2006 International Joint Conference on Neural Networks
Sheraton Vancouver Wall Centre Hotel, Vancouver, BC, Canada
July 16-21, 2006

Bayesian Training of Neural Networks Using Genetic Programming

Ts ili zi Marwala *Member, IEEE*

***Abstract*—Bayesian neural networks trained using Markov chain Monte Carlo (MCMC) and genetic programming in binary space within Metropolis framework is proposed. It is tested and compared to classical MCMC method and is observed to give better results than classical approach.**

I INTRODUCTION

THE use of Bayesian framework to train neural networks as been t e subject of researc uring t e previous eca e Some of t e tec niques t at ave been applie t us far to train neural networks using Bayesian framework inclu e Markov c ain Monte Carlo MCMC) met o [ ] an t e ybri Monte Carlo met o [ ] Bot t ese met o s ave been applie wit in t e framework of Metropolis et al. algorit m [ ] Markov c ain Monte Carlo met o as been applie to improve t e abilities of mat ematical mo els to pre ict t e ynamics an reliability of structures [4] MCMC was use for Bayesian curve fitting an applie to signal segmentation [5] to estimate regularization parameters for satellite image restoration [ ] an for inference of stoc astic volatility mo els of t e S P in ex [7]

All t ese applications t at ave been escribe above ave one aspect in common an t is is t at t ey ave been applie wit out paying particular attention to t e issue of ac ieving a global optimum posterior istribution function Ken all an Montana [8] ave note t at insi e every Markov c ain wit measurable transition ensity t ere is a iscrete state space Markov c ain struggling to escape from some local optimum istribution T is in essence in icates t at t e issue of global posterior istribution must not be taken for grante Several tec niques ave been implemente to ac ieve global optimum istributions suc as simulate annealing [9] an evolutionary computing to i entify global solutions in optimization problems [ ] T is paper t erefore proposes t e use of genetic programming to sample a posterior probability istribution of t e neural network weig ts T e proce ure propose operates in binary space an conventional evolutionary concepts of mutation crossover an repro uction are applie T is proce ure is t en compare to t e classical Markov c ain Monte Carlo met o t at samples states in floating point space

Manuscript receive December 5
T Marwala is a professor of Electrical Engineering at t e sc ool of Electrical an Information Engineering University of t e Witwatersran Jo annesburg Sout Africa p one: 7 7 7 7 7 fax: 7 4 9 9 e mail: t marwala@ee wits ac za)

II NEURAL NETWORKS

T is section escribes multi layer perceptron MLP) neural networks w ic are parameterise grap s t at make probabilistic assumptions about ata T e network arc itecture implemente in t is paper contains i en units an output units an as one i en layer T e relations ip between output y an input x may be written as follows [ ]:

$$y_k = f_{outer}\left(\sum_{j=1}^{M} w_{kj}^{(2)} f_{inner}\left(\sum_{i=1}^{d} w_{ji}^{(1)} x_i + w_{j0}^{(1)}\right) + w_{k0}^{(2)}\right) \qquad)$$

Here $w_{ji}^{(1)}$ an $w_{ji}^{(2)}$ in icate weig ts in t e first an secon layers respectively going from input i to i en unit j w ile $w_{j0}^{(1)}$ in icates t e bias for t e i en unit j Here M is t e number of i en units d is t e number of input units an k is t e in ex for t e output units In t is paper t e function $f_{outer}(\bullet)$ is linear w ile $f_{inner}(\bullet)$ is a yperbolic tangent function [ ] T e Bayesian met o i entifies t e istribution of weig ts in) t at look probable in t e lig t of ata

III BAYESIAN FRAMEWORK

In t is section a met o of i entifying t e network weig ts escribe in) is outline T e problem of i entifying t e network weig ts is pose in Bayesian form as follows [ ]:

$$P \; w \mid D) = \frac{P \; D \mid w) P \; w)}{P \; D)} \qquad)$$

w ere *P(w)* is t e probability istribution function of t e weig t space in t e absence of any ata also known as t e prior istribution function an $D \equiv (y_1,\ldots,y_N)$ is a matrix containing t e output ata T e quantity $P(w|D)$ is t e posterior probability istribution function after t e ata ave been seen $P(D|w)$ is t e likeli oo istribution function an $P(D)$ is t e evi ence an its function is to normalize t e posterior probability istribution function Equation) may be expan e to give [ ]:

$$P \; w \mid D) = \frac{}{Z_s}\exp\left(-\beta\sum_{n}^{N}\sum_{k}^{K}\{t_{nk} - y_{nk}\} \; - \frac{\alpha}{}\sum_{j}^{W} w_j\right) \qquad)$$

w ere

$$Z_S(\alpha,\beta)=\int exp\left(-\beta\sum_{n}^{N}\sum_{k}^{K}\{t_{nk}-y_{nk}\}^2-\frac{\alpha}{2}\sum_{j}^{W}w_j^2\right)dw \quad 4)$$
$$=\left(\frac{2\pi}{\beta}\right)^{N/2}+\left(\frac{2\pi}{\alpha}\right)^{W/2}$$

In) t e first term in t e exponent is t e likeli oo function an t e secon term is t e prior information n is t e in ex for t e training pattern β is t e ata contribution to t e error k is t e in ex for t e output units an α is t e coefficient of t e prior information T e secon term in) is t e regularization parameter Training t e network using Bayesian approac automatically penalizes ig ly complex mo els an also gives a probability istribution of t e output of t e networks

IV MCMC VIA METROPOLIS ALGORITHM

T e application of Bayesian approac to neural networks results in t e probability istribution functions of t e network outputs From t ese istribution functions t e average pre iction of t e neural network an t e variance of t at pre iction can be calculate T e probability istributions of t ese network weig ts are mat ematically escribe by) From) an by following t e rules of probability t eory t e istribution of t e output parameter *y,* is written as [ ]:

$$p\ \ y \mid x\ \ D)=\int p\ \ y \mid x\ \ w)p\ \ w \mid D)dw \qquad 5)$$

Equation 5) epen s on) an is ifficult to solve analytically ue to relatively ig imension of weig t space T us t e integral in 5) may be approximate as follows:

$$\tilde{y} \cong \frac{}{L}\sum_{i=I}^{R+L-}F\ \ w_i) \qquad)$$

Here F is t e mat ematical mo el t at gives t e output given t e input $\tilde{y}$ is t e average pre iction of t e Bayesian neural network R is t e number of initial states t at are iscar e in t e ope of reac ing a stationary posterior istribution function escribe in) an L is t e number of retaine states In t is paper MCMC met o is implemente by sampling a stoc astic process consisting of ran om variables $\{w_1,w_2,\ldots,w_n\}$ t roug intro ucing ran om c anges to weig t vector $\{w\}$ an eit er accepting or rejecting t e state accor ing to Metropolis et al algorit m given t e ifferences in posterior probabilities between two states t at are in transition [ ] T is algorit m ensures t at states wit ig probability form t e majority of t e Markov c ain Tra itionally t e MCMC was con ucte in floating point space an t is paper intro uces genetic sampling of Bayesian networks w ic is t e subject of t e next section

V MCMC: GENETIC PROGRAMMING AND METROPOLIS ALGORITHM

Genetic programming takes features from natural evolution an uses t ese to computationally solve practical problems Genetic algorit ms are examples of genetic programming an a proce ure t at is inspire by t ese is intro uce in t is section In t is paper some of t e features of genetic computing are applie to sample t e posterior istribution function in)

Genetic algorit ms were inspire by Darwin s t eory of natural evolution In natural evolution members of t e population compete wit eac ot er to survive an repro uce Evolutionary successful in ivi uals repro uce w ile weaker members ie As a result t e genes t at are successful are likely going to sprea wit in t e population T is natural optimisation met o as been successfully use to optimise complex problems [ ] [4] T is proce ure uses a population of binary string c romosomes an eac of t ese strings is t e iscretise representation of a point in t e searc space an t erefore as a fitness function t at is given by t e objective function On generating a new population t ree operators are performe :) crossover) mutation) an repro uction an t ese operators are a opte in genetic MCMC sampling T e crossover operator mixes genetic information in t e population by cutting pairs of c romosomes at ran om points along t eir lengt an exc anging over t e cut sections T is as a potential of joining successful operators toget er Crossover occurs wit a certain probability In many natural systems t e probability of crossover occurring is ig er t an t e probability of mutation occurring Simple crossover tec nique is use in t is paper [4] For simple crossover one crossover point is selecte binary string from beginning of c romosome to t e crossover point is copie from one parent an t e rest is copie from t e secon parent For example w en **11001** un ergoes simple crossover wit **111** it becomes **11001111.**

