Comparison between Hessian Free and Natural Gradient training for DNN acoustic models

Adnan Haider and Phil Woodland

Cambridge University Engineering Department

Abstract

- Compare Natural Gradient (NG) against Hessian Free (HF) and Dynamic Stochastic Average Gradient HF (DSAG-HF) [1] for Sequence training with large batch sizes.
- Effectiveness of both methods evaluated on BBC Multi-Genre Broadcast (MGB) 1 dataset.

Training of DNNs in ASR

- Frame-based training : Cross Entropy (CE) criterion.
- Sequence training:
  - Maximum Mutual Information (MMI)
    - maximise the sentence-level posterior probability of the correct utterance.
  - Minimum Bayes' Risk (MBR)
    -minimises the average expected loss computed over the hypothesis space.
    Typical Loss functions : phone error rate (MPE) or HMM state-id error (sMBR).

Hessian Free and Natural Gradient

- Hessian Free (HF) approach:
  - At each iteration, minimises a Taylor approximation of the objective function.
  - Uses a Gauss Newton approximation of the Hessian matrix.
- Natural Gradient (NG) [2] approach:
  - Corrects the gradient of $F(\theta)$ according to the local curvature of the KL-divergence surface.
  - Solves first order minimisation problem within a trust region.

Similarities between both methods

- Instead of minimising the objective function $F(\theta)$ directly, both methods, at each iteration minimise a quadratic of the form:

$$F(\theta_k) + \nabla F(\theta_k)\Delta\theta + \frac{1}{2}\Delta\theta^T J^T B J \Delta\theta \qquad (1)$$

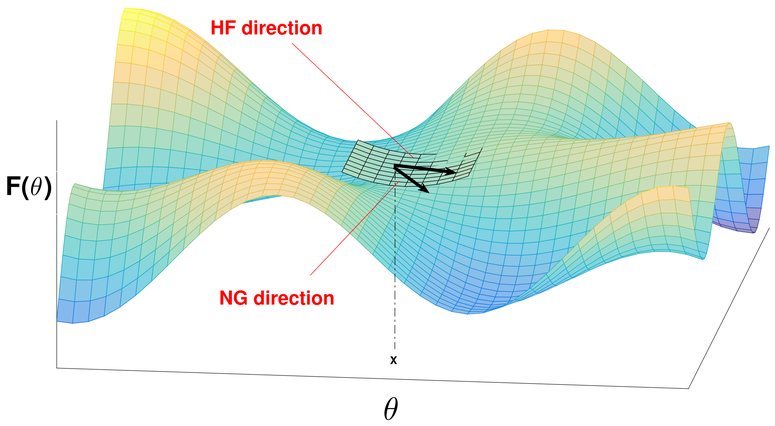
- J is the Jacobian of the linear output activations w.r.t θ.

Key differences

- Both methods primarily differ in the choice of the matrix B:

Method	CE training	Sequence training
HF	$\nabla^2 L_{\text{CE}}$	$\nabla^2 L_{\text{MBR/MMI}}$
NG	$-\nabla L_{\text{CE}} \nabla L_{\text{CE}}^T$	$\nabla L_{\text{MMI}} \nabla L_{\text{MMI}}^T$

Error surface of Sequence discriminative criterion



Conjugate Gradients

- The Newton direction yields the minimum of the quadratic:

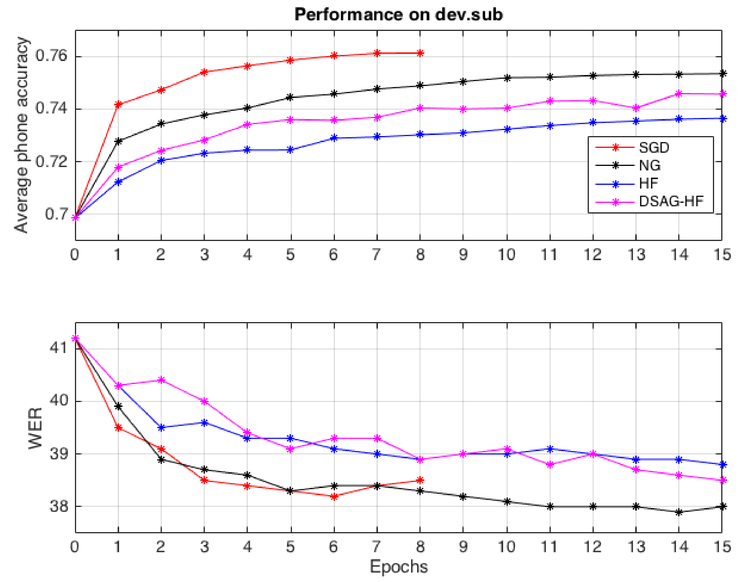
$$\Delta\theta = (J^T B J)^{-1} \nabla F(\theta)$$

- Inverting the matrix incurs a cost of $\mathcal{O}(D^3)$.
- Instead iteratively solve the system $J^T B J \Delta\theta = \nabla F(\theta)$ using Linear Conjugate Gradient.

Experimental Setup

- Experiments used ASRU 2015 MGB challenge data.
  - Training set : 200 hr MGB1 training dataset.
  - Validation : the official MGB1 dev.sub set (5.5 hours of audio sampled from 12 shows).
  - Test set : dev.sub2 audio from the remaining 35 shows in the MGB1 dev.full set.
- 158k word dictionary
- All systems were trained and decoded using extended version of HTK 3.5.
- DNN model architecture : 5 hidden layers of 1000 nodes sigmoid activation functions.
- DNN initialised using frame-based CE with Stochastic Gradient Descent (SGD).
- For both NG and variants of HF optimisation, batch sizes of roughly 25 hrs were used.
- For running CG, roughly 1% of the training set was sampled to compute matrix vector products.

Experimental Results



LM	SGD	HF	DSAG-HF	NG
158k Bigram	35.0	35.3	35.2	34.7
158k Trigram	29.3	29.3	29.2	29.0

Table 1: %WER differences between different optimisers on dev.sub with 158k vocabulary bigram/trigram LMs on dev.sub (200hr).

LM	CE	SGD	HF	DSAG-HF	NG
158k Trigram	33.1	30.8	31.0	30.8	30.5

Table 2: %WER differences between different optimisers on dev.sub2 with 158k trigram (200hr)

Conclusion

- Replacing 'Hessian' component of HF with empirical Fisher Information matrix leads to better and faster convergence in Sequence training.
- NG better addresses over-fitting due to mismatch of training criterion better than SGD or HF.
- On dev.sub2, NG's WER better than both SGD DSAG-HF (statistically significant at 1% level).

References

- P. Dognin & V. Goel, "Combining Stochastic Average Gradient and Hessian-free Optimization for Sequence Training of Deep Neural Networks", *Proc. ASRU, 2013,*
- *S. Amari, "Natural Gradient Works Efficiently in Learning,"* Proc. (NIPS), 1998.
- A .Haider & P. Woodland, "Sequence Training of DNN Acoustic Models With Natural Gradient", *Proc. ASRU, 2017*

E-Mail: mah@cam.ac.uk