Quantum Annealing meets Machine Learning

William Macready
The good news

- Exploiting quantum mechanics can dramatically accelerate certain computations
  - Factoring of an $n$ bit integer
    - Classically: $O\left(\exp\left(n^{1/3}\right)\left(\log n\right)^{2/3}\right)$
    - Quantum: $O\left(n^3\right)$ [Shor’s algorithm]
  - Blind search in database of $2^n$ items
    - Classically: $O\left(2^n\right)$
    - Quantum: $O\left(2^{n/2}\right)$ [Grover search]

The bad news

- It is difficult to build hardware that can support quantum algorithms
  - Largest experimentally realized version of Shor’s algorithm factored $21=7\times3$
The good news

• A recent computational model may offer a faster path to scalable quantum computation
  – Quantum annealing
  – A specialization of adiabatic quantum computation

• Certain problems (e.g. Grover search) can be accelerated now
  – In a nutshell: programmable hardware exploits quantum mechanics to quickly equilibrate to a Boltzmann-like distribution which can be rapidly sampled

• QA→ML:
  – new sampling and optimization capabilities may be used in machine learning applications

• ML→QA:
  – circumvent practical limitations of current hardware platforms
What’s ahead?

• QC introduction
• Quantum annealing
• Hardware implementation
  – benchmarking
• Domains of application (QC→ML):
  – Binary and structured classification
  – Sparse unsupervised learning
• Challenges (ML→QC):
  – Circumventing connectivity; richer models with hidden variables
  – Sampling when the sampling distribution is imperfectly known
  – Extending the range of applicability
**Idealized Quantum Mechanics** (zero temperature, no environment)

- **Key new ingredients:**
  - The state describing a physical system is a vector and measurements on the system are matrices which can potentially alter the state vector
  - QM is non-commutative

- **Single qubit system**
  - The qubit is the quantum analog of a bit and is described with a normalized 2-dimensional vector

If you measured a qubit in state $|\varphi\rangle$ you would observe 0 with probability $|\alpha_0|^2$ and 1 with probability $|\alpha_1|^2$.
Dynamics of many qubits

• With \( n \) qubits there are \( 2^n \) basis state vectors: \(|00 \cdots 00\rangle\) to \(|11 \cdots 11\rangle\)

• An arbitrary state is a normalized vector \(|\varphi\rangle = \sum_b \alpha_b |b\rangle\)
  
  \[ |\alpha_b|^2 \text{ is the probability of observing joint configuration } b = b_1 b_2 \cdots b_n \]

• An important operator acting on a state vector gives the energy, called the Hamiltonian, \( H \)
  
  \[ H \text{ is a Hermitian } 2^n \times 2^n \text{ matrix; in general } H(t) \text{ may vary with time} \]
  
  • Eigenvalues are real
  
  • \( H(t) \) determines how a state vector evolves in time:
    
    \[ \partial_t |\varphi\rangle = -iH(t)|\varphi\rangle \text{ [Schroedinger equation]} \]

  • When excess energy may be exchanged with an environment this dynamics acts to evolve state vectors to the eigenvector corresponding to lowest eigenvalue of \( H \) (minimize the energy)
Hamiltonians and Minimization

- We can solve an energy minimization problem $P$ by encoding the energy function on the diagonal of $H$

\[
H_P = \begin{bmatrix}
E_{0\ldots00} & 0 & 0 & 0 & 0 & 0 \\
0 & E_{0\ldots01} & 0 & 0 & \ldots & 0 \\
0 & 0 & E_{0\ldots10} & 0 & \ldots & 0 \\
0 & 0 & 0 & E_{0\ldots11} & 0 & \ldots \\
\vdots & \vdots & \vdots & \vdots & \ddots & \vdots \\
0 & 0 & 0 & 0 & \ldots & E_{1\ldots11}
\end{bmatrix}
\]

- lowest energy state $|b^\ast\rangle$ satisfies $H_P|b^\ast\rangle = E_b^* |b^\ast\rangle$; diagonalizing $H_P$ equivalent to minimizing $E_b$

- We’ll be focused on Ising energy functions:

\[
E_b = \sum_{i \in V} h_i b_i + \sum_{(i,i') \in E} J_{i,i'} b_i b_{i'}
\]

where $G = (V, E)$ is a graph of allowed variable interactions
Adding quantum mechanics...

