PROBABILISTIC MODELS IN MODERN AI

1. Foundations: What AI Means

[What is the goal of AI] [To construct dynamical systems that will process information, appropriately reacting to inputs, and to have some meta-program explain how this dynamical system will be shaped from an environment (either static, with data, or dynamic, or both)] [They will construct a representation of the data that allows for some downstream tasks that are not describable using usual programs] [Modern AI is about a search in the space of programs, and the identification of those that will do well] [Statistics as a field is about the same things, but the boring part, i.e. the emphasis on small problems with fairly trivial programs] [What is an LLM? It is an assignment of probabilities of tokens, but dynamical as a function of the past]

1.1. **Goals of Modern AI.**

- [Create programs that we cannot do otherwise] [Modern AI seeks to build programs that autonomously construct representations of data and act on them under uncertainty] [That can play games and do clever things] [There are quite many functions out there; how do we find them is usually the question]
- [Intelligence is the ability treat streams of information and to find interesting things in them (with regards to some goals)] [Some central questions are the ability to predict, compress, denoise; the ability to find some meaning]
- [Information theory is the foundation of what we do] [Information theory is about what can be done theoretically with information] [In practice, it may be a different question, but that's a baseline]
- [Modern AI is in my opinion at the intersection of information theory and practical optimization] [Information theory is in some sense the limit of what we can do]
- [The most exciting general tasks are prediction, compression, denoising]
- [We should transform tasks into optimization problem then] [Optimization problems applied to labeled or unlabeled data or to environments] [And then there will be the questions of what are the consequences of the choices made for this]
- [Let's start with prediction] [What does it even mean?]

1.2. **Machine Learning and Prediction.**

- [To find the right functions, we need an environment: either static, or dynamic]
- [The question in statistics is: what is the chance that we are right at predicting things]
- [The goal in AI is a bit broader, it is to construct things that work] [Information theory is: how well things could work for certain tasks, with the right algorithm]

- [Static environment] [Supervised Learning] [Unsupervised Learning] [Supervised, semi-supervised, unsupervised, self-supervised learning, reinforcement learning]
- [Dynamic Environment] [Supervised versus Reinforcement Learning]
- [All of this is in some sense about prediction]
- [What we need is a clear objective]

1.3. Prediction and Scoring Rules.

- [There are some simple tasks that are surprisingly deep] [If we learn how to do them, we can solve a lot of exciting problems]
- [Prediction, infilling, backward prediction]
- [Supervised learning is a particular case of prediction]
- [Prediction] [A fundamental task of AI is prediction] [Note that it is not the only task] [Other tasks can be considered]
- [Scoring Rules] [First and second-order relations]
- [Proper scoring rules] [Examples] [Expected score function] [Convexity of expected score function] [From convex function to scoring rule] {Legendre transform} [Savage representation] [Bregman divergence with respect to a convex function] [Asymmetry] [Minimization]
  - {Classification of proper scoring rules}
- [Why the logarithmic scoring rules are better] [Modularity]
  - {Locality} {Chain rule}
  - {Market scoring rules} {Modularity from Hanson's paper 'Logarithmic Market Scoring Rules for Modular Combinatorial Information Aggregation'} {Logarithmic is modular} {Modular implies logarithmic}

1.4. Compression.

- [If there is a single task that is now understood to be associated with intelligence/understanding, it is compression] [The ability to synthesize data is what physics is about] [There is some theoretical ideas as to why the shortest explanation is the best (Occam's Razor)]
- [Entropy, conditional entropy, cross-entropy, mutual information] [Basic properties]
- [Compression] [Summarization] [Synthesis] [Summarize the internet data]
- [Lossless vs Lossy] [Problem of lossy is how to define what we need] [Focus on lossless in this course, because it is easier, and it applies to text]
  - {Huffman, Shannon codings} {McMillan Inequality}
- [Source coding theorem] [Statement: if there is this much information in a channel, you can find a code that compresses it with that many bits] [Proof: one direction, one can use some kind of coding, the other direction is based on the typical sets]
- [Imperfect coding] [Cross-entropy and practical compression] [Kullback Leibler and Gibbs inequality] [Very related to prediction is compression]
- [Classical lossless compression]
  - {LZ Compression} {LZ compresses ergodic things optimally} {Burrows-Wheeler} {Invertibility of BW} {Good properties on text}
- [Arithmetic coding] [Idea]
  - {Optimality proof}
- [Bits back coding] [Idea] [Bits back coding chaining]

- [Asymmetric numeral systems] [Construction and inversion] [What is the point of this]
  - {Proof that construction and inversion work}
- [Now, if we allow for lossy compression, we get into the realm of denoising]

1.5. Denoising and Sampling.

- [Continuous entropy]
- [Denoising] [Error correction] [Distortion function] [Rate distortion and distortion rate] [Achievability] [Rate distortion theorem]
- [Sampling] [Related via diffusion models]
- [Variational auto-encoders] [Diffusion models]
- [Is there an optimal way to do that in theory?] [Turns out that the answer is an incomplete yes]
- [Now, the question is a little bit: how do we extract what we need from data?]