T e mutation operator picks a binary igit of t e c romosomes at ran om an inverts it T is as a potential of intro ucing to t e population new information Mutation occurs wit a certain probability In many natural systems t e probability of mutation is low i e less t an %) In t is paper binary mutation is use [4] W en binary mutation is use a number written in binary form is c osen an its value is inverte For an example: **1** may become **0**

Repro uction takes successful c romosomes an repro uces t em in accor ance to t eir fitness functions In t is paper Metropolis [ ] criteria is use as a repro uction met o By so oing t e least fit members are t erefore gra ually riven out of t e population of states t at form a Markov c ain T e sc ematic illustration of t e MCMC met o traine using genetic programming is s own in Fig In t is figure an initial sample weig t vector $\{w\}_n$ is generate

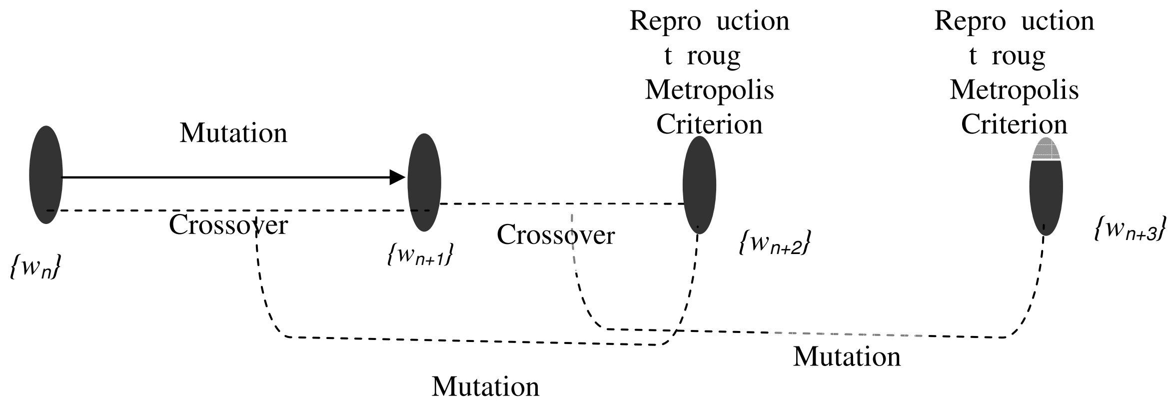

Fig A sc ematic illustration of genetic sampling for t e MCMC implementation

T en t e sample is converte into binary form using Gray met o [ ] T e sample is t en mutate to form a new sample vector $\{w\}_{n+1}$ T e new weig t vector $\{w\}_{n+1}$ un ergoes crossover wit its pre ecessor $\{w\}_n$ an mutates again to form a new network weig t vector $\{w\}_{n+2}$ T e weig t vector $\{w\}_{n+2}$ is converte into floating point an t en its probability is calculate T is network weig t vector is eit er accepte or rejecte using Metropolis criterion [ ] T ereafter states $\{w\}_{n+2}$ an $\{w\}_{n+1}$ in binary form un ergo crossover an are mutate to form $\{w\}_{n+3}$ T e genetic MCMC propose in t is section is ifferent from t e tra itional GA in t e following nature: a) T e genetic MCMC oes not generate a new population of genes at any given iteration i e generation in t e GA framework) as is t e case in GA but it generates one sample at eac iteration b) T e fitness function uses Metropolis criterion w ile in GA t is is not t e case an c) T e genetic MCMC as a ig er mutation rate t an GA T e genetic MCMC is ifferent from a stan ar MCMC in t e following way: a) T e ran om walk in t e classical MCMC is replace by a proce ure inspire by Darwin s t eory of evolution w ic entails cross over mutation an repro uction an b) It operates in floating point space

