**Monte Carlo methods**

My understanding is that ancestral sampling makes one pass through a directed pass, conditioning a sample of one variable on previous variables, while Gibbs' Sampling uses Markov chains and multiple passes. The thing I am confused about is how you can actually sample from an energy-based model. How exactly does conditioning one variable on other variables in the graph help to sample from it? For example when you have P(y | x), I understand that x and y are neighbors in the graph, but what actually happens? How is x drawn, and how does this help compute y in the graph?

You choose an initial value of x based off the marginal distribution, which is possible since there are no constraints (b/c x is the first variable being sampled so it is not dependent on other variables)

How do different number of feature maps arise between convnet layers?

* The weights applied to each of the input feature maps yield outputs which are summed up to give 1 output feature map (see convolution network notes)

exWhy is positive definiteness a requirement for the Hessian to be used in a lot of methods?

* What information do the eigenvalues reveal?

Parameters which encode conditional variance or some property of a prediction - how exactly do these work and how are they incorported? Do they affect the model's predictions

How is momentum incorporated in RMSprop / Adam?

Conjugate gradients - why do the search directions dt and dt-1 interfere w/ one another? (Review after directional derivatives)

* Line search => step size / learning rate

Does transpose w

Once flattened, each of the weights has its own partial derivative

What are these code branches?

Fig 16.9

Specifically, how d-separation works (when are two things d-separated)

How sampling actually works with ancestral sampling (intuition / how probabilities are used to compute the variables)

* The first variable is drawn from a marginal distribution
* In order to have accurate samples we must consider the dependencies between variables and conditional distributions, so when drawing a sample (a variable vector), we must see how x1 affects x2 and so on
* This is why we have connections, ideally in directed models which are more efficient

Posterior inference is computing probabilities of a sample, which can be done w/ sampling