Interspeech 2018
2-6 September 2018, Hyderabad

Combining Natural Gradient with Hessian Free Methods for Sequence Training

Adnan Haider, Philip C. Woodland

Cambridge University Engineering Dept., Trumpington St., Cambridge, CB2 1PZ U.K.
{mah90, pcw}@eng.cam.ac.uk

Abstract

This paper presents a new optimisation approach to train Deep Neural Networks (DNNs) with discriminative sequence criteria. At each iteration, the method combines information from the Natural Gradient (NG) direction with local curvature information of the error surface that enables better paths on the parameter manifold to be traversed. The method has been applied within a Hessian Free (HF) style optimisation framework to sequence train both standard fully-connected DNNs and Time Delay Neural Networks as speech recognition acoustic models. The efficacy of the method is shown using experiments on a Multi-Genre Broadcast (MGB) transcription task and neural networks using sigmoid and ReLU activation functions have been investigated. It is shown that for the same number of updates this proposed approach achieves larger reductions in the word error rate (WER) than both NG and HF, and also leads to a lower WER than standard stochastic gradient descent.

Index Terms: Natural Gradient, Hessian Free, NGHF, sequence training, overfitting, speech recognition.

1. Introduction

In recent years, Deep Neural Networks (DNNs) embedded within a hybrid Hidden Markov model (HMM) framework have become the standard approach to Automatic Speech Recognition (ASR) tasks [1]. While the structure of such DNN models gives rich modelling capacity and yields good performance, it also creates complex dependencies between the parameters which can make learning difficult via first order Stochastic Gradient Descent (SGD).

Natural Gradient (NG) [2, 3] descent is an optimisation method traditionally motivated from the perspective of information geometry, and works well for many applications [4, 5, 6] as an alternative to SGD. In our previous work [7], it was shown how the method when framed in a Hessian Free (HF) styled [8, 9] optimisation framework is more effective than either variants of SGD or HF for discriminative sequence training of hybrid HMM-DNN acoustic models. However, the efficacy of this form of NG training fails to extend to DNN models that utilise Rectified Linear Units (ReLUs) [10].

This paper proposes NGHF, an alternative optimisation method that combines both the NG and HF approaches to effectively train HMM-DNN models with discriminative sequence criteria. The NGHF method uses both the direction of steepest descent on a probabilistic manifold and local curvature information, and is effective for different feed-forward DNN architectures and choices of activation function. The method is evaluated on the Multi-Genre Broadcast (MGB) transcription task [11] and is shown to achieve larger reductions in the Word Error Rate (WER) for the same number of updates than both NG and HF, as well as lower WERs than SGD. The machinery needed to develop this framework relies on the concepts of manifolds, tangent vectors and directional derivatives from the perspective of information geometry. An overview of the necessary underlying concepts is provided in [12] but a more in-depth discussion can be found in Amari's infomation geometry text book [13].

The paper is organised as follows. Section 2 provides a brief overview of discriminative sequence training and compares standard derivative based optimisers with NG. Section 3 formulates the method of NGHF, and Sec. 4 discusses the effect of scaling directions with the Gaussian-Newton matrix. The experimental setup for ASR experiments is given in Sec. 5, with results in Sec. 6, followed by the conclusion.

2. Discriminative Sequence Training

ASR is a sequence to sequence level classification task where given an acoustic waveform $\mathcal{O}$, the goal is to produce the correct hypothesis sequence $\mathcal{H}$ through the use of an inference model $P_{\boldsymbol{\theta}}(\mathcal{H}|\mathcal{O})$. Let X denote the parameter manifold. As different realisations of DNN parameters lead to different probabilistic models $P_{\boldsymbol{\theta}}(\mathcal{H}|\mathcal{O})$, the manifold essentially captures the space of all probability distributions $\mathcal{M}$ that can be generated by a particular model. The goal of learning is to identify a viable candidate $f(\boldsymbol{\theta}, \mathcal{O}) \in \mathcal{M}$ that achieves the greatest reduction in the empirical loss w.r.t a given risk measure while generalising well to new examples. In ASR, the WER is the evaluation metric of interest which however corresponds to a discontinuous function of the model parameters. Hence, employing such a metric directly within a empirical risk criterion is not viable with standard derivative based optimisers. This forms the motivation behind the class of Minimum Bayes' Risk (MBR) objective functions [14, 15]:

$$F_{\text{MBR}}(\boldsymbol{\theta}) = \frac{1}{R}\sum_{r}^{R}\left[\sum_{\mathcal{H}} P_{\boldsymbol{\theta}}(\mathcal{H}|\mathcal{O}^r, \mathcal{M}) L(\mathcal{H}, \mathcal{H}^r)\right] \quad (1)$$

where $(\mathcal{H}^r, \mathcal{O}^r)$ represents the true transcription and feature vectors associated with utterance r, and L represents the loss function. In MBR training, instead of minimising the empirical loss for each utterance, the expected loss associated with each utterance in the training set is minimised. Such a function is a smooth function of the DNN model parameters and hence can be optimised by derivative based approaches. In practice, it is not feasible to consider the entire hypothesis space for each utterance without making simplifications to the HMM topology [16]. The standard approach is to encode confusable hypotheses for with each training utterance using an efficient lattice framework [17].

The premise behind all derivative based optimisation methods is Taylor's theorem. Assuming that the objective function

Adnan Haider has been funded by the IDB Cambridge International Scholarship. Many thanks to Dr Chao Zhang for helping us build the SGD baseline models used for this work.

10.21437/Interspeech.2018-2335

$F(\boldsymbol{\theta})$ is sufficiently smooth, Taylor's formula including terms up to the second order approximates the local behaviour of the objective function by the following quadratic function:

$$F(\boldsymbol{\theta}_k + \Delta\boldsymbol{\theta}) \simeq F(\boldsymbol{\theta}_k) + \Delta\boldsymbol{\theta}^T \nabla F(\boldsymbol{\theta}_k) + \frac{1}{2}\Delta\boldsymbol{\theta}^T H \Delta\boldsymbol{\theta} \quad (2)$$

where $\Delta\boldsymbol{\theta}$ corresponds to any offset within a convex neighbourhood of $\boldsymbol{\theta}_k$ and H is the Hessian. Instead of optimising the objective function directly, second order methods focus on minimising the approximate quadratic at each iteration of the optimisation process. The same approach is undertaken by first order methods which only consider the gradient $\nabla F(\boldsymbol{\theta}_k)$, i.e. the first order term. For the class of MBR objective functions, the gradient associated with the rth utterance at time t w.r.t the DNN output activations can be shown to be the component-wise multiplication of the vectors $\boldsymbol{\gamma}_t^r \odot \boldsymbol{L}$ [18, 19, 7], where $\boldsymbol{\gamma}_t^r$ represents the posterior probability associated with the states (DNN output nodes) at time t and the entries of $\boldsymbol{L}$ correspond to the local phone(state) level loss associated with these arcs within the consolidated lattice [15].

Solving (2) yields the critical point $\Delta\boldsymbol{\theta} = H^{-1}\nabla F(\boldsymbol{\theta}_k)$. This corresponds to a unique minimiser only when the Hessian H is positive definite. However, when the choice of models $\mathcal{M}$ is restricted to DNNs, the Hessian irrespective of the choice of objective function is no longer guaranteed to be positive definite. To address this issue, [20] showed that by approximating the Hessian with the Gauss-Newton [21] matrix, solving (2) guarantees an improvement in the training objective function. When the underlying model corresponds to a discriminative probabilistic model $P_{\boldsymbol{\theta}}(\mathcal{H}|\mathcal{O})$, a more natural optimisation method is the method of NG. In NG, the updates associated with each iteration correspond to the direction of steepest decent on the probabilistic manifold. In [2, 3, 22], such a direction is shown to equate to the critical point of (2) with the Hessian replaced by the Fisher Information matrix [13].

3. Formulating NGHF

In [12], it is shown that by deriving Taylor's second order approximation from the perspective of manifold theory, solving the minimisation problem of (2) becomes equivalent to solving the following minimisation problem in the tangent space $T_{\boldsymbol{\theta}_k}X$:

$$\underset{\Delta\boldsymbol{\theta}\in T_{\boldsymbol{\theta}_k}X}{\arg\min} \left[F(\boldsymbol{\theta}_k) + \langle \Delta\boldsymbol{\theta}, \nabla F(\boldsymbol{\theta}_k)\rangle + \frac{1}{2}\Delta\boldsymbol{\theta}^T H \Delta\boldsymbol{\theta} \right] \quad (3)$$

