Classification with Quantum Annealing on the D-Wave System

Quantum Computing for High Energy Physics workshop
CERN, 5-6 Nov 2018

Jean-Roch Vlimant
jvlimant@caltech.edu
California Institute of Technology
Outline

➢ Overview

• Quantum Annealing
• QA Machine Learning
• A Higgs dataset

➢ Experiments
➢ Outlooks
Overview
Solving a Higgs optimization problem with quantum annealing for machine learning

Alex Mott, Joshua Job, Jean-Roch Vlimant, Daniel Lidar & Maria Spiropulu

doi:10.1038/nature24047

Received: 04 April 2017  
Accepted: 28 July 2017  
Published online: 18 October 2017

- Computational science
- Experimental particle physics
- Qubits
- Theoretical particle physics

https://www.nature.com/articles/nature24047
Experiment

Quantum Annealing

Classification

SIGNAL

SIMULATION

BACKGROUND

AUROC

Size of training dataset ($10^3$)

D-Wave Classifier, OpenLab Q-HEP, J.-R. Vlimant

11/05/18
The D-Wave Computing System
The D-Wave Company

Welcome to the Future
Quantum Computing for the Real World Today

https://www.dwavesys.com/

1999  Founded
2011  D-Wave One : 128 qubits
2013  D-Wave Two : 512 qubits
2015  D-Wave 2X : 1000 qubits
2019? 5000 qubits ?
D-Wave 2X™

1098 qubits
Operates at 15mK
Anneals in 5-20 μs
qubit and qubit

Quantum Circuits
Series of quantum gates operating on a set of quantum states.

Quantum Annealing
Evolution of a quantum system to a low T Gibbs state

That's D-Wave!
Quantum Annealing
Adiabatic Quantum Annealing

- System setup with trivial hamiltonian $H(0)$ and ground state
- Evolve adiabatically the hamiltonian towards the desired Hamiltonian $H_p$
- **Adiabatic theorem**: with a slow evolution of the system, the state stays in the ground state.

$$H(t) = A(t)H(0) + B(t)H_p$$

D-Wave Hamiltonian
And
Chimera Graph
D-Wave Hamiltonian

\[ H_{\text{Ising}} = \sum_i h_i \sigma^z_i + \sum_{ij} J_{ij} \sigma^z_i \sigma^z_j \]

- External magnetic field
- Interactions

Runs over adjacent quBits
D-Wave qubit Adjacency

Active qubits in green
Coupling to 5-6 qubits
Inactive qubits in red
Not a fully connected graph
Model Embedding
Full Ising Model

- Create chains of spins through the chimera graph
- Split local fields across all qubits in the chain
- Tightly couple ($J_F = 6$)

- Non-unique embedding. Heuristic approach.
- Suppressing spin flip within chain as error correction.
- Use majority vote

→ Approximately full Ising Model with ~<40 spins

https://arxiv.org/abs/1210.8395
Ising Hamiltonian

\[ H_{\text{Ising}} = \sum_i h_i \sigma_i^z + \sum_{ij} J_{ij} \sigma_i^z \sigma_j^z \]

- External magnetic field
- Interactions
- Runs over all quBit pairs
D-Wave Sampling Solutions

User provides
\[( h_i, J_{ij} )\]

Multiple 5 μs annealing cycles

D-Wave provides sampling from lowest energy levels (approx. Gibbs)
\[( \{ \sigma_i \}_k, \mathcal{N}(\{ \sigma_i \}_k) )\]
Classification with Quantum Annealing
QA Machine Learning

Formalism to transform a binary classification into an Ising model hamiltonian optimization

K.L. Pudenz, D.A. Lidar

https://arxiv.org/abs/1109.0325
Define functions $h_i$ of the input variables into $[-1,1]$ such that

- $P(\text{signal}|h>0) > P(\text{bkg}|h>0)$
- $P(\text{bkg}|h<0) > P(\text{signal}|h<0)$

i.e. Most signal on $h>0$, most bkg on $h<0$

Define $w_i$ as binary linear combination of $h_i$

$$O(x) = \sum_i w_i h_i(x)$$

https://arxiv.org/abs/1109.0325
QAML Target/Objective

Define as a “target” function

\[ y(x) = \begin{cases} 
+1, & \text{if } x \in S, \\
-1, & \text{if } x \in B 
\end{cases} \]

Per event error

\[ E(x) = y(x) - \sum_{i=1}^{N} w_i h_i(x) \]

Full error

\[ \delta(\bar{w}) \propto \sum_{i,j} C_{ij} w_i w_j + \sum_i (\lambda - 2C_{iy}) w_i \]

