# Augmented Statistical Models: Exploiting Generative Models in Discriminative Classifiers

**Martin I. Layton**
Department of Engineering
University of Cambridge
Cambridge, CB2 1PZ
`ml362@eng.cam.ac.uk`

**Mark J. F. Gales**
Department of Engineering
University of Cambridge
Cambridge, CB2 1PZ
`mjfg@eng.cam.ac.uk`

In recent years, many algorithms have been proposed for discriminative classification of data. Popular examples are support vector machines (SVMs) [1] and conditional random fields (CRFs) [2]. These techniques make extensive use of fixed-dimensional mappings from the observation-space to a (often high-dimensional) feature-space. Unfortunately, for applications with variable-length sequences of observations – text processing, speech recognition and computational biology – it is not clear how these mappings should be defined.

For variable-length sequences, it is usual to estimate class-conditional latent-variable generative models, such as Gaussian mixture models (GMMs) and hidden Markov models (HMMs). Bayes' rule is then used to calculate the posterior probability of the class labels. This allows missing data and variable-length sequences to be handled in a simple yet robust manner. However, for many tasks, the independence and conditional-independence assumptions associated with standard latent-variable models are not correct and may degrade classification performance.

In [3], Jaakkola and Haussler proposed the Fisher score-space as a powerful method of incorporating the benefits of generative models within standard discriminative training algorithms for unsupervised learning. First a base (generative) model, $\hat{p}(\boldsymbol{O};\boldsymbol{\lambda})$, is estimated from the training examples using maximum likelihood (ML) or maximum mutual information (MMI) estimation. Next, the generative process is captured in a fixed-dimensional feature-vector using the tangent-space of the base model,

$$\boldsymbol{\phi}^{\mathsf{F}}(\boldsymbol{O};\boldsymbol{\lambda}) = \left[\,\nabla_{\boldsymbol{\lambda}} \ln \hat{p}(\boldsymbol{O};\boldsymbol{\lambda})\,\right] \tag{1}$$

This is then used as the feature-space for training a classifier on small amounts of labelled training data.

Later, Smith and Gales [4] presented an extension for supervised binary classification tasks: generative score-spaces. Instead of a single base model, class-conditional base models $\hat{p}(\boldsymbol{O};\boldsymbol{\lambda}^{(1)})$ and $\hat{p}(\boldsymbol{O};\boldsymbol{\lambda}^{(2)})$ are used, allowing problems such as wrap-around to be avoided [4]. To improve discrimination the log-likelihood ratio of the base models is also included. The resulting score-space is given by[1],

$$\boldsymbol{\phi}^{\mathsf{LL}}(\boldsymbol{O};\boldsymbol{\lambda}) = \begin{bmatrix} \ln \hat{p}(\boldsymbol{O};\boldsymbol{\lambda}^{(1)}) - \ln \hat{p}(\boldsymbol{O};\boldsymbol{\lambda}^{(2)}) \\ \nabla_{\boldsymbol{\lambda}^{(1)}} \ln \hat{p}(\boldsymbol{O};\boldsymbol{\lambda}^{(1)}) \\ -\nabla_{\boldsymbol{\lambda}^{(2)}} \ln \hat{p}(\boldsymbol{O};\boldsymbol{\lambda}^{(2)}) \end{bmatrix} \tag{2}$$

[1] The Fisher score-space is a special case of (2) where base model parameters are constrained to be equal.

where $\boldsymbol{\lambda} = \{\boldsymbol{\lambda}^{(1)}, \boldsymbol{\lambda}^{(2)}\}$. Since both Fisher and generative score-spaces are distance-based techniques, a score-space metric must be defined. An appropriate, maximally non-committal metric, is given by the Fisher information matrix [3,4].

