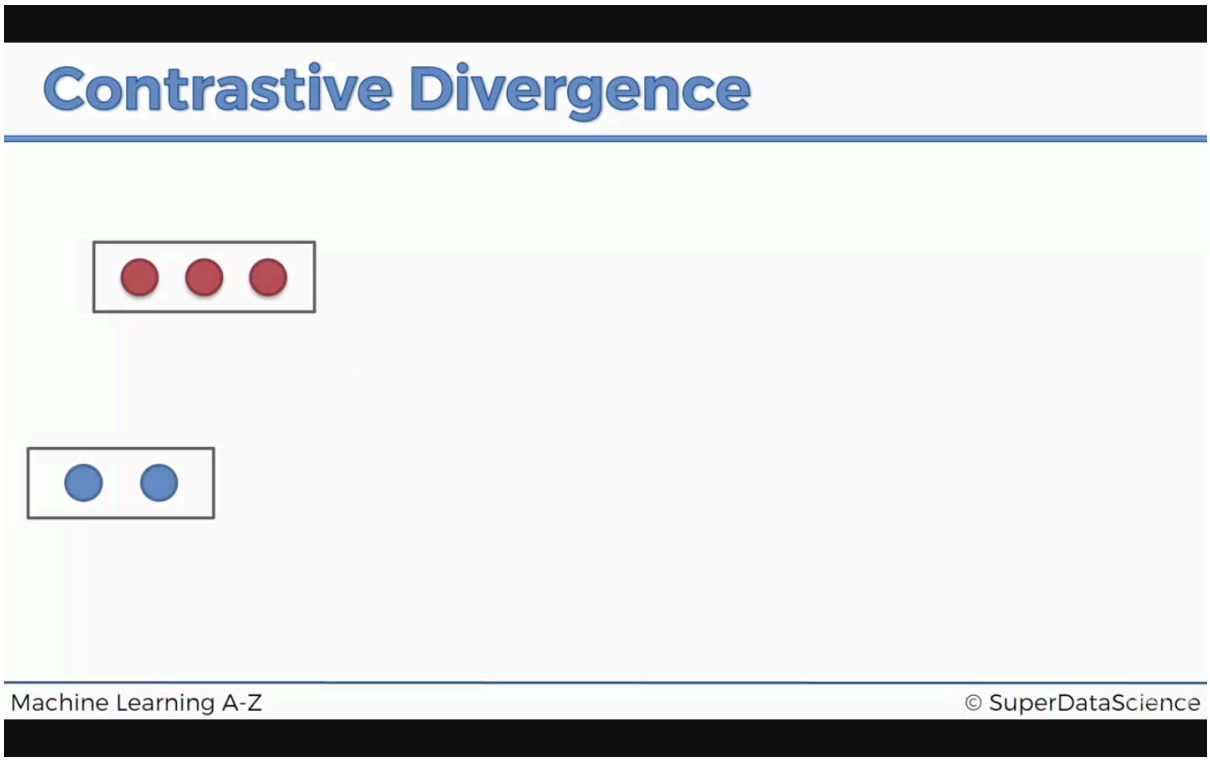
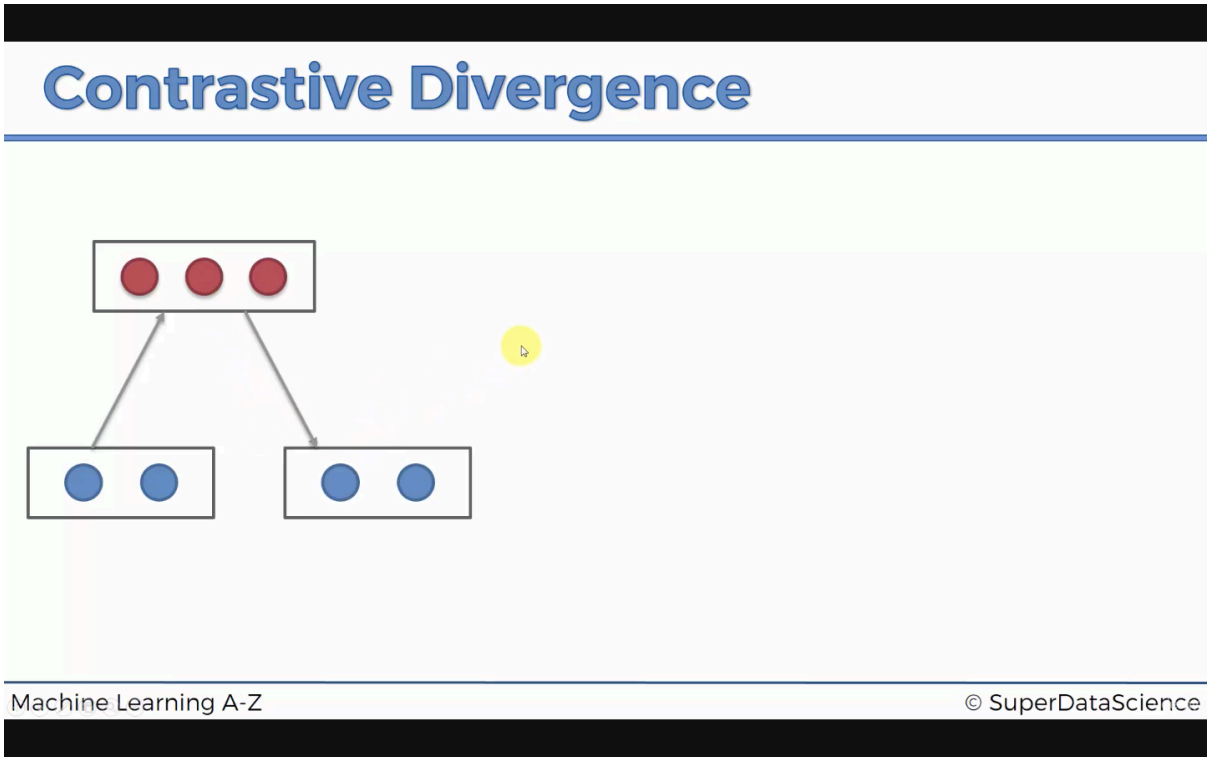
Contrastive divergence