- \( C_{ij} \) and \( C_{iy} \) are summations over the values of \( h_i \) over the training set
- \( \lambda \) is a parameter penalizing the number of non-zero \( w_i \)

https://arxiv.org/abs/1109.0325
QUBO
Quadratic Unconstrained Binary Optimization

\[ \delta(\vec{w}) \propto \sum_{i,j} C_{ij} w_i w_j + \sum_i (\lambda - 2C_{iy}) w_i \]

Simple conversion of binary weights to \( \pm 1 \)

\[ H_{\text{Ising}} = \sum_i h_i \sigma_i^z + \sum_{ij} J_{ij} \sigma_i^z \sigma_j^z \]
QAML End-to-End
QAML Discriminant

Signal and Background samples. Less than 40 features

Mask \( (w_i) \) of features contributing to

\[
\sum_{i=1}^{N} w_i h_i(x)
\]

continuous discriminant function. One mask per energy level.
A Higgs-$\gamma \gamma$/background dataset
Generated Samples

**SIGNAL**

Generated with PYTHIA 6.4 at 8TeV proton-proton c.o.m energy

**BACKGROUND**

Generated with SHERPA at 8TeV proton-proton c.o.m energy
- Photon pT of 32 GeV and 25 GeV for realistic trigger selection
- Di-photon mass [122.5, 127.5] GeV
- Higgs candidate $|\eta|<2.5$
Sample Size and Folding

- 300k signal + 300k background total sample
  - **Training set**
    - 20 stratified, independent splits of sizes 100, 1000, 5000, 10k, 15k, 20k events
    - Spread of classifier performance over the folds reported as the *uncertainty due to the choice of training sample, and initialization*.
  
  - **Testing set**
    - Remaining 100k+100k independent sample
    - *Statistical error* on the classifier performance estimated using bootstrapping
Characterizing Variables

<table>
<thead>
<tr>
<th>variable</th>
<th>description</th>
</tr>
</thead>
<tbody>
<tr>
<td>$p_T^{1}/m_{\gamma\gamma}$</td>
<td>transverse momentum of the highest $p_T$ photon divided by the invariant mass of the diphoton pair</td>
</tr>
<tr>
<td>$p_T^{2}/m_{\gamma\gamma}$</td>
<td>transverse momentum of the second-highest $p_T$ photon divided by the invariant mass of the diphoton pair</td>
</tr>
<tr>
<td>$(p_T^{1} + p_T^{2})/m_{\gamma\gamma}$</td>
<td>sum of the transverse momentum of the two photons divided by their invariant mass</td>
</tr>
<tr>
<td>$(p_T^{1} - p_T^{2})/m_{\gamma\gamma}$</td>
<td>difference of the transverse momentum of the two photons divided by their invariant mass</td>
</tr>
<tr>
<td>$p_{T\gamma}/m_{\gamma\gamma}$</td>
<td>transverse momentum of the diphoton system divided by its invariant mass</td>
</tr>
<tr>
<td>$\Delta \eta$</td>
<td>difference in $\eta = -\log \tan \left( \frac{\theta}{2} \right)$, where $\theta$ is the angle with the beam axis</td>
</tr>
<tr>
<td>$\Delta R$</td>
<td>sum in quadrature of the separation of and $\phi$, the azimuthal angle of the two photons ($\sqrt{\Delta \eta^2 + \Delta \phi^2}$)</td>
</tr>
<tr>
<td>$</td>
<td>\eta^{\gamma\gamma}</td>
</tr>
</tbody>
</table>
Weak Classifier Function

Define $v_{\text{shift}}$

- Based on $70^{\text{th}}$ and $30^{\text{th}}$ percentile of the signal distribution ($s_{70}, s_{30}$)
- If the percentile of background at $s_{70}$ is less than 70%, then translate to $s_{70}$ and invert the variable
- Else, check the percentile of background at $s_{30}$, and if more than 30%, then translate to $s_{30}$.
- Else, the two distributions are “too overlapping” and we discard the variable.

Define $h$

- $v_{+1}$ and $v_{-1}$ are the $10^{\text{th}}$ and $90^{\text{th}}$ percentile of $v_{\text{shift}}$

\[
h(v) = \begin{cases} 
  +1 & \text{if } v_{+1} < v_{\text{shift}}(v) \\
  \frac{v_{\text{shift}}(v)}{v_{+1}} & \text{if } 0 < v_{\text{shift}}(v) \leq v_{+1} \\
  \frac{v_{\text{shift}}(v)}{|v_{-1}|} & \text{if } v_{-1} < v_{\text{shift}}(v) \leq 0 \\
  -1 & \text{if } v_{\text{shift}}(v) < v_{-1}
\end{cases}
\]

Applied to all variables and their product (inverse if flipped)
## Weak Classifiers Numbering

