# Simulated Quantum Annealing of Feed-Forward Neural Network Synaptic Weights

Justin R. Fletcher, *Student Member, IEEE,* Michael J. Mendenhall, *Member, IEEE*

**Abstract**—In this paper we present a methodology for feed-forward neural network synaptic weight selection which employs a quantum-inspired variant of simulated annealing.
The algorithm presented in this paper does not require any explicit temperature schedule tuning, and runs continuously.

**Index Terms**—simulated quantum annealing, neural networks. stochastic gradient descent.

## 1 Introduction

.. the most common method of weight selection is back-propagation. Back propagation is a [direct] gradient descent method, which traverses the error surface along its maximum gradient path. The algorithm converges once a minimum located, because there is no longer a negative gradient for the algorithm to follow. While this method can be very successful, it often suffers from premature convergence. That is, the algorithm finds a local, rather than global, error surface minima.

... In this paper, we present an explanation for the utility of FSA in terms of quantum annealing. ... We show that quantum annealin is effective as a weight selection mechanism for feed-forward neural networks.

## 2 Literature Review

Simulated annealing was first applied to neural network weight selection

Though simulated annealing is effective at finding synaptic weight configurations which perform comparatively well [], the method has several significant shortcomings. Neural networks trained using simulated annealing are often converge slowly relative to other conventional methods of weight selection such as back-propagation []. Simulated annealing also requires the a priori specification of a temperature schedule, which must cool sufficiently slowly to ensure convergence to a global, rather than local, minimum. The amount of cooling time required is highly-problem dependent. These limitations motivated the study of alternative methods, which led to the work presented in this paper.

Simulated annealing is a stochastic optimization algorithm which can be used to find the global minimum of a cost function mapped from the configurations of a combinatorial optimization problem. The concept of simulated annealing was introduced in by Kirkpatrick et. al. in [1] as an application of the methods of statistical mechanics to the problem of discrete combinatorial optimization. Specifically, simulated annealing is an extension of the Metropolis-Hastings [2] algorithm which can be used to estimate the ground energy state of a many-body systems at thermal equilibrium. Kirkpatrick et. al. applied the Metropolis-Hastings algorithm sequentially, with decreasing temperature values in order to approximate a solid slowly cooling to low temperatures. Later work by Goffe [3], Corana et al. [4], and Lecchini-Visintini et. al. [5] extended simulated annealing to the continuous domain.

When considering only the influence of classical thermal fluctuations in particle energy levels, the probability of a particle traversing a barrier of height $\Delta V$ at a temperature $T$ is on the order of:

$$\mathcal{P}_t = e^{-\frac{\Delta V}{T}}. \tag{1}$$

This probability forms the basis for the Metropolis acceptance criterion which is given by:

$$\mathcal{P}_t = e^{-\frac{\Delta V}{T}}. \tag{2}$$

... In [] Szu and Hartley introduced the method of fast simulated annealing (FSA), which incorporates occasional, long jumps through the configurtation space. This provision allows for the possiblity of escaping local minima, and reduces the total computational effort required to reach a global minimum. Later work by Tsallis and Stariolo [], generalized both CSA and FSA into a single framework: generalized simulated annealing (GSA).

### 2.1 Paper Organization

Section 3 contains a discussion of the quantum-inspired simulated annealing algorithm that will be used to select

## 3 Problem Formulation and Overview

In this paper we present a methodology for feed-forward neural network weight selection which employs

- *J.F. and M.M. are with the Department of Electrical and Computer Engineering, Air Force Institute of Technology.*
- *The views expressed in this article are those of the authors and do not reflect the official policy or position of the United States Air Force, Department of Defense, or the U.S. Government.*

a quantum-annealing-inspired variant of simulated annealing. In the following sections CSA and simulated simulated annealing (QSA) are introduced and compared.

As discussed in Section 2, simulated annealing converges to a global, rather than local, minimum only if it is given sufficient time to converge. The amound of time required is problem dependent, and is often unacceptably long. One way to solve this problem is to create a mechanism by which the system can tunnel out of local minima to nearby, lower-cost configurations, thus reducing the amount of time needed to garuntee convergence to the global minimum. This is analogous to the quantum mechanical phenomenon of tunneling. In a quantum tunneling event a particle with energy $E$ incident upon a potential energy barrier of height $\Delta V > E$ has a non-zero probability of being found in, or past, the barrier. Classically, this behavior is forbidden. The probability of tunneling, $\mathcal{P}_t$, through a step barrier of height $\Delta V$ is described by:

$$\mathcal{P}_t = e^{-\frac{w\sqrt{\Delta V}}{\Gamma}} \tag{3}$$

where $\Gamma$ is the tunneling field strength [6]. Figure **??** depicts a one-dimensional example of quantum tunneling.

It is instructive to contrast Eq. 1 and Eq. 3. Both describe the same value, but the importance of the width and height of the traversed barrier in the two equations is considerably different. For systems in which quantum tunneling is possible, the probability of penetrating a barrier of height $\Delta V$ is increased by a factor of approximately $e^{\Delta V}$, for large values of $\Delta V$. This relationship is depicted graphically in Fig. **??** which shows the probability of barrier traversal for a system which allows quantum fluctuations, divided by the same probability for a system which only considers thermal fluctuations. Therefore, physical models which consider quantum effects are more likely predict penetration of tall, thin energy barriers than those which only include classical thermal effects.