This is the algorithm that allows restricted boltzmann machine to learn.

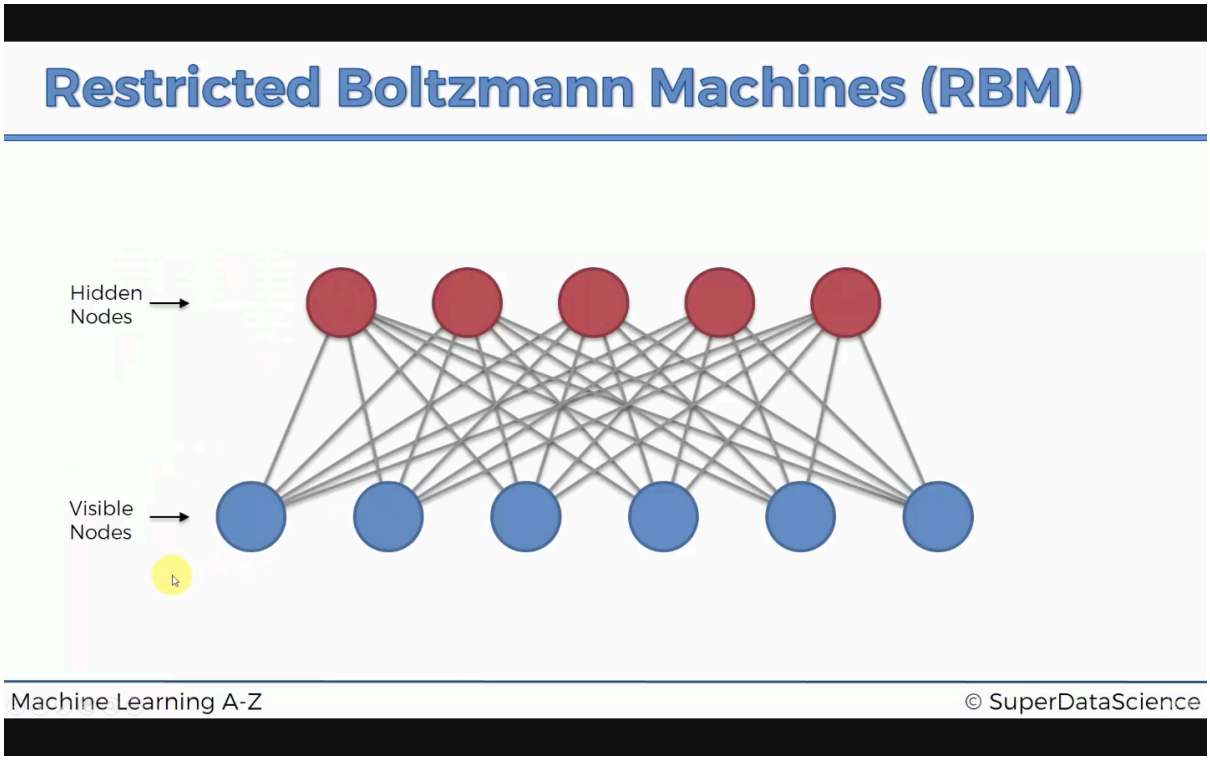


This is diagrammatic representation of our RBM. Blues are input nodes, and red are hidden nodes. Up until now we know about the learning process but how does RBM adjust its weights, is a question. In NN we used gradient descent through backpropagation but since in here we don't have directed network, Contractive Divergence comes in.

At start, once we put the inputs in the network, using some randomly assigned weights at the very start of system. the RBM calculates the hidden nodes. And then those hidden nodes by using exact same weights try to calculate the input nodes (or to reconstruct the input nodes)…

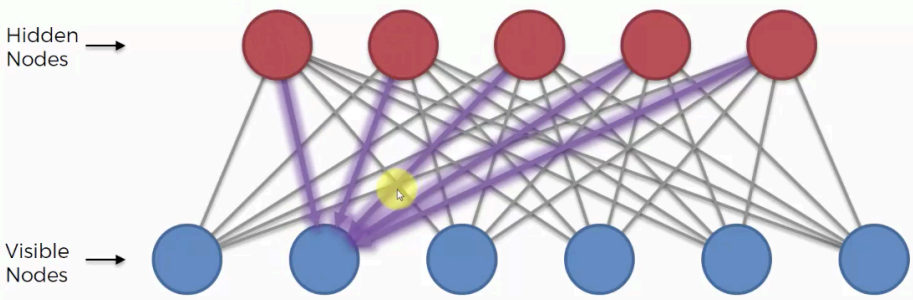


Note that reconstructed nodes are not going to be equal to original nodes.



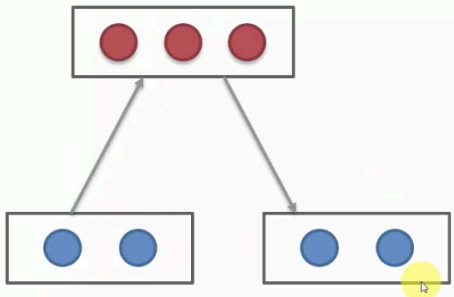
Now let's come back and to see why constructed nodes are not equal to original… well the reason for that is because these nodes are not initially interconnected.

Restricted Boltzmann Machines (RBM)

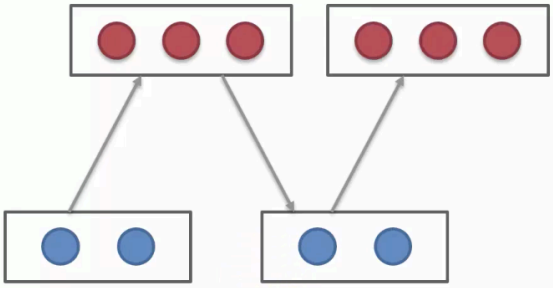


for example, let's consider the node 2; it gets restricted based on the values that all of these hidden nodes have in them. Important thing to learn is that these five hidden nodes are to construct only based on the visible node 2, if this was the case then the nodes would be identical. The hidden nodes are constructed based on all of the visible nodes.

