Toward the Implementation of a Quantum RBM

Misha Denil and Nando de Freitas
University of British Columbia, Computer Science

Motivation

- Boltzmann Machines are very general and powerful models of structure, but training an unrestricted Boltzmann Machine is very difficult.
- Restricted Boltzmann Machines make learning tractable by enforcing an advantageous conditional independence structure between layers in the model. Tractability comes at the cost of reduced expressive power.
- Boltzmann Machines are formally equivalent to the Ising model from statistical physics. This suggests the following training procedure:
  1. Transform the Boltzmann Machine into an Ising model.
  2. Set up a physical system which realizes the transformed problem and "run the physics" to allow the system to equilibrate.
  3. Measure the system to obtain a realization of states in the Ising model.
  4. Transform the Ising samples into samples from the Boltzmann Machine.
- D-Wave Systems has produced hardware to realize steps (2) and (3).

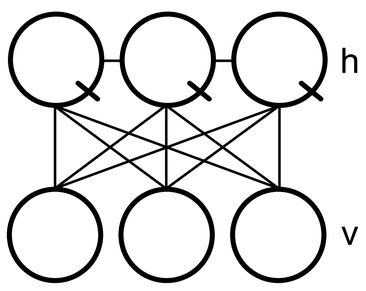
Boltzmann Machines

A Boltzmann Machine is a probabilistic graphical model defined on a complete graph of binary variables. The graph is partitioned into "visible" units $\mathbf{v}$, where values are observed during training and "hidden" units $\mathbf{h}$, where values must be inferred. The probability of observing a state in the Boltzmann Machine is governed by its energy function

$$E(\mathbf{v}, \mathbf{h}) = -\sum_{ij} W_{ij} v_i h_j - \sum_{jk} U_{jk} h_j h_k - \sum_{i\ell} L_{i\ell} v_i v_\ell$$

We study models with different graphical structures between the hidden units.

- Chain structured hidden connections, which are easy to sample from
- Chimera structured hidden connections, which reflect the architecture of the D-Wave machine.



Units intended to be realized on the hardware are marked with a small diagonal bar.

Quantum Annealing

Quantum Annealing solves problems by encoding them in a physical system

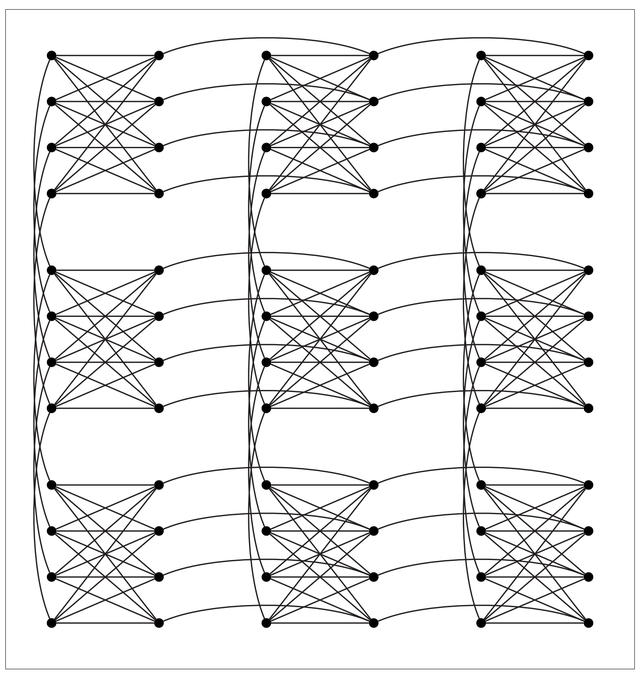
$$F[\rho] = \underbrace{\operatorname{tr}(\rho V)}_{\text{Problem term}} + \underbrace{\Gamma \operatorname{tr}(\rho K)}_{\text{Quantum term}} + \underbrace{T \operatorname{tr}(\rho \ln \rho)}_{\text{Classical term}}$$

- $\rho = |x\rangle\langle x|$ is the density matrix, where x is a distribution over the state of the system. If the state space has n bits then ρ has dimensions $2^n \times 2^n$.
- V is the problem Hamiltonian, which subsumes W, U and L in a Boltzmann Machine.
- K is a quantum disordering term which is minimized when all states are in uniform superposition.
- Γ is a scalar annealing parameter.
- T is the temperature.