- Quantum mechanics includes off-diagonal elements in $H$
  - Example realized in hardware acts to flip bits

\[
H = \begin{bmatrix}
E_{0\ldots00} & \Delta & \Delta & 0 & 0 \\
\Delta & E_{0\ldots01} & 0 & \Delta & 0 \\
\Delta & 0 & E_{0\ldots10} & \Delta & \vdots \\
0 & \Delta & \Delta & E_{0\ldots11} & 0 \\
\vdots & \vdots & \vdots & \vdots & \vdots \\
0 & 0 & 0 & 0 & E_{1\ldots11}
\end{bmatrix} = H_P + H_{od}
\]

Lowest eigenvector not aligned with any classical basis vector -- superposition.
Quantum annealing

• The optimization problem we want to solve is defined by $H_P$

• The inclusion of $H_{od}$ gives ground state eigenvectors which are linear combinations of classical states
  
  — Superposition: quantum mechanically we explore qubits assuming states which are both 0 and 1
  
  — This mechanism can be used to tunnel out of local minima in favour of better local minima

Diego de Falco and Dario Tamascelli [RAIRO-Theor. Inf. Appl. 45, 99 (2011)]
Use quantum effects to explore the search space

• Look to simulated annealing to exploit the exploration offered by quantum superposition

• Take time varying Hamiltonian

\[ H(t) = A(t/\tau)H_P + B(t/\tau)H_{od} \]

• Eigenbasis: \( H(t)\ket{\varphi_n(t)} = \lambda_n(t)\ket{\varphi_n(t)} \)

• Start in a ground state of \( H_{od} \)
  
  For this state all configurations \( \ket{b} \) are equally likely to be observed

• Slowly evolve ground state by turning up \( H_P \) and turning down quantum effects \( H_{od} \)
Quantum Annealing

Farhi et al., Science 292, 472 (2001)

\[ H(t) = A(t/\tau)H_P + B(t/\tau)H_{od} \]
What limits the speed of QA?

- Hardness of optimization problem manifested in a gap which may go to zero exponentially fast with the problem size

Like simulated (thermal) annealing:
   Equilibration time related to eigenvalue difference of transition matrix

**Evolution time:**

\[ \tau \approx \frac{\max_t |\langle \varphi_1(t) | H_{od} | \varphi_0(t) \rangle|}{gap^2} \]
How fast is QA?

• QA gives Grover’s quadratic speedup (Farhi et. al., Childs et. al.)
• QA easily simulates SA (Somma et. al.)
• There is also other experimental, numerical and theoretical evidence of speedups. (Brooke at. al., Kodawaki et. al., Matsuda et. al.)

Note: not simulating quantum annealing on classical hardware, but running on quantum hardware
A physical qubit

- Qubits are loops of superconducting wire (Josephson junctions)
- Direction of circulating current indicates the qubit states $|0\rangle$ and $|1\rangle$
- With external magnetic field we can bias towards one state or the other; linear terms in Ising model
- Auxiliary loop allows control of off-diagonal elements

Control the amount of superposition from quantum to classical bit; the $\Delta$ terms of $H_{od}$
Coupling qubits: a unit cell

- Qubits are stretched into long thin loops and coupled together
- Couplers give programmable pairwise coupling terms in Ising model
- Unit cell consists of 8 qubits
Tiling the chip with unit cells

4x4 array
C8 chip

• Next chip (available in September) has 8x8 array of unit cells
  — 512 qubits
  — Programmability: 512 h values; 1472 J values

• Duty cycle:
  — Programme h/J
  — Anneal
  — Readout

• Timing:
  — Programme + 1000 anneal/readout loops in <100ms

• Treewidth is 33
The full package

• Processor packaged on motherboard to connect to off chip elements
• Inputs coming from room temperature are filtered
• and system cooled to 20mK in a magnetically shielded environment (50000x smaller than earth’s magnetic field)
Practical realities: from ideal to realistic QM

• At non-zero $T$ an equilibrium system is described the density matrix: $\rho = \exp(-\beta H)/Z(\beta)$
  
  - Like probability density $tr(\rho) = 1$ and $\rho > 0$
  - Interactions in Hamiltonian’s are typically sparse and pairwise.
  - Quantum versions of conditional independence, Markov random fields, belief propagation etc.
  - Significantly complicated by the fact that “clique potentials” are operators and do not commute

• System never completely isolated from its environment
  
  - There is an interaction Hamiltonian with the environment and the hidden variables of the environment must be marginalized out
Prognosis: scalable quantum annealing?

- Speedups from quantum annealing still apply at non-zero temperature
  - In some cases inclusion of low temperature can help
  - At high temperature gains of QM are lost
  - Can get to low temperatures $E/k_B T \approx 3-5$

- Environmental coupling is more problematic
  - Shielding eliminates stray magnetic fields
  - Chip fabrication defects/impurities most significant
  - Modeling suggests current chip should work well at 512 qubits, but performance may degrade as chip scales unless chip imperfections can be reduced
  - Fortunately, noise reduction is linearly proportional to fidelity
    - If we can halve noise then we should obtain the same performance at 1024 qubits as available at 512 qubits
    - 10x noise reduction should be possible in the near term
Benchmarking

- Random Ising models on 4x4 chip
  - $h \in \{-3, -2, -1, 0, 1, 2, 3\}$
  - $J \in \{-3, -2, -1, 0, 1, 2, 3\}$ on hardware edges
- Exact grounds states determined by belief propagation / MIP
- Calculated run time to find ground state with 99% certainty