With such an approach, $\langle \Delta\boldsymbol{\theta}, \nabla F(\boldsymbol{\theta}_k)\rangle$ corresponds to the inner product between vectors $\nabla F(\boldsymbol{\theta}_k)$ and $\Delta\boldsymbol{\theta}$ in $T_{\boldsymbol{\theta}_k}X$ while $\Delta\boldsymbol{\theta}^T H \Delta\boldsymbol{\theta}$ represents a linear map $g : \boldsymbol{u} \in T_{\boldsymbol{\theta}_k}X \rightarrow \mathbb{R}$. Since X is a manifold, the inner product endowed on the tangent space $T_{\boldsymbol{\theta}}X$ need not be the identity matrix. The parameter manifold X can be equipped with any form of a Riemannian metric, a smooth map that assigns to each $\boldsymbol{\theta} \in X$ an inner product $I_{\boldsymbol{\theta}}$ in $T_{\boldsymbol{\theta}}X$. In our previous work [7], it was shown how for sequence discriminative training, an ideal choice of $I_{\boldsymbol{\theta}}$ corresponds to the expectation of the outer product of the Maximum Mutual Information (MMI) [17] gradient:

$$I_{\boldsymbol{\theta}} = E_{P_{\boldsymbol{\theta}}(\mathcal{H}|\mathcal{O})}\left[\left(\nabla \log P_{\boldsymbol{\theta}}(\mathcal{H}|\mathcal{O})\right) \left(\nabla \log P_{\boldsymbol{\theta}}(\mathcal{H}|\mathcal{O})\right)^T \right]$$

As $I_{\boldsymbol{\theta}}$ by definition is symmetric and positive definite, it is invertible by the spectral decomposition theorem. If the manifold X is now equipped with a Riemannian metric of the form of $I_{\boldsymbol{\theta}}^{-1}$, then the dot product $\langle \Delta\boldsymbol{\theta}, \nabla F(\boldsymbol{\theta}_k)\rangle$ in (3) corresponds to $\Delta\boldsymbol{\theta}^T I_{\boldsymbol{\theta}}^{-1} \nabla F(\boldsymbol{\theta}_k)$. Under such a metric, solving the minimising problem by considering only the first two terms in (3) equates to performing Natural Gradient on the parameter surface. In this paper, the entire quadratic function of (3) is considered when solving the minimisation problem in $T_{\boldsymbol{\theta}_k}X$:

$$\underset{\Delta\boldsymbol{\theta}\in T_{\boldsymbol{\theta}_k}X}{\arg\min} \left[F(\boldsymbol{\theta}_k) + \Delta\boldsymbol{\theta}^T I_{\boldsymbol{\theta}}^{-1} \nabla F(\boldsymbol{\theta}_k) + \frac{1}{2}\Delta\boldsymbol{\theta}^T H \Delta\boldsymbol{\theta} \right] \quad (4)$$

In practice, the expectation of the outer product of the MMI gradient is approximated by its Monte-Carlo estimate $\hat{I}_{\boldsymbol{\theta}}$ which is not guaranteed to be positive definite. Thus, its inverse is no longer guaranteed to exist. To address this issue, [12] provides the derivation of an alternative dampened Riemannian metric $\tilde{I}_{\boldsymbol{\theta}}^{-1}$ that is not only guaranteed to be positive definite but has the feature that its image space is the direct sum of the image and the kernel space of the empirical Fisher matrix $\hat{I}_{\boldsymbol{\theta}}^{-1}$.

The critical point of (4) is $\Delta\boldsymbol{\theta} = H^{-1} I_{\boldsymbol{\theta}}^{-1} \nabla F(\boldsymbol{\theta}_k)$ and corresponds to an NG direction regularised by multiplication with the inverse of the curvature matrix. In this work, the Hessian is approximated by the Gauss-Newton (GN) Matrix $G_{\boldsymbol{\theta}}$. Section 4 discusses the particular effect of scaling directions with $G_{\boldsymbol{\theta}}$. Computing the individual inverse matrix scalings directly is expensive in terms of both computation and storage. Hence in this paper, using the approach highlighted in [7, 8], the solution of individual inverse matrix scalings is approximated by solving equivalent linear systems using the Conjugate Gradient (CG) [23] algorithm. Apart from the obvious computational reasons, the use of CG presents two key advantages: the very first iteration of CG computes an appropriate step size for the direction the algorithm is initialised with. In the case of NGHF, this corresponds to the NG direction. Thus, at each iteration of NGHF, the resultant update found after two runs of CG conforms to $\Delta\boldsymbol{\theta} = w_1 \Delta\boldsymbol{\theta}_{NG} + w_2 \Delta\boldsymbol{\theta}_{HF}$, a weighted combination of the NG direction and conjugate directions computed using local curvature information. Secondly, when applied to solve the proposed linear system $\tilde{I}_{\boldsymbol{\theta}} \Delta\boldsymbol{\theta} = \nabla F(\boldsymbol{\theta}_k)$, the very first directions explored by the algorithm are guaranteed to be the directions which constitute the image space of $\hat{I}_{\boldsymbol{\theta}}$ [24].

