a generic graphical model into a junction tree is not computationally feasible (large tree-width).
- Understand that belief propagation on loopy graphs, i.e. loopy belief propagation (BP) currently offers *no guarantees* on convergence.
- Understand loopy BP empirically results in either:
    - Convergence – results in a good approximation assessed by comparison with brute-force marginalisation of variables on synthetic datasets.
    - Oscillation.
- Understand that in the absence of convergence guarantees, and hence intractable marginal posteriors, an approximate solution to inference can be sought.
- Reinforce understanding of the problem setup whereby an approximation of an intractable probability distribution of interest is sought.
- Reinforce understanding of appropriate formalisms of the above, through the KL-divergence, lower bound on log-likelihood, entropy, and significantly, the (Gibbs) free energy.

- Understand how a joint distribution on a tree-structured GM can be specified in terms of singleton node and pairwise node probabilities, and a node degree term.
- Understand how this can be used to derive free energy and entropy terms for tree-structured GMs.
- Understand that the Bethe free energy can be viewed as an approximation or surrogate for [INSERT WHEN CLARIFIED].
- Understand that the singleton and pairwise marginals of interest can be obtained as the solution to a constrained minimisation problem - minimising the Bethe free energy subject to normalisation and local consistency constraints.
- Understand that with some re-parametrisation, belief propagation equations on loopy graphs are obtained.
- Appreciate this insight historically – ad-hoc, empirical investigation of loopy BP came to invite a theoretical contextualisation that related it to constrained optimisation using principles from statistical physics.

Week 7.

Lecture 12 – Theory of Variational Inference: Marginal Polytope, Inner and Outer Approximation.

The first half of this lecture covers content of slides titled "L11a – Variational Inference: Loopy Belief Propagation". The second half covers slides titled "L12 – Theory of Variational Inference: Marginal Polytope, Inner and Outer Approximation".

- Understand that the theory of variational inference provides perspective that relates loopy BP with MF approximation.
- Understand the principle of variational methods is to represent a quantity of interest as a solution to an optimisation problem; and then to approximate the former via solving a relaxed version of the latter.
- Understand that a motivation for variational methods is also computational – allowing for use of local, incremental, iterative update algorithms.
- Understand that exponential families and convex analysis are key components of formulating inference problems in UGMs in a variational form.
- Understand how to represent UGMS in an exponential family form.
- Understand the notion of effective canonical parameters and their relation to the log-partition function.
- Understand that certain computations on the exponential family components yield quantities of interest for inference.
- Understand that computing the expectation of the mean parameters (sufficient statistics) given canonical parameters yields the marginal of interest.
- Understand that computation of the normaliser yields the log-partition function.
- Understand that this discloses a connection between parameter estimation and inference.
- Understand the mathematical specification of the following:
    - Exponential families (mean parameters, canonical parameters, log-partition function).
    - Convexity, conjugate duality; Legendre transform/conjugate dual of the log-partition function, Shannon entropy.
    - Constraint set of all realisable mean parameters –convex polytope, half-plane representation of a convex polytope, marginal polytope.

    - Gradient mapping between mean and canonical parameters.
    - Relationship between conjugate dual of the log-partition and Shannon entropy.
- Understand the nuances of how the above elements relate to form a variational framework to understand both exact and approximate inference in graphical models.

- Understand the variational formulation of the log-partition function, with particular emphasis on the marginal polytope and negative entropy terms.
- Understand the computational challenges of characterising the marginal polytope and conjugate dual function.
- Understand how exact inference problems can be placed within the variational framework e.g. Bernoulli, Ising model, general exponential family.
- Understand that the MF approximation approximates the marginal polytope with a non-convex inner bound, but uses the exact form of entropy.
- Understand the notion of a tractable subgraph.
- Understand how tractable subgraphs can be used to restrict the set of canonical parameters, and hence construct the inner approximation to the marginal polytope.
- Understand how tractable subgraphs can be used compute the exact entropy.
- Understand the specification of the variational problem being solved by the MF approximation.
- Understand how the non-convexity of the inner approximation to the marginal polytope, and geometric considerations, can offer insight on the properties of the solution obtained via the MF approximation.