## Augmented Models

In the previous section, both Fisher and generative score-spaces were viewed as a method of mapping variable-length sequences of observations into a fixed-dimensional feature-space suitable for classification. Alternatively, these features can be used as sufficient statistics, $\boldsymbol{T}(\boldsymbol{O};\boldsymbol{\lambda})$, for a statistical model. One such model is the *augmented statistical model*[2] [6,7],

$$p(\boldsymbol{O};\boldsymbol{\lambda},\boldsymbol{\alpha}) = \frac{1}{\tau(\boldsymbol{\lambda},\boldsymbol{\alpha})}\exp\left(\boldsymbol{\alpha}^T\boldsymbol{T}(\boldsymbol{O};\boldsymbol{\lambda})\right) \tag{3}$$

$\boldsymbol{\alpha}$ are augmented parameters and $\tau(\boldsymbol{\lambda},\boldsymbol{\alpha})$ is the normalisation term (an expectation over observation sequences). The sufficient statistics are given by the vector form of base model derivatives of orders 0 through $\rho$ [6,8].

$$\boldsymbol{T}(\boldsymbol{O};\boldsymbol{\lambda}) = \begin{bmatrix} \ln\hat{p}(\boldsymbol{O};\boldsymbol{\lambda}) \\ \nabla_{\boldsymbol{\lambda}}\ln\hat{p}(\boldsymbol{O};\boldsymbol{\lambda}) \\ \vdots \\ \frac{1}{\rho!}\mathrm{vec}(\nabla^{\rho}_{\boldsymbol{\lambda}}\ln\hat{p}(\boldsymbol{O};\boldsymbol{\lambda})) \end{bmatrix} \tag{4}$$

Although it may seem strange to embed a generative model within another statistical model, this is a perfectly valid operation since the generative model is used only to generate sufficient statistics. Compared to arbitrary statistics (such as $\boldsymbol{o}$ and $\boldsymbol{o}^2$), statistics from generative models are advantageous since they are tuned to match the distribution of the data, thus providing a better representation of the underlying source [3]. It is interesting to constrast the nature of dependencies incorporating in augmented models to those of the base model. Since no new statistics are introduced (only new functions of the base model statistics), independence assumptions of the base model are retained. This is not the case, however, for conditional independence assumptions. In particular, derivatives of latent-variable models are a function of the posterior probabilities of the latent states. Since these are dependent on all observations and all latent states, conditional independence is broken.

Unfortunately, with this additional modelling power, augmented models can be difficult to train since the normalisation term often has no closed-form solution. This typically makes ML and MMI estimation of augmented parameters infeasible. Instead a two-stage training algorithm may be used. First, the optimal base model parameters, $\tilde{\boldsymbol{\lambda}}$, are estimated using standard ML or MMI training. This fixes the values of the sufficient statistics, yielding the optimisation,

$$\tilde{\boldsymbol{\alpha}} = \arg\max_{\boldsymbol{\alpha}} \sum_{i=1}^{n} \mathcal{F}(y_i, \boldsymbol{T}(\boldsymbol{O}_i;\tilde{\boldsymbol{\lambda}});\boldsymbol{\alpha}) \tag{5}$$

where $\mathcal{F}(\cdot)$ is the objective function. Augmented parameters can then be trained using one of two discriminative techniques: maximum margin (MM) or conditional maximum likelihood (CML) estimation. Although neither have closed-form solutions, they are both convex and so have unique global solutions. The resulting models have half their parameters trained generatively ($\boldsymbol{\lambda}$) and half trained discriminatively ($\boldsymbol{\alpha}$).

Note that it is also possible to optimise the base model parameters using MM and CML [8]. However, this breaks the convexity of the objective function (by allow the statistics to vary) resulting in a highly complex objective function with many local maxima. Further details of this process are given in [8].

[2] Augmented models also have an elegant interpretation in terms of a $\rho$-th order Taylor series expansion [5,6] about distributions of the base model [6,7].

### Maximum Margin Estimation

A common discriminative training criterion when dealing with high-dimensional feature-spaces is maximum margin (MM) estimation. One popular implementation is the support vector machine (SVM) [1]. Unfortunately, SVMs are inherently binary classifiers and so are normally restricted to binary classification tasks. Given two augmented models, $p(\boldsymbol{O};\boldsymbol{\lambda}^{(1)},\boldsymbol{\alpha}^{(1)})$ and $p(\boldsymbol{O};\boldsymbol{\lambda}^{(2)},\boldsymbol{\alpha}^{(2)})$, the decision boundary that minimises the probability of error is given by Bayes' decision rule,