4. Scaling Directions with the GN Matrix

When DNN models are employed to solve the inference problem, $G_{\boldsymbol{\theta}}$ can be shown to take the particular form of $J_{\boldsymbol{\theta}}^T \nabla^2 \hat{L}_{\boldsymbol{\theta}} J_{\boldsymbol{\theta}}$ where

- $\nabla^2 \hat{L}_{\boldsymbol{\theta}}$ is the Hessian of the loss function w.r.t the DNN linear output activations with individual entries being functions of $\boldsymbol{\theta}$.
- $J_{\boldsymbol{\theta}}$ is the Jacobian of the DNN output activations w.r.t $\boldsymbol{\theta}$.

To keep the notation uncluttered, the dependency on $\boldsymbol{\theta}$ will be dropped for the remainder of this section when dealing with the individual factors of the product $J_{\boldsymbol{\theta}}^T \nabla^2 \hat{L}_{\boldsymbol{\theta}} J_{\boldsymbol{\theta}}$. As both $\nabla^2 \hat{L}$ and the product $J^T \nabla^2 \hat{L} J$ are real and symmetric, by the spectral decomposition theorem: $J^T \nabla^2 \hat{L} J \equiv J^T U \Lambda U^T J \equiv V \Sigma V^T$. Under this factorisation, each eigenvector $\boldsymbol{v}_j$ of $J^T \nabla^2 \hat{L} J$ can be interpreted as a particular weighted sum of the gradients of the output activations of the DNN w.r.t θ. Switching from the standard basis to the basis spanned by $U^T J$, updates conforming to directions of steepest descent can be expressed as:

$$\Delta\theta = \sum_j^k \eta \left(\boldsymbol{v}_j^T \nabla F_{\mathrm{MBR}} \right) U_{j,i} \frac{\partial \tilde{F}_i}{\partial \theta}$$

where η is the learning rate and $\tilde{F}$ represents the objective function with the domain constrained to the DNN output layer. With respect to this basis, scaling with the inverse of the GN matrix effectively corresponds to rescaling the steps taken along individual $\boldsymbol{v}_j$ by a factor $1/\mu_j$:

$$\Delta\theta = \sum_{j}^{k} \frac{1}{\mu_j} \left(\boldsymbol{v}_j^T \nabla F_{\text{MBR}} \right) U_{j,i} \frac{\partial \tilde{F}_i}{\partial \theta} \tag{5}$$

where μ_j is the eigenvalue associated with $\boldsymbol{v}_j$ in V. Recall that $\nabla^2 \tilde{L}$ can alternatively be presented as $\nabla.(\nabla \tilde{F})$. Therefore, eigenvectors $\boldsymbol{u}_j$ in U with large eigenvalues correspond to directions that can induce large changes in the gradient of $\nabla \tilde{F}$. By establishing a one-to one correspondence between eigenvectors of $\nabla^2 \tilde{L}$ with eigenvectors of $J^T A J$, it can be seen that re-scaling with $1/\mu_j$ effectively de-weights back propagation information carried by those paths through the network that can induce large changes in $\nabla \tilde{F}$. In the context of discriminative training, this ensures that the DNN frame posterior distribution does not become overly sharp.

5. Experimental Setup

The various DNN optimisation approaches were evaluated on data from the 2015 Multi-Genre Broadcast ASRU challenge task (MGB1) [11]. In this work, systems were trained using a 200 hour training set[1]. The official MGB1 dev.sub set was employed as a validation set and consists of 5.5 hours of audio data. To estimate the generalisation performance of candidate models, a separate evaluation test set dev.sub2 was used. This comprises roughly of 23 hours of audio from the remaining 35 shows belonging the MGB1 dev.full set. Further details related to the data preparation can be found in [25].