### 3.1 Classical and Quantum Simulated Annealing

The formulation of QSA described in this paper arises from the construction of a neighborhood function which approximates the transition probabilities of a quantum system traversing a potential energy surface (PES). This concept is nearly mathematically identical to the fast simulated annealing model described in []. However, the QSA algorithm presented in this paper was arrived at independently, though an attempt to incorporate quantum mechanical phenomena into neural network weight selection.

### 3.2 Annealing Synaptic Weights

In order to apply the techniques of QSA and CSA, we must first formulate the problem of feed-forward neural network synaptic weight selection as a combinatorial optimization problem. Each synaptic weight in a feed-forward neural network may be encoded as a real-valued element in a 3-dimensional relation matrix, denoted as $W_{ijk}$. In this encoding scheme, for a given layer, $k$, of the matrix the row and column indexes indicate the pre-synaptic and post-synaptic neurons, respectively. The absence of a synaptic connection is indicated by a value of $0$ in the matrix element cooresponding to that synaptic connection. A nonexistent synapse can be caused by the absence of either the presynaptic or post-synaptic neuron, or by the absence of a connection between the neurons. This weight encoding scheme is depicted graphically in Fig. **??**. The weight matrix, $\boldsymbol{\mathcal{W}}$, therefore encodes a configuration in the solution space of the problem, which in turn corresponds to some cost value $\mathcal{C}(\boldsymbol{\mathcal{W}})$. We may now attempt to minimize $\mathcal{C}(\boldsymbol{\mathcal{W}})$ by perturbing $\boldsymbol{\mathcal{W}}$. With this framework in place, we can apply the techniques of CSA and QSA to the problem of weight selection.

### 3.3 Neighborhood Functions

All variations of simulated annealing require the specification of a neighborhood function, which determines the way in which new states may be generated from the current state. Let us define a generic neighborhood function, $\mathcal{N}$, which operates on $\boldsymbol{\mathcal{W}}$, thereby changing its value. In the following section we detail several possible realizations of $\mathcal{N}$, some of which constitute an approximation to system performing quantum tunneling under the influence of an annealable artificial temperature.

#### *3.3.1 Classical Neighborhood Functions*

We define $\mathcal{N}_c$ to be the classical neighborhood function, which is defined as

$$\mathcal{N}_c(\boldsymbol{\mathcal{W}}) = \boldsymbol{\mathcal{W}} + \alpha\boldsymbol{\mathcal{A}} \tag{4}$$

where $\alpha$ is the learning rate and $\boldsymbol{\mathcal{A}}$ is a matrix with dimensionality equal to that of $\boldsymbol{\mathcal{W}}$. Each element of $\boldsymbol{\mathcal{A}}$ is generated from a distibution over the range $[-1, 1]$. $\boldsymbol{\mathcal{A}}$ is restricted such that for each element of $\boldsymbol{\mathcal{W}}$ that is zero the cooresponding element in $\boldsymbol{\mathcal{A}}$ is zero, and that $\boldsymbol{\mathcal{A}}$ must be normalized such that the magnitude of the matrix elements sum to $1$. These restrictions ensure two useful properties of the classical neighborhood function, that:

1) $\mathcal{N}_c(\boldsymbol{\mathcal{W}})$ will produce a weight matrix, $\boldsymbol{\mathcal{W}}'$, which will have an $L^2$ distance of exactly $\alpha$ from the original matrix, $\boldsymbol{\mathcal{W}}$, in the weight space.
2) This traversal distance will be distributed anisotropically over the weights, with the anisotropicity determined by the distribution used to genererate $\boldsymbol{\mathcal{A}}$. (See Sec. 3.3.3 for more information.)

These properties are useful in that they allow for strict control over the systems traversal of the cost surface.

#### 3.3.2 Quantum Neighborhood Function

The classical neighborhood function present in Sec. 3.3.1 is one of many which could be employed in CSA. In order to incorporate quantum tunneling into this model, we must have some mechanism which allows the neighborhood function to generate neighbor states which are "across" a cost surface barrier. Since it is impossible to know the cost funciton value of a configuration which has not yet been evaluated, we must construct a neighborhood function which is able to jump to these configurations. This mechanism is often called a trial jump. We define our quantum neightborhood function to be

$$\mathcal{N}_q(\boldsymbol{\mathcal{W}}) = \boldsymbol{\mathcal{W}} + \alpha G_\Gamma(X)\boldsymbol{\mathcal{A}} \tag{5}$$

where $\alpha$ and $\boldsymbol{\mathcal{A}}$ are defined as they are in Eq. 4, $X$ is a uniform random variable over the range $[0, 1]$, and $G_\Gamma(X)$ is a parameterized generation function used to produce a value from the visiting distribution. As in [7], the visiting distribution is defined as the probability distribution function of the trial jump distance.