- Understand how the variational perspective can offer insight on exact inference on trees, and on approximate inference on loopy graphs, under the same framework.
- Understand sum-product message passing/belief propagation on trees from a variational perspective.
- Understand that the solution of the variational formulation of the log-partition function for tree-structured GMs yields the message passing updates.
- Understand that the Bethe approximation approximates the marginal polytope with a polyhedral outer bound, and a non-convex Bethe entropy approximation,
- Understand that these approximations specify the Bethe variational problem over pseudo-marginals, and the constrained solution of this via Lagrangians yields loopy BP.

<u>Lecture 13 – Approximate Inference: Monte Carlo and sequential Monte Carlo methods.</u>

The first half of this lecture covers content of the slides titled "L12 – Theory of Variational Inference: Marginal Polytope, Inner and Outer Approximations". The second half of this lecture covers slides titled "L13 - Approximate Inference: Monte Carlo and sequential Monte Carlo methods."

- Understand the distinction between exact/closed-form representations and sample-based representations of probability distributions.
- Understand the principles behind Monte Carlo methods – producing a stochastic representation of a probability distribution with the use of sampling.
- Understand that the asymptotic properties of these representations.
- Understand that inference queries (marginal posteriors, expectations) can be approximated on this basis.
- Understand the challenges of sampling based techniques.

- Understand ancestral sampling from a Bayesian network.
- Understand that a specific limitation of sampling and sampling-based techniques – rare events.
- Understand the context of the rejection sampling algorithm.
- Understand the mathematical procedure behind rejection sampling, with emphasis on a target distribution, proposal distribution, unnormalised distribution, and comparison function.
- Understand the limitations of rejection sampling, e.g. excessive rejection rate, sensitivity to proposal and comparison function, and pathological behaviour in high-dimensions.
- Understand the principles, motivation, and mathematical specification of adaptive rejection sampling.
- Understand the issues with using uniform sampling to evaluate expectations as a motivation for the importance sampling algorithm.
- Understand the mathematical procedure behind importance sampling.
- Understand that importance weights are constructed from likelihood ratios.
- Understand that normalised importance sampling can be used in situations where the target distribution can only be evaluated up to a normalisation constant.
- Understand the statistical properties of unnormalised and normalised importance sampling.
- Understand how rejection and importance sampling are normally used for sampling from univariate distributions.
- Understand the motivation and procedures involved in the weighted resampling algorithm.
- Understand the motivation for sequential Monte Carlo sampling algorithms for SSMs.
- Understand how resampling is used in the mathematical specification of
- Understand particle filters provide a means of extending SSMs to more complex probability distributions, for which posteriors are analytically intractable.
- Understand the mathematical derivation, and graphical illustration of particle filters through recursive time and measurement update steps.
- Understand that particle filtering can be used for switching SSMs.

- *Understand the specification of Rao-Blackwellised sampling.*

Week 8.

Lecture 14 – Approximate Inference: Markov Chain Monte Carlo.

- Understand how the limitations of Monte Carlo sampling methods (direct, rejection, importance) motivate the use of an adaptive proposal distribution.
- Understand that Markov Chain Monte Carlo (MCMC) methods use an adaptive proposal distribution.
- Understand the mathematical specification of the Metropolis-Hastings (MH) algorithm, such as the acceptance probability.
- Understand that the acceptance probability in MH is the ratio of importance sampling weights.
- Appreciate the mechanics of MH, the pseudocode, and the concept of "burn-in".
- Reinforce understanding of the following concepts needed to formally specify Markov chains: Markov property, transition kernels, time-homogeneity, state distribution, transition, stationary distributions.
- Understand the following properties of Markov chains: irreducibility, aperiodicity, ergodicity; time-reversibility, detailed-balanced conditions.