All experiments were conducted using an extended version of the HTK 3.5 toolkit [26, 27]. This paper focuses on training standard fully connected DNNs and Time Delay Neural Networks (TDNNs) [28] using both ReLU and sigmoid activations. The architecture used for DNNs consisted of five hidden layers each with 1000 nodes. For TDNNs, the network topology consisted of seven hidden layers each with 1000 hidden units. The context specification used for the various TDNN layers is as follows: [-2, +2] for layer 1, $\{-1, 2\}$ for layer 2, $\{-3, 3\}$ for layer 3, $\{-7, 2\}$ for layer 4 and [0] for the remaining layers. For both models, the output layer consisted of 6k nodes and corresponds to context dependent sub-phone targets formed by conventional decision tree context dependent state tying [29]. For DNNs, the input to the model was produced by splicing together 40 dimensional log-Mel filter bank (FBK) features extended with their delta coefficients across 9 frames to give a 720 dimensional input per frame. While for TDNNs, only the 40 dimensional log-Mel filter bank features were considered. For all experiments, the input features were normalised at the utterance level for mean and at the show-segment level for variance [25].

All models were trained using lattice-based MPE training [14]. Prior to sequence training, the model parameters were initialised using frame-level CE training. To track the occurrence of over-fitting due to training criterion mismatch at intermediate stages of sequence training, decoding was performed on the validation set using the same weak pruned biased LM used to create the initial MPE lattices. To evaluate the generalisation performance of the trained models, a 158k word vocabulary trigram LM was used to decode the validation and test set.

[1] Note that most results in [25] use a larger 700h training set, stronger language models and other setup differences.

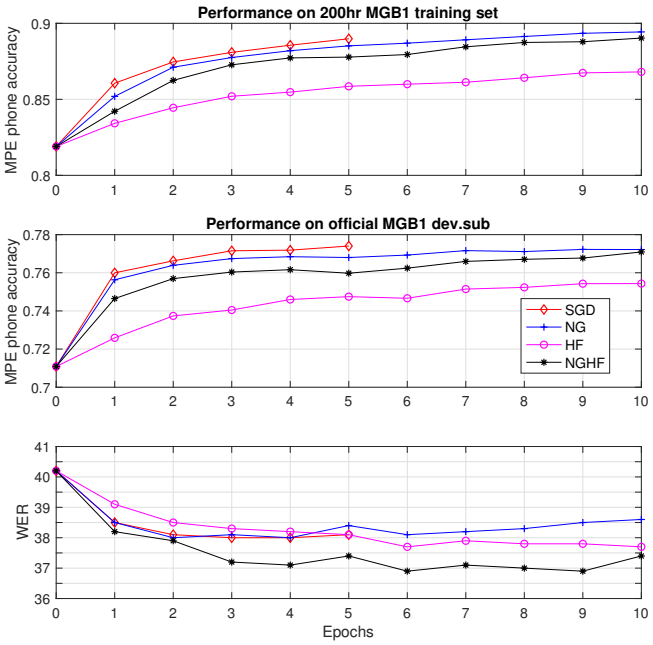

Figure 1: *Evolution of MPE phone accuracy criterion on the training and validation (dev.sub) sets with ReLU based DNN (top 2 graphs). Also (lower graph) WER with MPE LM on dev.sub as training proceeds.*

Training configuration for SGD: The best results with SGD were achieved through annealing of the learning rates at subsequent epochs. The initial learning rates were found through a grid search. For TDNNs, using momentum was found to yield the best WERs.

Training configuration for NGHF, NG & HF: The recipe described in [7] was used: gradient batches corresponding to roughly 25 hours and 0.5 hrs of audio were sampled for each CG run. In all experiments, running each CG run beyond 8 iterations was not found to be advantageous. The CG computations varied between 18% to 26% of the total computational cost.