In the D-Wave machine...

- V is set by the user.
- Γ is large at the beginning of computation and small at the end.
- The classical term is very small ($T \approx 0$).

D-Wave Hardware



The D-Wave hardware realizes an Ising model with a type of graphical structure called a Chimera.

A Chimera(M, N, L) graph is formed by connecting an $M \times N$ grid of $L \times L$ dense bipartite graphs.

The picture on the left shows the connectivity pattern of a Chimera$(3, 3, 4)$ graph.

The D-Wave machine realizes a larger Chimera$(4, 4, 4)$ graph. Future versions of the hardware will implement larger Chimera graphs.

$$\text{Chimera}(N, M, L) = \mathbf{I}_M \otimes \mathbf{I}_N \otimes \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix} \otimes \mathbf{1}_L + \mathbf{L}_M \otimes \mathbf{I}_N \otimes \begin{bmatrix} 1 & 0 \\ 0 & 0 \end{bmatrix} \otimes \mathbf{I}_L + \mathbf{I}_M \otimes \mathbf{L}_N \otimes \begin{bmatrix} 0 & 0 \\ 0 & 1 \end{bmatrix} \otimes \mathbf{I}_L$$

Backward Filtering Forward Sampling

Suppose we have a Restricted Boltzmann Machine with chain structured hidden units whose energy function is given by

$$E(\mathbf{v}, \mathbf{h}) = -\sum_{ij} W_{ij} v_i h_j - \sum_{|j-k|<2} U_{jk} h_j h_k - \sum_i L_{ii} v_i \ ,$$

Because of how the hidden connections have been restricted, the conditional distribution over the hidden units factors as

$$p(\mathbf{h}|\mathbf{v}) = p(h_1 = 1|\mathbf{v}) \prod_{j=2}^{|\mathbf{h}|} p(h_j = 1|h_{1:j-1}, \mathbf{v}) = p(h_1 = 1|\mathbf{v}) \prod_{j=2}^{|\mathbf{h}|} p(h_j = 1|h_{j-1}, \mathbf{v}) \ .$$

The marginal distribution for $p(h_1 = 1|\mathbf{v})$ can be found via a backwards pass of belief propagation. Starting at the end of the chain, for each $j = |\mathbf{h}| - 1, \ldots, 1$ we compute

$$m_{j+1\to j}(h_j) = \sum_{h_{j+1}} \phi_{j+1}(h_{j+1}) \psi_{j+1}(h_j, h_{j+1}) m_{j+2\to j+1}(h_{j+1})$$
$$\psi_{j+1}(h_j, h_{j+1}) = (U_{j+1,j} + U_{j,j+1}) h_j h_{j+1}$$
$$\phi_j(h_j) = \sum_i W_{ij} v_i + U_{jj}$$

The marginal distribution $p(h_1 = 1|\mathbf{v})$ is then given by

$$p(h_1 = 1|\mathbf{v}) \propto \phi_1(h_1) m_{2\to 1}(h_1)$$

With this information we can start with $p(h_1 = 1|\mathbf{v})$ and use the factorization above to sample sequentially from the joint distribution over hidden units.