Contrastive Divergence

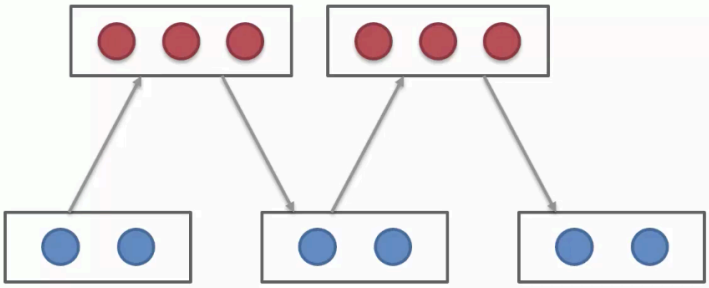


Contrastive Divergence

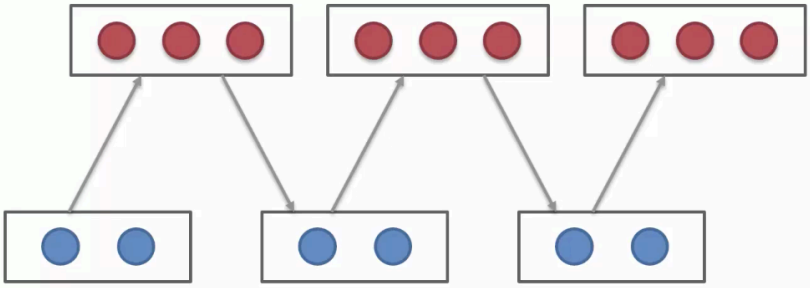


In here, we reconstruct the values of our input into the RBM and we get some hidden values.

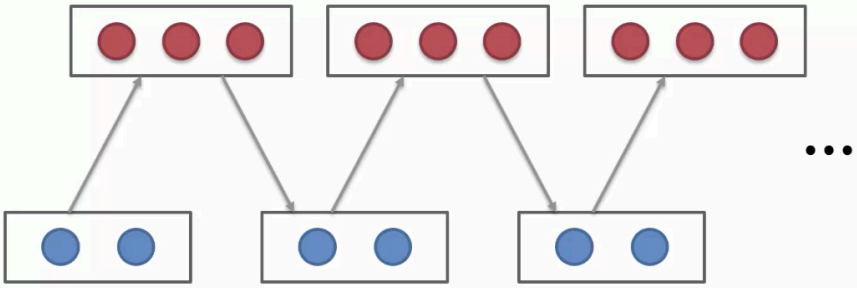
Contrastive Divergence



Contrastive Divergence

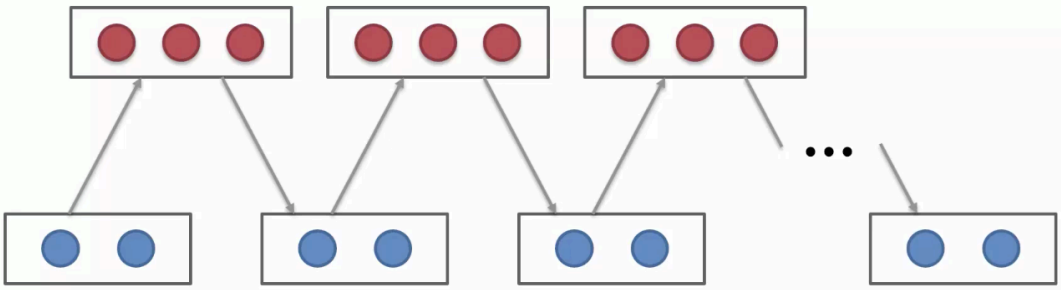


Contrastive Divergence



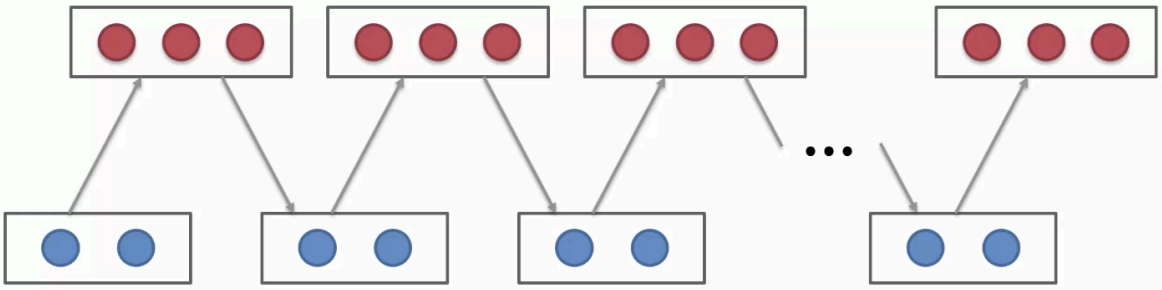
This whole process is called Gibbs Sampling