6. Experimental Results

Figure 1 compares the performance of various optimisation methods on training a ReLU based 200hr HMM-DNN model while Table 1 shows the performance of these optimisers for a ReLU based HMM-TDNN system.

method	#epochs	#updates	phone acc.		WER with
			train	dev.sub	MPE LM
CE	N/A	N/A	0.870	0.754	36.9
SGD	4	4.64×10^5	0.888	0.760	36.4
NG	4	32	0.913	0.789	36.2
HF	4	32	0.899	0.783	35.9
NGHF	4	32	0.911	0.791	**35.6**

Table 1: *Performance of different optimisers on the TDNN-ReLU model. WERs on dev.sub.*

It can be seen that among all the optimisers, NGHF is the most effective in achieving the largest WER reductions on dev.sub with the weak MPE LM. At each iteration, the update produced by the method conforms to $\Delta\boldsymbol{\theta} = w_1 \Delta\boldsymbol{\theta}_{NG} + w_2 \Delta\boldsymbol{\theta}_{HF}$, a weighted combination of NG direction and conjugate directions computed using local curvature information. In Fig. 1, it can be seen that by utilising information from both the

KL divergence in the probabilistic manifold and local curvature information, the proposed method follows a path where optimising the MPE criterion better correlates with achieving reductions in WER. With the ReLU based TDNN as evident from Table 1, this same feature can be observed. NGHF achieves better generalisation performance for both the MPE criterion and the WER on the validation set. To investigate whether these WER reductions hold with stronger LMs and the relative gains are not constrained to only ReLU based systems, equivalent systems using sigmoids were trained. Table 2 compares the performance of the various optimisers on the validation set with the different models using the 158k LM.

Model	CE	MPE			
		SGD	NG	HF	NGHF
DNN-ReLU	30.9	29.9	29.8	28.9	**28.1**
TDNN-ReLU	28.6	28.5	28.7	28.1	**27.5**
DNN-sigmoid	31.9	29.3	29.0	29.3	**29.0**
TDNN-sigmoid	28.5	27.1	26.9	27.0	**26.6**

Table 2: *WERs on MGB1 dev.sub with 158k trigram LM.*

From Table 2, it can be seen that again models using NGHF achieve the largest reductions in WER. For ReLU based models, NGHF achieves a relative Word Error Rate Reduction (WERR) of 9% with the DNN and 4% with the TDNN. Whereas with the sigmoid based models, the method achieves a relative WERR of 6% with the DNN and 7% with the TDNN. Compared to SGD, NGHF is especially effective with the ReLU based models. For the DNN, the method achieves a relative WERR of 6% over SGD, while with the TDNN the relative WERR is 4%.

Finally, the generalisation performance of the trained models were estimated by performing Viterbi decoding on dev.sub2 using 158k LM. Results are shown in Table 3.

Model	CE	MPE			
		SGD	NG	HF	NGHF
DNN-RELU	32.3	31.9	31.4	30.6	**29.8**
TDNN-RELU	30.6	29.8	30.6	29.6	**29.3**
DNN-sigmoid	33.2	30.8	30.5	30.9	**30.5**
TDNN-sigmoid	29.9	28.6	28.2	28.4	**27.9**

Table 3: *WERs on MGB1 dev.sub2 with 158k trigram LM.*

It can be observed that as before the model trained with NGHF achieves the largest reductions in WER. With sigmoid based models, the method can be seen to be achieve WERR reductions of 8% with the DNN and 7% with the TDNN. Over standard SGD, the proposed method achieves relative WERR of 7% with the ReLU-DNN model and a 2% with the ReLU based TDNN model.

6.1. Investigating overfitting due to Criterion Mismatch

ReLU based systems failed to achieve similar WERRs as sigmoid based systems from sequence training with either SGD or NG. After a few epochs, improvements made on the MPE criterion failed to correlate with lower WERs. This effect was particularly noticeable with the TDNN model. It was observed that this emergence of criterion mismatch is correlated with the sharp decrease in the average entropy of DNN output frame posteriors (Fig. 2). MBR training broadly speaking tries to concentrate probability mass: a sufficiently flexible model trained to convergence with MBR will assign a high probability to those hypotheses that have the smallest loss. This means that during