VI CASE STUDY : SIMULATED STUDY

In t is case stu y Bayesian neural network t at is traine using genetic MCMC is use for regression problems T e same regression problem is solve using t e classical MCMC w ic generate states in floating point space an accept or reject t e state using Metropolis et al met o T e results of t e two met o s are t en compare T e simulate ata are generate from a noisy sine function wit a stan ar eviation of T is is t e same function t at was use by Nabney [5] Twenty ata points are generate aroun $x=0.25$ Regression analysis is con ucte for t e omain x [ ] T e MLP networks constructe ave one input five i en units an one output units T e optimal number of i en units was obtaine by stu ying t e relations ip between t e number of i en units an t e generalization error T is was con ucte by setting t e number of i en units to fall between an 8 an assessing t e generalization error T e i en layer activation functions are a yperbolic tangent functions w ile t e output activation functions are linear functions

On implementing Bayesian training t e coefficient of t e ata contribution to t e error β is set to w ile t e prior coefficient α is set to T e manner in w ic t ese parameters fit into t e Bayesian framework is escribe by) T e number of retaine states L is w ile t e number of iscar e states R is an t ese values fit into t e Bayesian framework t roug)

For t e genetic part of t e simulation t e rate of mutation is % an t e rate of crossover is 7 % It s oul be note t at t e rate of mutation propose ere is ig er t an t at of stan ar genetic algorit m T e propose Bayesian met o via genetic programming as a ran om component searc an t erefore may be viewe as being equivalent to t e ran om walk t at is execute in t e stan ar Bayesian sampling In ee t e propose proce ure may in principle be equivalent to t e stan ar ran om walk owever it takes into account of t e efficient sampling in binary space w ic as been observe in stan ar genetic algorit m It must be note t at t e rate of mutation c osen ere is lower t an t e rate of crossover w ic is in accor ance to many natural systems W en implementing t e genetic framework t roug genetic algorit m bit binary numbers are use T e boun s of t e magnitu es of t e components of t e weig t vectors are [4 4] T e results obtaine w en Bayesian networks are traine using genetic programming are s own in Fig

T e mean square error MSE) obtaine from t is figure is 7 T e results obtaine w en t e Bayesian networks are traine using classical MCMC are s own in Figure T is gives an MSE of 55 W en Fig is compare to Fig it is observe t at genetic approac to MCMC performs better t an t e met o t at uses MCMC met o t at operates in floating point space because it gives lower average errors T e grap s owing t e errors as a function of samples accepte from state to state is s own in Fig 4