- Understand that ergodic Markov chains have a limiting distribution that is the stationary distribution, and the significance of this for convergence of MCMC algorithms.
- Understand that time-reversibility is a sufficient condition for a Markov chain to have a stationary distribution.
- Understand how to use these results to give a theoretical guarantee that MH converges to a stationary distribution that is our target distribution.
- Understand that there are currently no theoretical results on the duration required for convergence of MH to occur (burn-in), and that convergence is assessed heuristically.
- Understand the mathematical specification of the Gibbs sampling algorithm, in particular, as sequential sampling from conditional probability distributions.
- Understand that from a GM perspective, Gibbs sampling reduces to sampling from conditional probability distributions of a node given its Markov blanket.
- Understand how collapsed Gibbs sampling allows for sampling from topic models, thereby providing an inference algorithm.
- Understand how Gibbs sampling is a special case of MH.
- Understand practical issues and heuristic tools when implementing MCMC methods:
- Evaluating the suitability of a proposal distribution:
    - Understand the trade-off between acceptance rate and variance of proposal distributions.
    - Know some baseline heuristics on acceptance rates for selecting proposal distributions.
    - Understand how autocorrelation functions, autocorrelation plots, sample size inflation factor, and effective sample size can be used to empirically assess the suitability of a proposal distribution.
- Stopping "burn-in":
    - Understand how time-series plots of multiple MH runs can be used to characterise well-mixed and poorly-mixed chains.
    - Understand how log-likelihood plots can be used to monitor convergence.
- Understand that "vanilla MCMC" is characterised by random walk behaviour.
- Understand how optimisation and gradient-based methods can be leveraged to augment MCMC with Hamiltonian dynamics.
- Understand the mathematical specification of Hamiltonian Monte Carlo, in particular, position, momentum, potential energy, kinetic energy, Hamiltonian.
- Understand that Hamiltonian MCMC can yield better mixing and convergence properties.
- Understand MCMC with Langevin dynamics as a special case of Hamiltonian Monte Carlo methods.

- *Understand that MCMC can be augmented by making the proposal distribution a variational approximation to the target distribution (variational MCMC).*
- *"Finding scientific topics", Griffiths and Steyvers (2004)*

<u>Lecture 15 – Statistical and algorithmic foundations of deep learning.</u>

- Appreciate the historical significance of the paper by McCulloch and Pitts (1943).
- Reinforce understanding of the perceptron, perceptron learning algorithm, backpropagation.
- Reinforce understanding of the modern building blocks of deep neural networks e.g. activations, layers, loss functions.

- Appreciate the similarities and differences between graphical models and deep neural networks, with emphasis on representational function of the graph, and parameter estimation and inference.
- Understand the distinctions between the conceptual utility of graphs/networks in graphical models and deep neural networks.
- Understand that a number of neural networks are graphical models: Boltzmann machines, Restricted Boltzmann machines (RBM), sigmoid belief networks (SBN), deep belief networks (DBN), deep Boltzmann machines (DBM).
- Understand the mathematical specification of RBMs, in particular, the joint distribution, log-likelihood, and log-likelihood gradient.
- Understand that RBMs are fully connected MRFs over a bi-partite graph with a layer of hidden and observed variables.
- Understand how exact and approximate sampling methods are used as a subroutine in stochastic gradient descent when training RBMs.
- Understand the mathematical specification of SBNs.
- Understand that SBNs are simple BNs over binary variables with CPDs represented by sigmoid functions.
- Understand how "explaining away" provides insight on (approximate) inference and parameter estimation in SBNs, in particular, convergence of Gibbs sampling.
- Understand how training an RBM via alternating block Gibbs sampling is equivalent to training an infinitely deep SBN with tied weights.
- Understand the mathematical specification of DBNs.
- Understand that DBNs are hybrid graphical models consisting of RBMs stacked on top of sigmoid networks with untied weights.
- Understand how "explaining away" in DBNs renders exact inference problematic.
- Understand the training procedure of DBNs – greedy layer-wise pre-training followed by fine-tuning.
- Understand that DBMs are fully undirected MRFs which can be trained similarly to RBMs via MCMC.
- Appreciate that those in the deep neural network community are more concerned with empirical performance of the network on a downstream task, and with exploration of architectures and their components.
- Appreciate that those in the graphical models community are more concerned with correctness of inference and parameter estimation, and convergence.
- Appreciate that there may be further and richer, more fruitful connections between graphical models and deep learning.