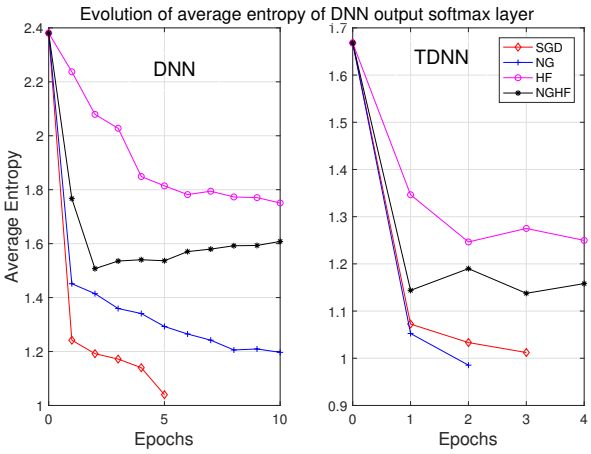

Figure 2: *Evolution of average entropy of DNN output activations during MPE training with ReLU based DNN models. Left graph is for DNNs and the right graph for TDNNs.*

the course of training, the posterior distribution of states $\boldsymbol{\gamma}_t^r(i)$ associated with high local losses $\boldsymbol{L}(i)$ is gradually reduced. With hyper-parameters such as LM and acoustic model scale factors fixed, this sharp decrease in the DNN output entropy directly reflects that this distribution $\boldsymbol{\gamma}_t^r$ is becoming overly sharp, which was found to be detrimental to the decoding performance.

With sigmoid models, the criterion mismatch was found to be less severe when using first order methods (Table 2). It was observed that the average entropy of the DNN frame posteriors with sigmoid models was much larger at the start of sequence training. This is expected as ReLUs allow a better flow of gradients during back-propagation resulting in better CE trained discriminative models. However, as observed in Fig. 2, this also results in sharper frame posterior distributions. From Fig. 2, it can also be seen how scaling with the GN matrix helps regularise the entropy of the DNN frame posteriors. When compared to HF, the proposed NGHF approach is better in finding a balance between improving the MPE criterion and regularising the entropy changes of the DNN frame posteriors.

To improve generalisation performance, techniques such as dropout [30] and L2 regularisation with SGD sequence training were investigated. However, both of these techniques were unsuccessful in alleviating this overfitting due to criterion mismatch. To improve NG training, the use of Tikhonov damping as advised by Martens [9] was also investigated to help regularise the NG updates. Taking comparatively more conservative steps along conjugate directions at the expense of slower learning was observed to regulate the decrease in the average entropy of DNN frame posteriors. However, the damped optimiser failed to achieve better convergence.

7. Conclusion

This paper has introduced a new optimisation method to effectively train HMM-DNN models with discriminative sequence training. The efficacy of the method has been shown to be agnostic with respect to both the choice of feed forward architecture and choice of DNN activation functions. When applied within a HF styled optimisation framework, the proposed methods enjoys the same benefits as HF but leads to better convergence than NG, HF and SGD. Future work will involve extending the proposed framework to training DNN architectures with recurrent topologies.

8. References

[1] G. Hinton, L. Deng, D. Yu, G. Dahl, A. Mohamed, N. Jaitly, A. Senior, V. Vanhoucke, P. Nguyen, T.N. Sainath & B Kingsbury, "Deep Neural Networks for Acoustic Modeling in Speech Recognition: The Shared Views of Four Research Groups", *IEEE Signal Processing Magazine*, vol. 29, no. 6, pp. 82–97, 2012.

[2] S. Amari, "Natural Gradient Works Efficiently in Learning", *Proc. Neural Information Processing Systems (NIPS)*, 1998.

[3] S. Amari, "Neural Learning in Structured Parameter Spaces–Natural Riemannian Gradient", *Proc. Advances in Neural Information Processing Systems (NIPS)*, 1997.

[4] R. Pascanu &Y. Bengio, "Revisiting Natural Gradient for Deep Networks", *arXiv preprint arXiv:1301.3584*, Jan. 2013.

[5] G. Desjardins, K. Simonyan & R. Pascanu, "Natural Neural Networks", *Proc. Advances in Neural Information Processing Systems (NIPS)*, 2015.

[6] D. Povey, X. Zhang & S. Khudanpur, "Parallel Training of DNNs with Natural Gradient and Parameter Averaging", *arXiv preprint arXiv:1410.7455*, Oct. 2014.

[7] A. Haider & P.C. Woodland, "Sequence Training of DNN Acoustic Models with Natural Gradient", *IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU)*, Dec. 2017.

[8] B. Kingsbury, T.N. Sainath & H. Soltau, "Scalable Minimum Bayes Risk Training of Deep Neural Network Acoustic Models using Distributed Hessian-Free Optimisation", *Proc. Interspeech*, 2012.