Contrastive Divergence



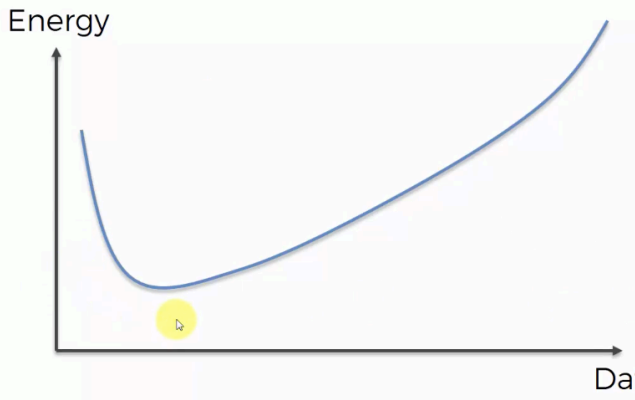
At the end, we get some reconstructed input values which if we feed them into the RBN…

Contrastive Divergence



and try to reconstruct them again, we get the same input (it goes back to its input). So, in our final scenario, our network is modelling exactly our input. Remember in this process we haven't changed the weights

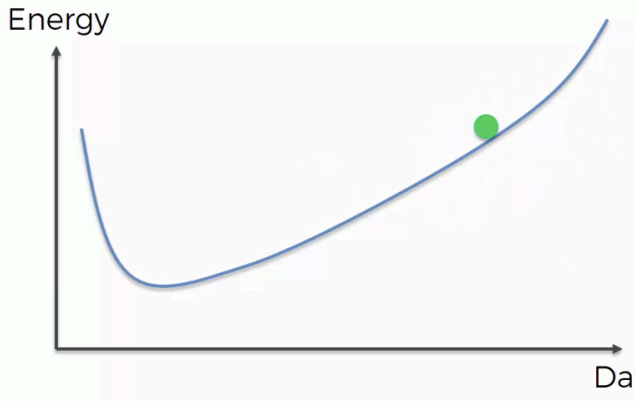
Contrastive Divergence



$$\frac{\partial \log p(\mathbf{v}^0)}{\partial w_{ij}} = <v_i^0 h_j^0> - <v_i^\infty h_j^\infty>$$