For small $N$ annealing time scaling linearly on 4x4 hardware

Early version of 8x8 hardware
Annealing time

A 108 variable Ising problem

Conclusion: prob. distribution peak shifts to right as interp. time increases
Consistent with adiabatic evolution

S. Boixo, Z. Wang, D. Lidar
Putting QA to work

• **<speculation>**
  
  – There will be QA hardware more widely available in the next 5 years that can address sparse Ising problems of up to 5000-10000 variables
  
  – Time to low energy solutions likely to be dramatically faster than is possible using classical hardware
  
  – The machines will be stochastic; i.e. returned values will be samples from some distribution

  </speculation>

• These machines will have constraints on the types of problems that can be natively addressed

  – Sparsely connected, but treewidth may be high (i.e. $tw > 120$)
  
  – Optimization will be unconstrained
  
  – Pairwise interactions
  
  – Problems requiring high precision specification of $h/J$ will be more difficult
  
  – There will be no closed form description of the sampling distribution
QA→ML: applications of QA

• Lots of optimization in ML, but the vast majority is continuous optimization
  – Relatively little exploitation of combinatorial optimization

• A few things we + collaborators have tried:
  – Structured classification
    • SSVM: \( y(x) = \arg\min_y \{ \langle h(x) | y \rangle + \langle y | J(x) | y \rangle \} \)
      – Use standard approach to learn \( h(x) \) and \( J(x) \) from training set; subgradients evaluated by quantum annealing
      – Convex optimization algorithms need to be slightly improved to accommodate potentially noisy subgradients
    • CRF: \( P(y|x) \approx \exp\{-\langle h(x) | y \rangle - \langle y | f(x) | y \rangle\} \)
      – Gradient with respect to fitting parameters requires expectations which we evaluate in hardware using importance sampling

  – Binary classification with new regularization (Neven et al)
    • \( y = \text{sign}(\langle w | c(x) \rangle) \) where weights \( \{ w_\alpha \} \) are Boolean valued, and \( \{ c_\alpha(x) \} \) are weak classifiers
    • Regularize using \( R(w) = \|w\|_0 = \langle 1 | w \rangle \)
    • Use squared loss \( L(w) = \sum_i [m_i(w) - 1]^2 \) where the margin is \( m_i(w) = y_i(\langle w | c(x_i) \rangle) \) then minimizing \( L(w) + \lambda R(w) \) is an Ising optimization problem for the optimal weights \( w \)

  – Unsupervised L0 dictionary learning
    • Factor a matrix \( X = DW \) by minimizing \( \|X - DW\|_{Fro} + \lambda \|W\|_0 \); all elements of \( W \) are Boolean-valued
    • Block coordinate descent on \( D \) then \( W \); each column of \( W \) is an Ising optimization
ML → QA: outstanding problems

• Extend applicability of QA hardware
  - Given a fixed factor graph develop methods to optimize objectives defined with different factor graphs
  - Blackbox optimization: develop methods for objectives not having a factor graph
    • i.e. black box optimization where objective function is code without a closed form expression

• Monte Carlo methods
  - Hardware is stochastic and we can sample i.i.d. very quickly
  - Unfortunately, the sampling distribution is not known exactly; although to lowest order it is roughly Boltzmann
Circumventing a sparse pairwise factor graph

- Native problems are pairwise and sparse
- Can always reduce higher-order interactions to pairwise, but at the cost of additional qubits
  - Qubits are a scarce resource: for certain problem types are there more efficient reductions?
- We can simulate connectivity by slaving qubits
  - Strong ferromagnetic couplings \(-\lambda s_i s_j (\lambda > 0)\) sets \(s_i = s_j\) in low energy solutions
  - New variables mediate interactions creating qubit “wires”
  - Not scalable as finding embeddings is NP hard
  - What to do?
Problem decomposition

• Even 10 000 qubits may be too small for many applications

• What are good approaches for decomposing large optimization problems down to a sequence of smaller problems
  
  — Lagrangian relaxation: ok for relatively simple problems; not very effective for harder problems
Monte Carlo

- Hardware acts as a source of fast i.i.d. samples from a tunable Boltzmann-like distribution
  - However, we do not have a closed form description of the sampling distribution
  - Are there methods to exploit hardware to adaptively shape the h/J input parameters to certain tasks?
    - Creating a proposal distribution for MCMC
    - Evaluating expectations
    - Estimating partition functions
Summary

• Quantum annealing machines offer opportunities for new classes of “tractable” problems
  – What new learning algorithms can be constructed that rely on solving sparsely connected combinatorial optimization problems?
  – Can Monte Carlo algorithms take advantage of samples from Ising models that are roughly Boltzmann distributed?

• For broadest applicability a number of key problems need to be addressed:
  – How can we effectively apply pairwise fixed-connectivity solvers to the solution of higher-order models and/or models with alternate variable connectivity?
  – How can we decompose larger problems into smaller manageable chunks

• Not new problems, but certainly new incentives for tackling some of these issues