1.6. Algorithmic Information Theory.

- [The information theory point of view studies how right we could ever be] [The information theoretic lens looks at data and tells us how much there is to do]
- [There is the question of what we could ever know] [Note that it is not because an information is in principle available that it can be found]
- [Why does the information point of view prove to be more useful?] [It abstracts away the specific algorithm that we use; as opposed to statistics, which tends to focus on constructed quantities] [Information theory is about what could be achieved, theoretically]
- [Kolomogorov-Solomonoff-Chaitin foundations] [Motivation]
- [Kolmogorov complexity definition] [Conditional Kolmogorov complexity] [Universality of Kolmogorov complexity] [Upper bound on conditional complexity by length] [Upper and lower bounds on Kolmogorov complexity] {What we think is unlikely as a random configuration} [The shortest description is upper bounded by the entropy] [The entropy is upper bounded by entropy by Kraft's inequality and source coding]
- [Martin-Löf probability] [Key Concepts] [Examples]
- [Minimum Description Length] [Circuit Length Description]
- [Understand the world generally speaking] [How to process information in a computable way]
- [Information theory comes with a few tasks: compressing and denoising] [Compressing is understanding]
- [Solomonoff induction] [What Solomonoff's induction is about] [Why it is uncomputable]
  - {Kolmogorov Sampler By Donoho}
- [Once we have done this thing, there is the question of what we could do with that] [Later led to AiXi]

1.7. Towards Universal AGI.

- [AiXi: a theory of RL based upon Solonomoff ideas]
- [Gödel Machines]
- [Now in practice, what do we do]

1.8. What this course will be about.

- [What we do in practice: neural network models that are trained on data]
- [Explain]

2. Neural Networks

2.1. Architecture.

- [The idea is to compose linear maps and nonlinear maps]
- [We take vectors as inputs, and we get vectors as outputs] [To do something interesting, we then need some code that uses that] [For instance, for LLMs, the output would be used to sample tokens]
- [Universal approximation results]
- [Why the architecture is not the only important thing]
- [Early misconceptions about neural networks]

2.2. Optimization.

- [Need to fix a selection process with an objective] [The true objective may be different, but we need to select a decent surrogate objective]
- [Saddlepoint problem] [Random initialization] [Gradient descent] [Adam]
- [The specification of a model must involve the optimization task]
- [How long do we run the optimization?]
- [Abstract formulation of gradient with a kernel]
- [Question of large neural networks] [Wrong conjectures]

2.3. Infinite-Width Limit.

- [Naive infinite-width limit blows up]
- [Activation kernel scaling regime]
- [Neural tangent kernel description]
- [Law of large numbers]
- [Stability during training]

2.4. Kernel Description.

- [The infinite width of neural networks in the kernel regime]
- [The activation kernel]
- [Random features]
- [Gaussian process prior and posterior for kernels]

2.5. Consequences.

- [Global minima]
- [Double-descent phenomenon]
- [Generalization]
- [Fine-tuning regime]

3. Large Language Models

3.1. Auto-Regressive Language Models.

- [Loss function]
- [Information extraction]

3.2. LSTMs, GRUs, Transformers, Mamba.

- [Transformers] [Started with Neural Turing Machines] [Then Bert] [Then GPT]

3.3. Information Theory.

3.4. Compression.

- [Arithmetic Encoding]

3.5. Arrows of Time.

4. Diffusion Models

4.1. Framework.

4.2. Variational Auto-Encoders.

4.3. Denoising.

4.4. Stochastic Calculus.

4.5. Diffusion and Bits Back Coding.

5. Alternative Paradigms

5.1. GANs.

5.2. Causality.

5.3. Bayesian Flow Networks.

6. Reinforcement Learning: Towards AGI

6.1. AiXi.

6.2. Capability Measures.

6.3. Games.

6.4. Transfer Learning.

6.5. Generality.

6.6. Alife Ideas.