Covariance Updates for Discriminative Training by Constrained Line Search

Peter Bell, Simon King

Centre for Speech Technology Research, University of Edinburgh, UK
peter.bell@ed.ac.uk, simon.king@ed.ac.uk

Abstract

We investigate the recent Constrained Line Search algorithm for discriminative training of HMMs and propose an alternative formula for variance update. We compare the method to standard techniques on a phone recognition task.

Index Terms: speech recognition, acoustic models, discriminative training, line search

1. Introduction

In HMM-based systems for automatic speech recognition, discriminative training of acoustic model parameters, which seek to maximise a discriminative objective function such as MMI, has been consistently shown to yield benefits over conventional ML training. Training is performed by iteratively maximising a weak-sense auxiliary function, $\mathcal{F}$ [1]. Following [2], methods for discriminative training typically incorporate a smoothing term into the auxiliary function that, when large enough, ensures the function is convex and is a lower bound for the objective function close to the initial parameters. The process is known as the extended Baum-Welch (EBW) algorithm.

Recently, Liu et al [3] proposed a constrained line search (CLS) algorithm for MMI training. No smoothing function is required: instead, at each iteration the auxiliary function is maximised subject to a constraint on the Kullback-Liebler divergence (KLD) between the old and new parameter sets. This can be approximated by quadratic constraints on the updated mean, log variance and weights. The resulting optimisation is solved by simply limiting the length of the update vector to the radius of the constraint set.

Liu et al reported that CLS achieved consistent performance improvements over EBW on TIDIGITS and Switchboard. However, on TIDIGITS, very little performance change was observed when just mean parameters were updated, compared to updating all parameters. We propose alternative formulae for covariance updates for CLS, and compare performance to EBW on the TIMIT phone recognition task.

2. Covariance updates

CLS typically achieves a much larger performance increase than EBW after the first iteration, in particular. We conjecture that this may be because, unlike EBW, the mean update is constrained independently of the variance, allowing bigger step sizes. However, this presents problems for the variance update since the gradient, $\nabla\mathcal{F}(\sigma)$, and consequently the critical point, is dependent on the new mean. Furthermore, the original auxiliary function is not, in general, convex in σ, and may be unsuitable at the new mean. To remedy this, we propose setting part of the denominator in the auxiliary to be linear in $\log\sigma^2$ and also to update σ using a variant of Newton's method (a quadratic approximation for $\mathcal{F}$). The linearisation is described in [1]. At initial parameters (μ_o, σ_o^2), with β^d, $\mathbf{x}^d$, S^d representing the zeroth, first, and second order denominator statistics, respectively (S being centralised about μ_0) the linear term in one dimension is given by

$$-\frac{1}{2}\log\sigma^2(\beta^d - \frac{S^d}{\sigma_o^2}) + \mu\frac{(\beta^d\mu_0 - \mathbf{x}^d)}{\sigma_0^2} \qquad (1)$$

A factor E weights the contribution of the linear term. The variance update in one dimension may be expressed in terms of $L = \log\sigma^2$ [3]. Given a mean update of $\Delta\mu$, the step size computed for all dimensions is as follows, after which the quadratic constraints are imposed directly on the resulting update vector:

$$-\left(\frac{\partial\mathcal{F}}{\partial L} + \Delta\mu\frac{\partial^2\mathcal{F}}{\partial\mu\partial L}\right)\left(\frac{\partial^2\mathcal{F}}{\partial L^2}\right)^{-1} \qquad (2)$$

3. Experiments

We performed experiments on the TIMIT 39-phone recognition task, using a a standard 3-state HMM for each phone, with emission probabilities modelled as 12-component full-covariance GMMs, initialised by HMM training. Diagonalising transforms were applied before each parameter update; the transforms were not updated. The KLD constraints, ρ, were set as in [3], and an identical I-Smoothing constant applied to all systems. The table below shows test set results for various training routines with mean (m) and variance (v) updates, for 1–6 iterations of estimation. The baseline (ML) accuracy is 71.0%.

Routine	1	2	3	4	5	6
EBW m	71.1	71.2	71.3	71.4	71.3	71.3
EBW m+v	71.2	71.5	71.4	**71.5**	71.4	71.4
CLS m	**71.6**	71.2	71.3	71.4	71.4	71.5
CLS m+v, E=0.0	71.2	71.3	71.0	71.0	70.9	70.9
CLS m+v, E=0.5	**71.8**	71.6	71.2	71.4	71.1	71.3
CLS m+v, E=1.0	71.5	71.4	71.1	71.3	71.1	71.3

These results support earlier findings that CLS is able to outperform EBW, with fewer iterations. However, we found the CLS algorithm very susceptible to over-training on this system. Updating the variance using CLS, it is necessary to include a linear part in the denominator term of $\mathcal{F}$; performance is also sensitive to the weight E. The reliance of CLS on several heuristics, including E and ρ, is a disadvantage.

4. References

[1] D. Povey, "Discriminative training for large vocabulary speech recognition," Ph.D. dissertation, Cambridge University, 2003.

[2] Y. Normandin and S. D. Morgera, "An improved MMIE training algorithm for speaker-independent, small vocabulary, continuous speech recognition," in *Proceedings ICASSP*, Toronto, 1991.

[3] P. Liu, C. Liu, H. Jiang, F. K. Soong, and R.-H. Wang, "A constrained line search approach to general discriminative HMM training," in *Proc. ASRU*, Kyoto, 2007.