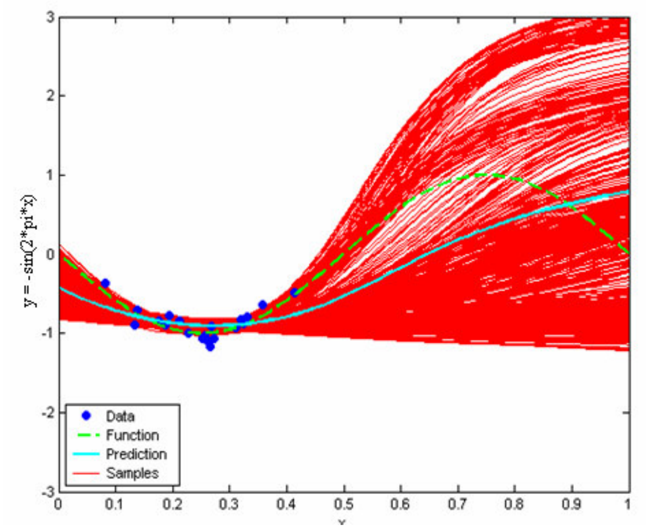

Fig Results obtaine w en Bayesian networks are traine via genetic programming

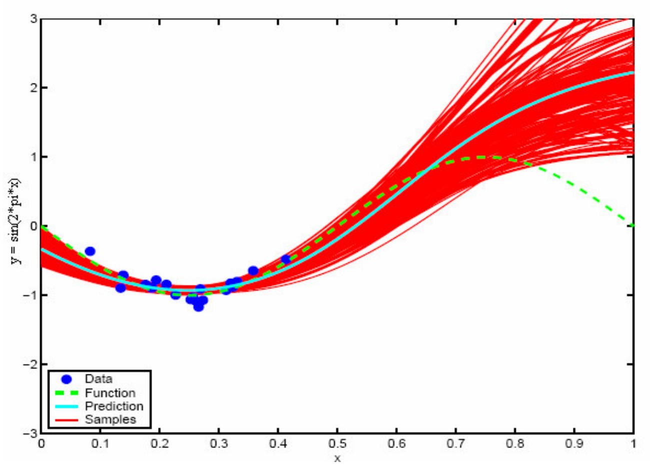

Fig Results obtaine w en Bayesian networks are traine via tra itional MCMC

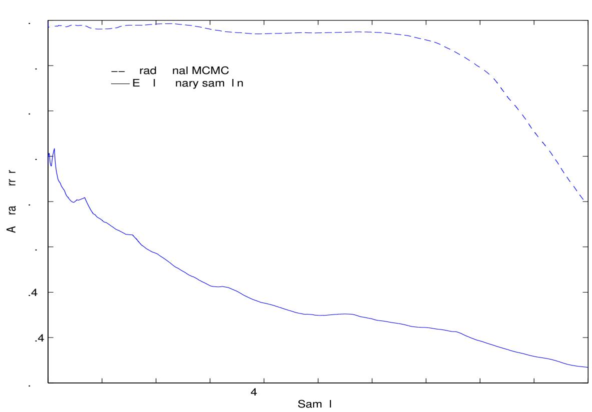

Fig 4 Pre iction errors versus samples

T is grap s ows t at t e rate of convergence to a stationary posterior istribution is faster w en using genetic approac t an w en using t e stan ar MCMC met o T is is because it as been proven t at sampling t roug binary space is more efficient t an sampling t roug floating point space T e reason for t is is t at sampling t roug binary space is able to explore a larger part of t e weig t space t an if t e process is con ucte in floating point space T e acceptance rate of states for MCMC met o t at operates in floating point space is 8 w ereas w en using MCMC met o base on genetic programming is 7 T is in icates t at t e MCMC met o base on genetic programming is able to explore states t at form a Markov c ain better t an t e MCMC met o t at operates in floating point space

VII CASE STUDY : ARTIFICIAL TASTER

Now t at t e simulations ave been con ucte we now apply t e proce ure to a practical problem of a evelopment of an artificial taster Artificial taster as been t e subject of researc for some time Some of t e works on t is subject are a evelopment of a taster base on proton transfer mass spectroscopy to successfully taste mozzarella c eese [ ] a soli state electronic taster for beverage analysis [7] a taste sensor base on lipi coate crystal microbalance to evaluate beer bo y an smoot ness [8] an wine flavour taster t at uses multivariate statistics [9] In t is paper t e met o propose is use to construct an artificial taster an compare to t e classical approac T e artificial taster is basically an infrastructure t at relates t e c aracteristics of beer measure in t e laboratory to taste score measure from a panel of professional tasters T ese c aracteristics capture parameters t at uman beings are sensitive to on tasting beer an t ese are: alco ol level sugar level calle present extract an real extract pH iron acetal e y e imet yl sulp i e et yl acetate iso amyl acetate total ig er alco ols colour an bitterness as inputs to t e artificial taster t at estimates t e average taste score given by a panel of professional tasters To realise t e artificial taster MLP neural network is use T e MLP as inputs 7 i en no es an output T e genetic parameters use in t e previous example are use in t is case T e c aracteristics of beer an t e correspon ing taste score from an average score from a panel of professional tasters are use to construct an artificial taster T ese tasters eac give a taste score t at ranges from for a really ba beer to for a goo beer W en an artificial taster w ic is efine in t is paper as

Table Samples nee e for convergence

Met o	Number of Samples for Convergence	Average Percentage Errors
Classical Approac	5	9 %
Evolutionary	4	7 %

neural networks an measure analytical ata from beer is use to pre ict taste scores t e results obtaine are s own in Table an Fig 5 T e i ea of giving a taste a numerical number is quite ifficult to justify p ilosop ically but psyc op ysics as provi e appropriate measurement tec niques for subjective p enomena suc as taste an t is is t e framework t at is a opte in t is paper

T ese results s ow t at t e genetic approac propose in t is paper converges faster t an t e classical MCMC Secon ly t ese results s ow t at t e genetic programming gives lower average errors t an t e classical approac T is