$$\frac{P(\omega_1|\boldsymbol{O})}{P(\omega_2|\boldsymbol{O})} = \frac{P(\omega_1)\tau(\boldsymbol{\lambda}^{(2)},\boldsymbol{\alpha}^{(2)})\bar{p}(\boldsymbol{O};\boldsymbol{\lambda}^{(1)},\boldsymbol{\alpha}^{(1)})}{P(\omega_2)\tau(\boldsymbol{\lambda}^{(1)},\boldsymbol{\alpha}^{(1)})\bar{p}(\boldsymbol{O};\boldsymbol{\lambda}^{(2)},\boldsymbol{\alpha}^{(2)})} \underset{\omega_2}{\overset{\omega_1}{\gtrless}} 1 \tag{6}$$

where $P(\omega_1)$ and $P(\omega_2)$ are class priors and $\bar{p}(\boldsymbol{O};\boldsymbol{\lambda}^{(\omega)},\boldsymbol{\alpha}^{(\omega)})$ denotes an unnormalised augmented model. Taking the natural logarithm of both sides and rearranging yields [7],

$$\left\langle \boldsymbol{w}, \boldsymbol{\phi}^{\mathsf{LL}}(\boldsymbol{O};\boldsymbol{\lambda})\right\rangle + b \underset{\omega_2}{\overset{\omega_1}{\gtrless}} 0\,; \quad \boldsymbol{w} = \begin{bmatrix} 1 \\ \boldsymbol{\alpha}^{(1)} \\ \boldsymbol{\alpha}^{(2)} \end{bmatrix}; \quad b = \ln\left[\frac{P(\omega_1)\tau(\boldsymbol{\lambda}^{(2)},\boldsymbol{\alpha}^{(2)})}{P(\omega_2)\tau(\boldsymbol{\lambda}^{(1)},\boldsymbol{\alpha}^{(1)})}\right] \tag{7}$$

It is clear from equation (7) that this represents a linear hyperplane[3] in the generative score-space $\boldsymbol{\phi}^{\mathsf{LL}}(\boldsymbol{O};\boldsymbol{\lambda})$. Maximum margin estimation of this hyperplane yields a discriminatively trained decision boundary suitable for classification. In addition, under some minor constraints (see [7]), values of $\boldsymbol{\alpha}$ can be extracted from $\boldsymbol{w}$, yielding an augmented model with maximum margin estimated augmented parameters $\boldsymbol{\alpha}$.

### Conditional Maximum Likelihood

In the previous section, MM estimation of augmented parameters was discussed for binary classification tasks. Alternatively, when multiclass classification is required, *conditional augmented (C-Aug) models* can be defined. Instead of modelling observation likelihoods, these directly model the posterior probability of the class labels, $\omega$,

$$P(\omega|\boldsymbol{O};\boldsymbol{\lambda},\boldsymbol{\alpha}) = \frac{1}{Z(\boldsymbol{\lambda},\boldsymbol{\alpha})}\exp\left(\boldsymbol{\alpha}^T\boldsymbol{T}(\omega,\boldsymbol{O};\boldsymbol{\lambda})\right) \tag{8}$$

where $\boldsymbol{\lambda} = \{\boldsymbol{\lambda}^{(\omega)}\}$, $\boldsymbol{\alpha} = \{\boldsymbol{\alpha}^{(\omega)}\}$, $\forall\omega \in \Omega$ (the set of all class labels), and $Z(\boldsymbol{\lambda},\boldsymbol{\alpha})$ is the normalisation term[4]. Sufficient statistics are similar to those for generative augmented models and are given by, $\boldsymbol{T}_{\omega'}(\omega,\boldsymbol{O};\boldsymbol{\lambda}) = \left\{\delta_{\omega=\omega'}\boldsymbol{T}(\boldsymbol{O};\boldsymbol{\lambda})\right\}_{\omega\in\Omega}$.