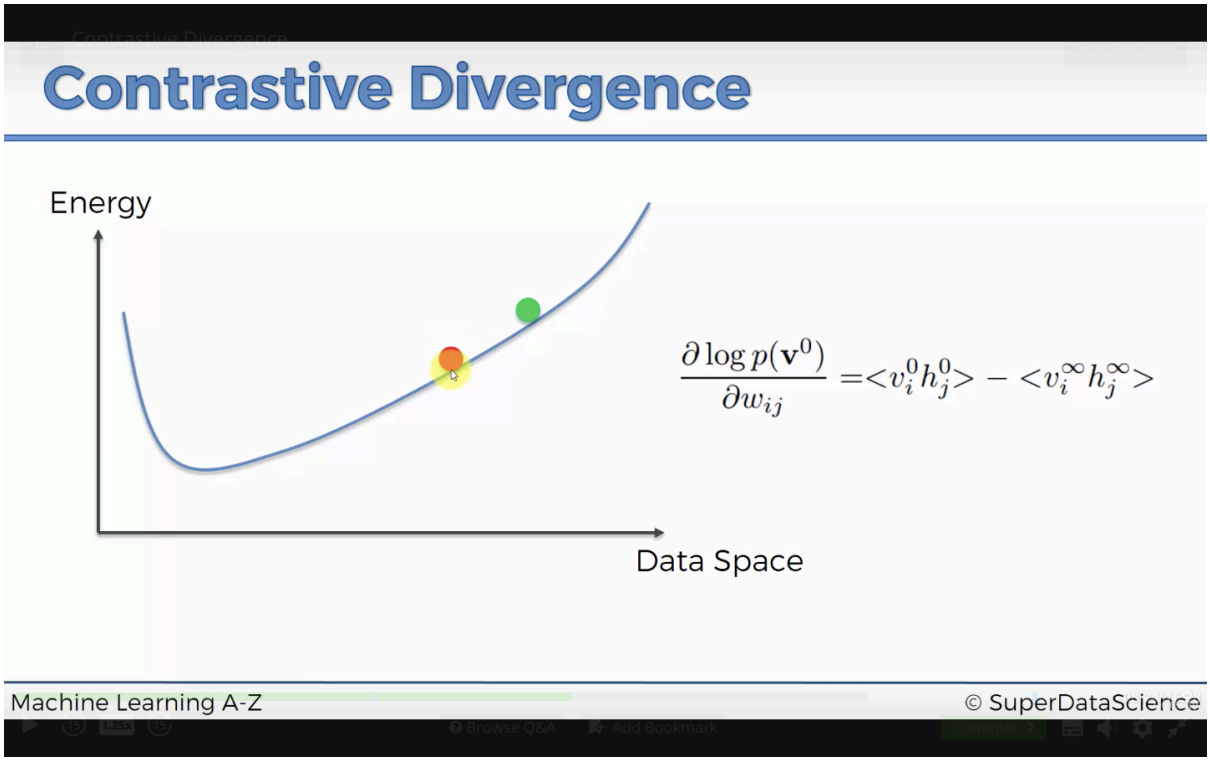
In case of curve, this is what it looks like. This is a gradient formula for adjusting, modifying the curve, to make sure the state is inside an energy minimum. In RBM, energy is defined through weights (weights are fixed).
Gradient of log probability of a certain state / weights in our system = <initial state of our system which is 'visible layer'^{0} 'hidden layer'^{0}> -