The visiting distribution used in this paper for quantum neighborhood functions is the exponential distibution, for which the generation function is given by

$$G_\Gamma(X) = \frac{-\ln(X)}{\gamma} \tag{6}$$

where $\gamma$, the scale parameter for the distribution, is defined as $(1-\Gamma)$. The exponential distribution was selected for this effort so that as close an analogy as possible with quantum mechanical reality could be maintained. As indicated by Eq. 3 the probability of making a transition through an potential energy barrier decreases exponentially with the width of the barrier; this is analogous to the trial jump distance in the quantum neighborhood function.

The introduced parameter $\Gamma$ denotes the strength of the tunneling field as in [6], and thus controls the likelihood of a long trial jump. $\Gamma$ is supported on $[0, 1)$. A large value of $\Gamma$ corresponds to frequent, long range quantum trial jumps. It is the stochastic control parameter $\Gamma$ value that connects the quantum neightborhood function to qauntum annealing. Much like FSA [**?**], this provides mostly local search with occasoinal global-scale searches. Unlike FSA and GSA, QSA includes an annealable stochastic control parameter specfically to control the behavior of the frequency of global serach instances.

... Consider the hypothetical cost surface depicted in Fig. **??**.

#### 3.3.3 Anisotropicity

In the course of developing this framework, it was noted that it is sometimes advantageous to move along an error surface in exactly one dimension, as show in Fig. **??** This corresponds to modifying a single synaptic weight a single epoch. Though it is possible that this could occur by chance, the likelihood of a single-weight modification is inversely proportional to the number of weights in the network. In physical science a force acting preferentially in a subset of the total dimensionality of the degrees of freedom in which it could act is referred to as an anisotropy. This concept has been extended to the weight selection scheme presented in this paper in the form of a value, $a$, specifying the anisotropicity of a weight matrix perturbation. This value serves as a nonlinear transform on the perturbation matrix, thus concentrating the perturbation in a subset of the available degrees of freedom. Given as

By inspection, when $a$ is $0$, the perturbation will be unchanged and will be distributed evenly across all degrees of freedom, on average. Conversely, $a$ is $1$, all the perturbation will occur in a single dimension, thus changing only one synaptic weight.

### 3.4 Cost Functions

### 3.5 Annealing Schedules

## 4 Implementation Details

## 5 Experiments and Results

## References


[1] S. Kirkpatrick, C. D. Gelatt, and M. P. Vecchi, "Optimization by simulated annealing," *SCIENCE*, vol. 220, no. 4598, pp. 671–680, 1983.

[2] N. Metropolis, A. W. Rosenbluth, M. N. Rosenbluth, A. H. Teller, and E. Teller, "Equation of State Calculations by Fast Computing Machines," vol. 21, pp. 1087–1092, June 1953.

[3] W. L. Goffe, G. D. Ferrier, and J. Rogers, "Global optimization of statistical functions with simulated annealing," *Journal of Econometrics*, vol. 60, pp. 65–99, 1994.

[4] A. Corana, M. Marchesi, C. Martini, and S. Ridella, "Minimizing multimodal functions of continuous variables with the “simulated annealing” algorithmcorrigenda for this article is available here," *ACM Trans. Math. Softw.*, vol. 13, no. 3, pp. 262–280, Sept. 1987. [Online]. Available: http://doi.acm.org/10.1145/29380.29864

[5] A. Lecchini-Visintini, J. Lygeros, and J. Maciejowski, "Simulated Annealing: Rigorous finite-time guarantees for optimization on continuous domains," *ArXiv e-prints*, Sept. 2007.

[6] S. Mukherjee and B. K. Chakrabarti, "Multivariable optimization: Quantum annealing and computation," *European Physical Journal Special Topics*, vol. 224, p. 17, Feb. 2015.

[7] C. Tsallis and D. A. Stariolo, "Generalized simulated annealing," *Physica A: Statistical and Theoretical Physics*, vol. 233, no. 1-2, pp. 395–406, Nov. 1996. [Online]. Available: http://www.sciencedirect.com/science/article/B6TVG-3YK5TC8-2H/1/d040f1408073d6a09dc185f38673e3dd


PLACE PHOTO HERE

**Justin R. Fletcher** received the B.S degree in Computer Engineering from Embry Riddle Aeronautical University in 2012 and is currently pursuing the M.S. degree in Computer Science from the Air Force Institute of Technology (AFIT). His research interests include neural networks, quantum mechanics, and stuff. [do this later]

PLACE PHOTO HERE

**Michael J. Mendenhall** received the B.S. degree in Computer Engineering from Oregon State University, Corvallis, OR in 1996, the M.S. degree in Computer Engineering from the Air Force Institute of Technology (AFIT), Wright-Patterson AFB, OH, in 2001, and the Ph.D. degree in Electrical Engineering from Rice University, Houston, TX, in 2006. Currently, he is an Assistant Professor of Electrical Engineering at AFIT. He recieved the Dr. Leslie M. Norton teaching award by the AFIT student association and an honorable mention for the John L. McLucas basic research award at the Air Force level, both in 2010. His research interests are in hyperspectral signal/image processing, hyperspectral signature modeling, and computational intelligence.