<p>| | | | | | | | | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
<td>5</td>
<td>6</td>
<td>7</td>
<td>8</td>
<td>9</td>
<td></td>
</tr>
<tr>
<td>$p_T^1$</td>
<td>$p_T^2$</td>
<td>$\Delta R$</td>
<td>$p_{\gamma\gamma}^T$</td>
<td>$p_T^1 + p_T^2$</td>
<td>$p_T^1 - p_T^2$</td>
<td>$\Delta \eta$</td>
<td>$\eta_{\gamma\gamma}$</td>
<td>$(p_T^1 + p_T^2)\eta_{\gamma\gamma}$</td>
<td></td>
</tr>
<tr>
<td>10</td>
<td>11</td>
<td>12</td>
<td>13</td>
<td>14</td>
<td>15</td>
<td>16</td>
<td>17</td>
<td>18</td>
<td></td>
</tr>
<tr>
<td>$\frac{p_T^2}{p_T^1 - p_T^2}$</td>
<td>$\frac{p_T^2}{\Delta \eta}$</td>
<td>$p_T^2 \eta_{\gamma\gamma}$</td>
<td>$\frac{1}{\Delta R p_{\gamma\gamma}^T}$</td>
<td>$p_T^1 + p_T^2$</td>
<td>$\frac{1}{\Delta R(p_T^1 - p_T^2)}$</td>
<td>$\frac{1}{\Delta R \Delta \eta}$</td>
<td>$\frac{\eta_{\gamma\gamma}}{\Delta \eta}$</td>
<td>$\frac{1}{(p_T^1 - p_T^2)\Delta \eta}$</td>
<td></td>
</tr>
<tr>
<td>19</td>
<td>20</td>
<td>21</td>
<td>22</td>
<td>23</td>
<td>24</td>
<td>25</td>
<td>26</td>
<td>27</td>
<td></td>
</tr>
<tr>
<td>$p_T^1 p_T^2$</td>
<td>$\frac{p_T^1}{\Delta \eta}$</td>
<td>$p_T^1 (p_T^1 + p_T^2)$</td>
<td>$\frac{p_T^1}{p_T^1 - p_T^2}$</td>
<td>$\frac{1}{\Delta \eta}$</td>
<td>$\frac{p_T^1}{\eta_{\gamma\gamma}}$</td>
<td>$\frac{p_T^2}{\Delta \eta}$</td>
<td>$\frac{\eta_{\gamma\gamma}}{p_T^1 - p_T^2}$</td>
<td></td>
<td></td>
</tr>
<tr>
<td>28</td>
<td>29</td>
<td>30</td>
<td>31</td>
<td>32</td>
<td>33</td>
<td>34</td>
<td>35</td>
<td>36</td>
<td></td>
</tr>
<tr>
<td>$\eta_{\gamma\gamma}^T$</td>
<td>$\frac{p_T^2 (p_T^1 + p_T^2)}{p_T}$</td>
<td>$\frac{p_T^2 + p_T^2}{p_T \eta_{\gamma\gamma}^T}$</td>
<td>$\frac{1}{p_T \Delta \eta}$</td>
<td>$\frac{1}{p_T^1 - p_T^2}$</td>
<td>$p_T^1 + p_T^2$</td>
<td>$\frac{1}{\Delta \eta}$</td>
<td>$\eta_{\gamma\gamma} \Delta \eta$</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Experiments
Outline

➔ Train classical classifiers as a baseline measurement of performance.

➔ Evaluate the exact solution of the problem using simulating annealing of the Ising model.

➔ Scan for $\lambda$, penalty on number of weak classifiers.

➔ Classification performance depending on the size of the training set.

➔ Scan on the fraction of exited states included in the classifier.
Baseline Classifiers
Classical Baseline

➔ XGBoost (XGB)
  • Extremely efficient library for training decision trees
  • http://xgboost.readthedocs.io
  • Discovered during the higgs-ml challenge
    https://www.kaggle.com/c/higgs-boson
  • Moderately optimize the hyper-parameters

➔ Deep Neural Network (DNN)
  • Simple fully connected model 2 layers 1000 nodes
  • https://keras.io/
    http://deeplearning.net/software/theano/
  • Moderately optimize the hyper-parameters
Simulated Annealing
Ising Model Heuristic Solution

- Monte-Carlo based method to find ground state of energy functions
- Random walk across phase space
  - accepting descent
  - accepting ascent with probability $e^{-\frac{\Delta E}{kT}}$
- Decrease $T$ with time

Applied to the QUBO problem, and finds the **ground state** reasonably well. SA in the legends.
Variable Importance
Weak Classifier Penalty

\[
\delta(\vec{w}) \propto \sum_{i,j} C_{ij} w_i w_j + \sum_i (\lambda - 2C_{iy}) w_i
\]