Contrastive Divergence

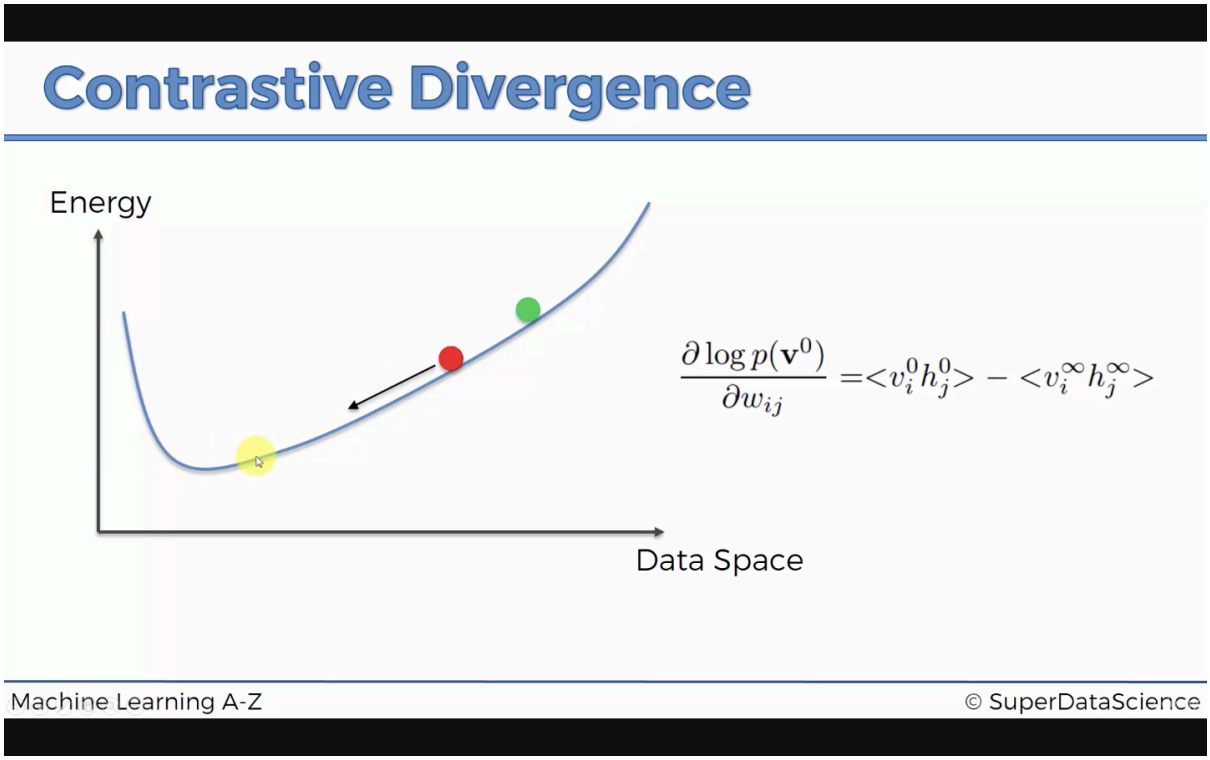


$$\frac{\partial \log p(\mathbf{v}^0)}{\partial w_{ij}} = <v_i^0 h_j^0> - <v_i^\infty h_j^\infty>$$

this is our initial state of system (or <v_i^0 h_j^0>) or in Gibbs sampling is when we are constructing for the first time.

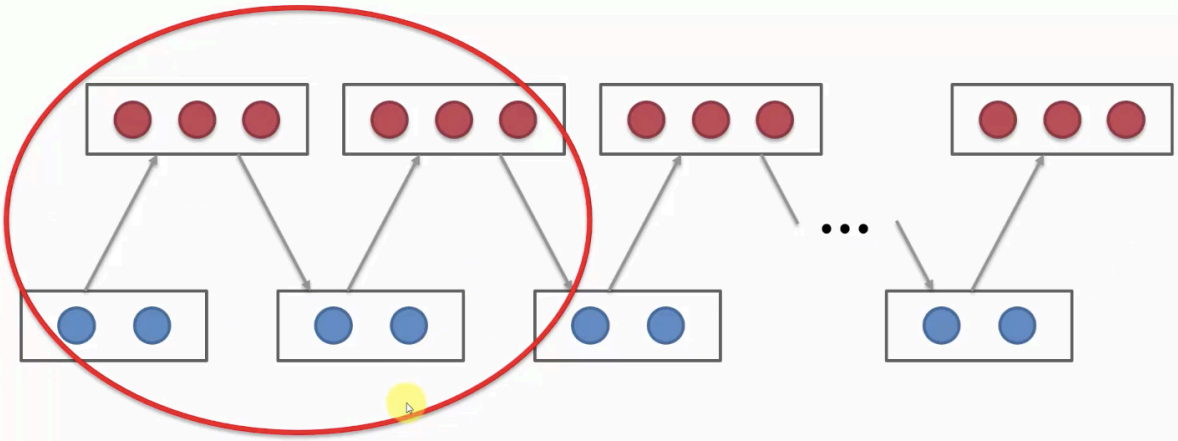


This is when we are constructing from second visible node.



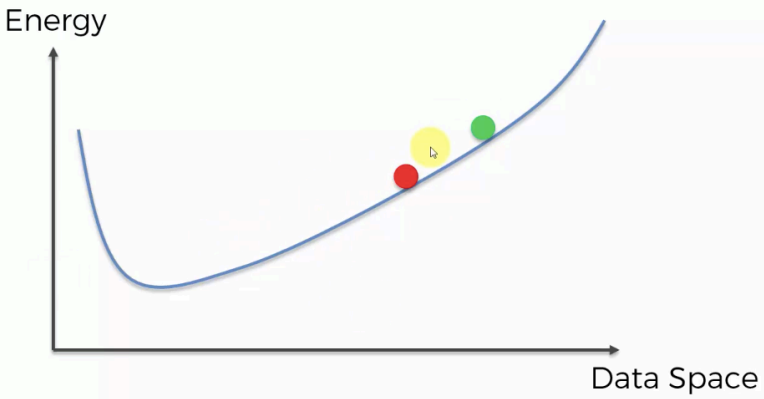
a system which is govern by its energy will always try to end up in the lowest energy state possible (like the gas example). At the very end (our last visible node) the energy going to be at the lowest or in this plot, it's going to be at the bottom of curve.
When we reach to this, we can use the formula to calculate the difference between the initial and last state to tell you how adjusting our weights will affect the log probability of our initial state. So the objective is to modify the curve in a way that our initial state in at the minimum energy.