[9] J. Martens, "Deep Learning via Hessian-Free Optimisation", *Proc. International Conference on Machine Learning (ICML)*, 2010.

[10] V. Nair & G. Hinton, "Rectified Linear Units Improve Restricted Boltzmann Machines", *Proc. International Conference on Machine Learning (ICML)*, 2010.

[11] P. Bell, M. Gales, T. Hain, J Kilgour, P. Lanchantin, X Liu, A McParland, S. Renals, S. Saz, N Wester & P. Woodland, "The MGB Challenge: Evaluating Multi-Genre Broadcast Media Recognition", *Proc. IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU)*, 2015.

[12] A.Haider, "A Common Framework for Natural Gradient and Taylor based Optimisation using Manifold Theory", *arXiv preprint* 2018.

[13] S. Amari, " Information Geometry and its Applications", *Springer 2016*, ch. 1, pp. 3-27.

[14] D. Povey & P.C. Woodland, "Minimum Phone Error and I-smoothing for Improved Discriminative Training," *Proc. ICASSP*, 2002.

[15] M. Gibson & T. Hain, "Hypothesis Spaces for Minimum Bayes' Risk Training in Large Vocabulary Speech Recognition", *Proc. Interspeech*, 2006.

[16] D. Povey, V. Peddinti, D. Galvez,, P. Ghahremani, V. Manohar, X. Na,Y. Wang & S.Khudanpur, "Purely Sequence-Trained Neural Networks for ASR based on Lattice-Free MMI", *Proc. Interspeech*, 2016.

[17] P.C. Woodland & D. Povey, "Large Scale Discriminative Training of Hidden Markov Models for Speech Recognition", *Computer Speech and Language*, vol. 16. pp. 25-47, 2002.

[18] K. Vesely, A. Ghosal, L. Burget & D. Povey, "Sequence-Discriminative Training of Deep Neural Networks", *Proc. Interspeech*, 2013.

[19] M. Shannon, "Optimizing Expected Word Error Rate via Sampling for Speech Recognition", *Proc. Interspeech*, 2017.

[20] T.N. Sainath, B. Kingsbury & H. Soltau, "Optimization Techniques to Improve Training Speed of Deep Neural Networks for Large Speech Tasks", *IEEE Trans. on Audio, Speech, and Language Processing*, vol. 21, no. 11, pp. 2267–2276, 2013.

[21] N. Schraudolph, "Fast Curvature Matrix-Vector Products for Second-Order Gradient Descent", *Proc. Neural Information Processing Systems (NIPS)*, 2002.

[22] L. Bottou, F. Curtis & J. Nocedal, *Optimization Methods for Large-Scale Machine Learning*, arXiv preprint arXiv:1606.04838, 2016.

[23] J.R. Shewchuk, "An Introduction to the Conjugate Gradient Method Without the Agonizing Pain", 1994.

[24] J. Nocedal & S. Wright, "Numerical Optimization", *2nd Edition, Springer series in Operations Research*, ch. 5, pp. 115.

[25] P.C. Woodland, X. Lui, Y. Qian, C. Zhang, M.J.F. Gales, P. Karanasou, P. Lanchantin & L. Wang, "Cambridge University Transcription Systems for the Multi-Genre Broadcast Challenge", *Proc. IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU)*, 2015.

[26] C. Zhang & P.C. Woodland, "A General Artificial Neural Network Extension for HTK", *Proc. Interspeech*, 2015.

[27] S.J. Young, G. Evermann, M.J.F. Gales, T. Hain, D. Kershaw, X. Liu, G. Moore, J.J. Odell, J. Ollason, D. Povey, A. Ragni, V. Valtchev, P.C. Woodland & C. Zhang, *The HTK Book (for HTK 3.5)*, Cambridge University Engineering Department, 2015.

[28] V. Peddinti, D. Povey and S.Khudanpur "A Time Delay Neural Network Architecture for Efficient Modeling of Long Temporal Contexts" *Proc. InterSpeech* 2015.

[29] S.J. Young, J.J. Odell & P.C. Woodland "Tree-Based State Tying for High Accuracy Acoustic Modelling", *Proc HLT*, 1994.

[30] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever & R.Salakhutdinov, "Dropout: A Simple Way to Prevent Neural Networks from Overfitting", *The Journal of Machine Learning Research*, vol. 15, no. 1, pp. 1929-1958, 2014.