Penalize for using many weak classifiers
Surviving Weak Classifiers

<table>
<thead>
<tr>
<th>λ</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>9</th>
<th>10</th>
<th>11</th>
<th>12</th>
<th>13</th>
<th>14</th>
<th>15</th>
<th>16</th>
<th>17</th>
<th>18</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>0.01</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>0.02</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>0.05</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>0.1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>0.2</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>0.4</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>0.8</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>λ</th>
<th>19</th>
<th>20</th>
<th>21</th>
<th>22</th>
<th>23</th>
<th>24</th>
<th>25</th>
<th>26</th>
<th>27</th>
<th>28</th>
<th>29</th>
<th>30</th>
<th>31</th>
<th>32</th>
<th>33</th>
<th>34</th>
<th>35</th>
<th>36</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>20</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>19</td>
<td>0</td>
<td>20</td>
<td>0</td>
<td>3</td>
<td>0</td>
<td>20</td>
<td>19</td>
<td>7</td>
<td>0</td>
<td>15</td>
<td>0</td>
<td>19</td>
<td>0</td>
</tr>
<tr>
<td>0.01</td>
<td>20</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>19</td>
<td>0</td>
<td>20</td>
<td>0</td>
<td>2</td>
<td>0</td>
<td>20</td>
<td>19</td>
<td>6</td>
<td>0</td>
<td>15</td>
<td>0</td>
<td>19</td>
<td>0</td>
</tr>
<tr>
<td>0.02</td>
<td>20</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>19</td>
<td>0</td>
<td>20</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>20</td>
<td>19</td>
<td>4</td>
<td>0</td>
<td>15</td>
<td>0</td>
<td>19</td>
<td>0</td>
</tr>
<tr>
<td>0.05</td>
<td>20</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>19</td>
<td>0</td>
<td>20</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>20</td>
<td>16</td>
<td>1</td>
<td>0</td>
<td>11</td>
<td>0</td>
<td>19</td>
<td>0</td>
</tr>
<tr>
<td>0.1</td>
<td>20</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>20</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>20</td>
<td>1</td>
<td>0</td>
<td>5</td>
<td>0</td>
<td>16</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>0.2</td>
<td>18</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>20</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>20</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>0.4</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>20</td>
<td>3</td>
<td></td>
</tr>
<tr>
<td>0.8</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>19</td>
<td>0</td>
</tr>
</tbody>
</table>

Table: Number of times the weak classifier of a given variable is used in the ground state solution, as a function of the penalty.

Three major variables (2, 13, 28): \( p_T^2, (\Delta R p_T^{\gamma\gamma})^{-1} \), and \( \frac{p_T^2}{p_T^{\gamma\gamma}} \)

Relates to the creation of a heavy particle (Higgs) with less transverse energy than typical QCD in the same mass range.
Classification Performance
Sample Size of 100 Events

Two components in the error band:
2. Spread over the folds.
Sample Size of 20k Events

Two components in the error band. Both negligible:
1. Stat error of the test set.
2. Spread over the folds.
Evolution With Training Size

Two components in the error band.
1. Stat error of the test set.
2. Spread over the folds.
Hybrid Classifier
In presence of Excited States

- Excited states bring new mask of weak classifier to build discriminators
- Building a discriminator including excited states
  - Take x% of the levels above the ground state
  - For each evaluate the average ROC on the training folds
  - At each signal efficiency, pick the energy state that has the most rejecting discriminator
Comparison DW/SA

Ground state only
$E < E_0 + 2\%$ excited states
$E < E_0 + 5\%$ excited states

Integral of ROC difference: DW-SA

Training size

100  1000  5000  10000  15000  20000
Chimera Spin Chains

- Quantum Annealing gets better at finding the ground truth ground state with strong chain strength.
- SA does not include spin chains.
- Feedback to machine developers
Summary

• First application of D-Wave quantum annealing capability to High Energy Physics use case. Raises interesting questions.

• Scope of the Quantum Annealing on the D-Wave computing device is solving the Ising model. Limited but powerful.

→ Potential applicability of Quantum Annealing to other problems.
This project is supported in part by the United States Department of Energy, Office of High Energy Physics Research Technology Computational HEP and Fermi Research Alliance, LLC under Contract No. DE-AC02-07CH11359. The project is also supported in part under ARO grant number W911NF-12-1-0523 and NSF grant number INSPIRE-1551064. The work is supported in part by the AT&T Foundry Innovation Centers through INQNET, a program for accelerating quantum technologies. We wish to thank the Advanced Scientific Computing Research program of the DOE for the opportunity to first present and discuss this work at the ASCR workshop on Quantum Computing for Science (2015). We acknowledge the funding agencies and all the scientists and staff at CERN and internationally whose hard work resulted in the momentous H(125) discovery in 2012.

With special thanks to Joshua Job.