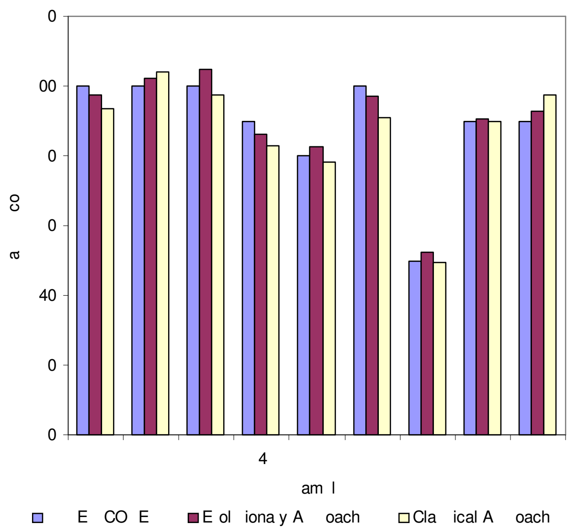

Fig 5 T e taste score for a given taste sample T e top grap is for t e entire 7 test ata w ile t e bottom grap is a close view of t e top grap

is because of t e fact t at genetic programming is able to explore a wi er searc space more efficiently t an t e classical approac

VIII CONCLUSION

T e met o is teste on simulate ata T e results obtaine are compare to t ose obtaine from MCMC met o t at operates in floating point space It is conclu e t at t e MCMC met o base on genetic programming gives better results t an MCMC met o t at operates in floating point space on mo eling simulate ata an artificial taster

REFERENCES

[ ] R E Kass B P Carlin A Gelman an M Neal "Markov Monte Carlo in practice: A roun table iscussion" *American Statistician* vol 5 pp 9 998

[ ] R M Neal "An improve acceptance proce ure for t e ybri Monte Carlo algorit m" *J. Computational Physics* vol pp 94 994

[ ] N Metropolis A W Rosenblut M N Rosenblut A H Teller an E Teller "Equations of state calculations by fast computing mac ines" *Journal of Chemical Physics* vol pp 87 9 95

[4] T Marwala an S Sibisi "Finite element up ating using Bayesian framework an mo al properties" *Journal of Aircraft* vol 4 pp 75 78 5

[5] E Punskaya C An rieu A Doucet an W J Fitzgeral "Bayesian curve fitting using MCMC wit applications to signal segmentation" *IEEE Transactions on Signal Processing* vol 5 pp 747 758

[ ] A Jalobeanu L Blanc Férau an J Zerubia "Hyperparameter estimation for satellite image restoration using a MCMC maximum likeli oo met o " *Pattern Recognition* vol 5 pp 4 5

[7] S C ib F Nar ari an N S ep ar "Markov c ain Monte Carlo met o s for stoc astic volatility mo els" *Journal of Econometrics* vol 8 pp 8

[8] W S Ken all an G Montana "Small sets an Markov transition ensities" *Stochastic Processes and their Applications* vol 99 pp 77 94

[9] R M Neal "Probabilistic inference using Markov c ain Monte Carlo met o s" *University of Toronto Technical Report CRG-TR-93-1* Toronto Cana a 99

[ ] T Marwala "Finite element mo el up ating using wavelet ata an genetic algorit m" *Journal of Aircraft* vol 9 pp 7 9 7

[ ] C M Bis op *Neural Networks for Pattern Recognition* Oxfor University Press Oxfor UK 995

[ ] J Hollan *Adaptation in natural and artificial systems* University of Mic igan Press 975

[ ] Z Mic alewicz *Genetic algorithms + data structures = evolution programs* Springer Verlag 99

[4] D E Gol berg "Genetic algorit ms in searc optimization an mac ine learning" *Addison-Wesley* Rea ing MA 989

[5] I Nabney *Netlab: algorithms for pattern recognition* Springer Berlin

[ ] F Gasperi "T e mozzarella c eese flavour profile: a comparison between ju ge panel an proton transfer reaction mass spectroscopy" *Journal of the Science of Food and Agriculture* vol 8 pp 57

[7] L Lvova S S Kim SS A Legin Y Vlasov J S Yang G S C a an H Nam "All soli state electronic tongue an its application for beverage analysis" *Analytica Chimica Acta* vol 4 8 pp 4

[8] Y Vlasov A Legin an A Ru nitskaya "Electronic tongues an t eir analytical application" *Analytical and Biological Chemistry* vol 7 pp 4

[9] A C Noble an S E Ebeler "Use of multivariate statistics in un erstan ing wine flavor" *Food Reviews International* vol 8 pp