Contrastive Divergence



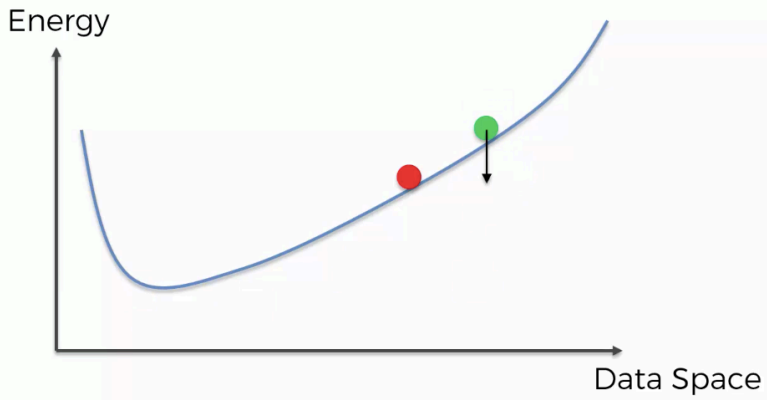
Since the whole process is too long, Jeffrey Hinton proposed a shortcut which in here, it's sufficient only to take the first 2 passes (and not to wait until it converges to the end) to adjust our curve.
We call this CD1 (Contrastive diversion one pass), we might even here CD2, CD3 and so on.

Contrastive Divergence

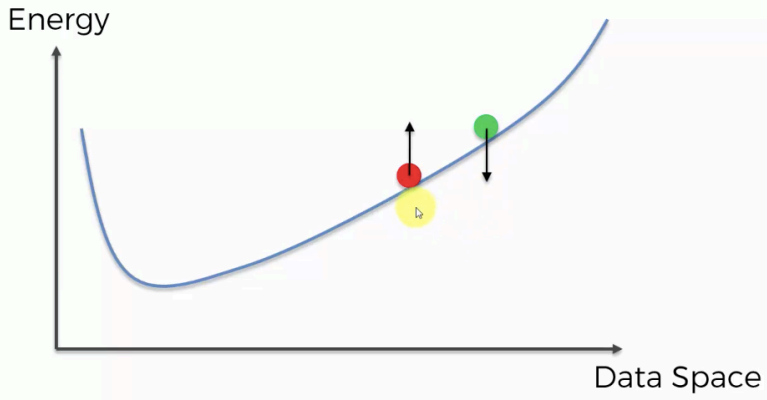


This is enough to know in which way we're rolling.
So to sum up, in here because it's an energy based system, we are adjusting our curve so our initial state is at minimum energy.

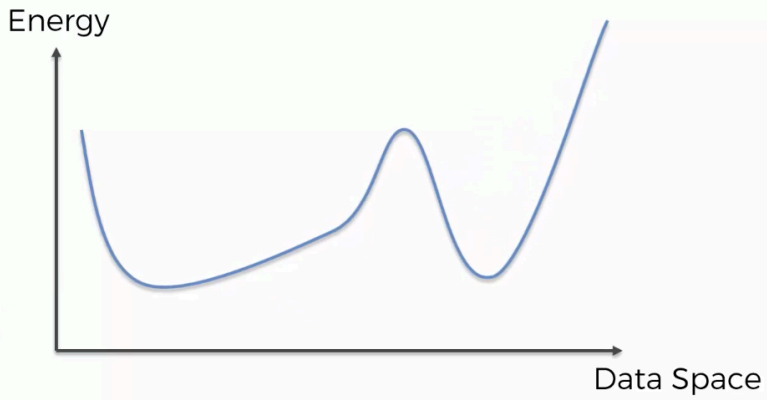
Contrastive Divergence



Contrastive Divergence



Contrastive Divergence



Contrastive Divergence