Although C-Aug models appear similar to the generative augmented models discussed in (3), they are in reality very different (the normalisation term is calculated as the expectation *over the class labels*). In particular, since the number of classes is typically small, $Z(\boldsymbol{\lambda},\boldsymbol{\alpha})$ can be calculated explicitly making direct training of model parameters possible. One such (discriminative) training criterion is conditional maximum likelihood (CML). For training examples $\boldsymbol{O}_i$ with labels $y_i \in \Omega$, $i \in \{1,\ldots,n\}$, the values of $\boldsymbol{\lambda}$ and $\boldsymbol{\alpha}$ that maximise the conditional likelihood of the class labels are given by,

$$\{\tilde{\boldsymbol{\lambda}},\tilde{\boldsymbol{\alpha}}\} = \arg\max_{\boldsymbol{\lambda},\boldsymbol{\alpha}} \sum_{i=1}^{n} \ln P(y_i|\boldsymbol{O}_i;\boldsymbol{\lambda},\boldsymbol{\alpha}) \tag{9}$$

Although this has no closed-form solution, it is a convex problem with a single global solution. Optimisation is therefore a simple matter of selecting an appropriate algorithm and waiting for convergence. In this paper, stochastic gradient descent is used. It is important to note that although the conditional distribution is always valid, the generative model associated with it (by Bayes' rule) may not be.

[3] Due to the definition of the bias $b$, there is some interaction between the base statistical model parameters $\boldsymbol{\lambda}$ and the augmented parameters $\boldsymbol{\alpha}$.

[4] For clarity, the normalisation term is denoted $Z(\cdot)$ instead of $\tau(\cdot)$ to emphasise that the expectation is calculated over the classes instead of over the observation sequences.

## Experimental Results

Preliminary results are presented for the TIMIT phone classification task. Base models (HMMs) with three hidden states and either ten or twenty mixture-components were trained. C-Aug feature-spaces were then constructed using derivatives of the base models with respect to the means, variances and component-priors. Neither feature whitening nor language models were used.

| Classifier | Criterion | | Components | |
|---|---|---|---|---|
| | $\lambda$ | $\alpha$ | 10 | 20 |
| HMM | ML | – | 29.4 | 27.3 |
| C-Aug | ML | CML | 25.6 | – |
| HMM | MMI | – | 25.3 | 24.8 |
| C-Aug | MMI | CML | 24.1 | – |

Table 1: Classification error (%) on the TIMIT core test set

As shown in Table 1, baseline ML estimated HMMs yield an error rate of 29.4%. Increasing the number of parameters (by adding components) yields a performance gain of 1.9%. However, if instead, parameters are added using the augmented model framework (with ML statistics), improved gains of 3.8% are achieved. Similar results are achieved for MMI estimation and statistics.

These results demonstrate how the additional flexibilty of augmented models allows them to outperform standard HMM baselines. However, despite good performance, there is evidence that the CML criterion caused overtraining: the 10-component C-Aug (MMI) model had a training error of 16.8%. It is therefore expected that performance will improve with regularisation. Additionally, in these experiments the state segmentation of examples was fixed by the base model; research is needed to evaluate segment optimisation techniques.

## Conclusion

In this paper, augmented models are proposed as a powerful form of statistical model that combine the benefits of generative and discriminative techniques. In particular, discriminate models are trained using statistics derived from generative models. A two-stage optimisation process allows augmented parameters to be estimated according to a convex objective function yielding a unique global solution. In preliminary experiments, the resulting half generative ($\lambda$), half discriminative ($\alpha$) C-Aug models outperformed both ML and MMI trained HMMs.

### References

[1] V.N. Vapnik. *Statistical Learning Theory*. John Wiley & Sons, 1998.

[2] J. Lafferty, A. McCallum, and F. Pereira. Condition random fields: Probabilistic models for segmenting and labeling sequence data. In *ICML*, pages 591–598, 2001.

[3] T. Jaakkola and D. Haussler. Exploiting generative models in disciminative classifiers. In *NIPS*, pages 487–493, 1999.

[4] N. Smith and M. Gales. Speech recognition using SVMs. In *Advances in Neural Information Processing Systems 14*, pages 1197–1204, 2002.

[5] S. Amari and S. Wu. *Methods of Information Geometry*. Oxford University Press, 2000.

[6] N.D. Smith. *Using Augmented Statistical Models and Score Spaces for Classification*. PhD thesis, University of Cambridge, September 2003.

[7] M.J.F. Gales and M.I. Layton. SVMs, score-spaces and maximum margin statistical models. In *Beyond HMM workshop, ATR*, 2004. http://mi.eng.cam.ac.uk/˜mjfg/BeyondHMM.pdf.

[8] M.I. Layton and M.J.F. Gales. Augmented statistical models for ASR. Technical Report CUED/F-INFENG/TR.540, Cambridge University Engineering Department, Nov 2005.