Parameter Warping

- If we ask the hardware for samples from $P(x|\theta)$ we get samples from $P(x|W(\theta))$, where W is an unknown non-linear function.
- $W : \mathbb{R}^{832} \to \mathbb{R}^{832}$, so estimating it is very hard (and we actually care about its gradient, which is even worse to estimate).
- W is caused by interactions between proximate parameter realizations in hardware, so we can take
- This is an engineering problem and will be solved in time, but it is still something we need to deal with right now.
- Our solution is to use an alternative objective function which can be optimized as a black box.
- Black box optimization allows us to avoid knowledge of W because the relationship between parameters and objective values is treated as completely opaque.
- Our optimizer minimizes the one step reconstruction error on the training set.

Simultaneous Perturbation Stochastic Approximation

- An algorithm for approximate gradient based optimization of noisy, differentiable, black box functions.
- Requires that the objective be differentiable, but does not require explicit access to the objective gradient.

Given an objective function $J(\boldsymbol{\theta})$ and an initial value $\boldsymbol{\theta}_0$, at time t during the optimization SPSA preforms the update

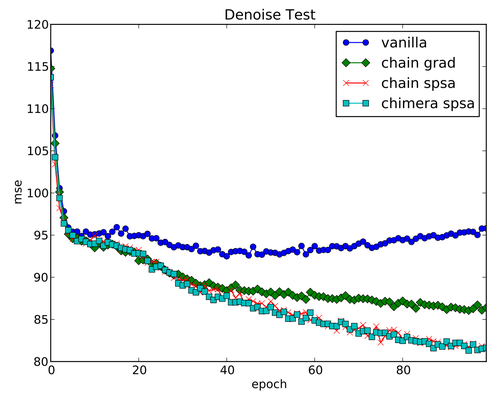
$$\boldsymbol{\theta}_{t+1} = \boldsymbol{\theta}_t - a_t g_t(\boldsymbol{\theta}_t)$$

where $g_t(\boldsymbol{\theta}_t)$ is a stochastic estimate of the gradient of $J(\boldsymbol{\theta}_t)$ given by

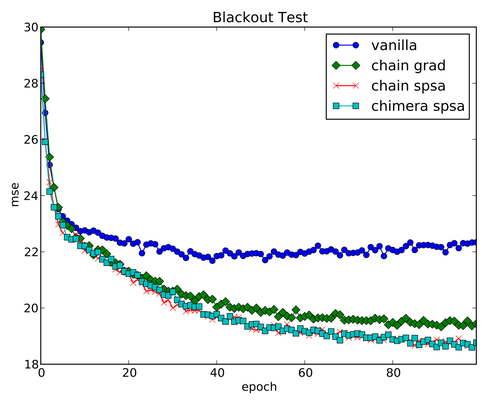
$$\nabla_\theta J(\boldsymbol{\theta}_t) \approx g_t(\boldsymbol{\theta}_t) = \frac{1}{2c_t}(J(\boldsymbol{\theta}_t + c_t \boldsymbol{\Delta}_t) - J(\boldsymbol{\theta}_t - c_t \boldsymbol{\Delta}_t)) \boldsymbol{\Delta}_t^{-1}$$

where $\boldsymbol{\Delta}_t^{-1}$ denotes the elementwise inverse of the vector $\boldsymbol{\Delta}_t$. At each step, each element of $\boldsymbol{\Delta}_t$ is chosen independently from $\{-1, 1\}$ with equal probability.

Experiments



Reconstruction error on denoising task



Reconstruction error on blackout task

- SPSA based methods achieve lowest error overall on both tasks.
- Adding chain structured latent connections improves performance over the baseline.

Selected References

- G. Rose and W. Macready. *An Introduction to Quantum Annealing*. D-Wave Systems, 2007.
- M. Johnson, M. Amin, S. Gildert, T. Lanting, F. Hamze, N. Dickson, R. Harris, A. Berkley, J. Johansson, P. Bunyk, et al. *Quantum annealing with manufactured spins*. Nature, 473(7346):194198, 2011.
- J. Spall. *Introduction to stochastic search and optimization: estimation, simulation, and control*. 2003.