Sparse Maximum A Posteriori Adaptation

Peder A. Olsen, Jing Huang, Vaibhava Goel, Steven J. Rennie

Department of Speech and Language Algorithms,
T. J. Watson Research Center, IBM,
1101 Kitchawan rd, Yorktown Heights, NY 10598, USA
{pederao,jghg,vgoel,sjrennie}@us.ibm.com

***Abstract*—Maximum A Posteriori (MAP) adaptation is a powerful tool for building speaker specific acoustic models. Modern speech applications utilize acoustic models with millions of parameters, and serve millions of users. Storing an acoustic model for each user in such settings is costly. However, speaker specific acoustic models are generally similar to the acoustic model being adapted. By imposing sparseness constraints, we can save significantly on storage, and even improve the quality of the resulting speaker-dependent model. In this paper we utilize the ℓ_1 or ℓ_0 norm as a regularizer to induce sparsity. We show that we can obtain up to 95% sparsity with negligible loss in recognition accuracy, with both penalties. By removing small differences, which constitute "adaptation noise", sparse MAP is actually able to improve upon MAP adaptation. Sparse MAP reduces the MAP word error rate by 2% relative at 89% sparsity.**

I. Introduction

The state of the art estimation of an acoustic model depends on the type and amount of training data available. With large amounts of data (100-10,000 hours) the training is typically done with maximum likelihood or discriminative training techniques, [1]. The resulting acoustic models frequently consist of 10^5–10^6 gaussian components. Such large acoustic models cannot be estimated from small amounts of data. To adapt an acoustic model to a speaker given a limited amount of data (10 seconds to 10 minutes), linear regression methods are usually employed, [2], [3]. Multiple class-based linear transforms can be used as the amount of data grows, but ultimately the framework of maximum a posteriori (MAP) model adaptation, [4], is needed to reach the best performance for medium amounts of data (20 minutes to 10 hours).

MAP adaptation for Gaussian Mixture Models (GMMs), [4], utilizes conjugate Bayesian priors for the gaussian components (Dirichlet for mixture weights, Normal distribution for mean and Wishart for covariances) to re-estimate the parameters of the acoustic model, starting from a speaker independent or canonical acoustic model. The number of parameters estimated can be very large compared to the amount of data, and although the Bayesian prior provides smoothing, small movements of model parameters can still constitute noise. In this paper we consider removing the small differences between the old acoustic model and the speaker specific acoustic model. We aim for two objectives: compressing the information we need to store and reducing the word error rate. Both objectives are met through the use of sparse differences.

To make the parameter differences sparse we employ sparsity inducing penalties, and in particular we employ the ℓ_q penalties. For ℓ_0 we employ the counting "norm"

$$\|\mathbf{x}\|_0 = \#\{i : x_i \neq 0\}. \tag{1}$$

and for $q = 1$ we the regular ℓ_1 norm: $\|\mathbf{x}\|_1 = \sum_i |x_i|$.

In general there are well known techniques to minimize smooth convex functions with an additional $\|\mathbf{x}\|_1$ term, [5]. The $\|\mathbf{x}\|_0$ term, in contrast, is not convex and convex problems with an additional ℓ_0 penalty are known in general to be NP-hard, [6]. However, the problems considered in this paper are tractable and have elementary analytic solutions. Consequently the resulting algorithms are very efficient and straightforward to implement.

A. Important Related Work

In [7], it was demonstrated that an exponential language model trained with an $\ell_2^2 + \ell_1$ penalty can predict its own test performance to a remarkable accuracy. This penalty is known as the elastic net penalty, [8]. This insight has led to better language models, such as Model M, [9]. In this work, our ultimate goal is to understand what the analogous statement should be for acoustic models, and indeed our best results are for exponential acoustic models with an approximate $\ell_2^2 + \ell_1$ penalty. However, at this point, we make no claims to performance prediction.

B. Choice of Parameterization

In this work we consider sparse regularizers for the following parameterizations of a gaussian:

- **Moment Variables:** This is the usual representation with the parameters being $\boldsymbol{\xi} = \left(\begin{smallmatrix}\mu \\ v\end{smallmatrix}\right)$. We refer to the collection of variables for all the mixtures as $\boldsymbol{\Xi} = \{\omega_g, \boldsymbol{\mu}_g, \mathbf{v}_g\}_{g=1}^G$, where ω_g are the mixture weights, and G is the total number of gaussians in the acoustic model.
- **Exponential Family Variables:** Here we use the exponential family representation $\boldsymbol{\theta} = \left(\begin{smallmatrix}\frac{\mu}{v} \\ -\frac{1}{2v}\end{smallmatrix}\right) = \left(\begin{smallmatrix}\psi \\ -\frac{1}{2}p\end{smallmatrix}\right)$. We will say more about this later. This representation is especially efficient for likelihood computations. For this parameterization we refer to the collection of variables as $\boldsymbol{\Theta} = \{\omega_g, \boldsymbol{\psi}_g, \mathbf{p}_g\}_{g=1}^G$.

C. The Normal Distribution as an Exponential Family

We can write the one-dimensional gaussian as an exponential family as follows:

$$\mathcal{N}(x; \mu, v) = \frac{e^{-\frac{(x-\mu)^2}{2v}}}{\sqrt{2\pi v}} = \frac{e^{\boldsymbol{\theta}^\top \boldsymbol{\phi}(\mathbf{x})}}{Z(\boldsymbol{\theta})}. \tag{2}$$

With the features $\boldsymbol{\phi}$ and parameters $\boldsymbol{\theta}$ chosen as follows

$$\boldsymbol{\phi}(x) = \begin{pmatrix} x \\ x^2 \end{pmatrix} \qquad \boldsymbol{\theta} = \begin{pmatrix} \frac{\mu}{v} \\ -\frac{1}{2v} \end{pmatrix} = \begin{pmatrix} \psi \\ -\frac{1}{2}p \end{pmatrix}, \tag{3}$$

we get the following log-partition function

$$\log Z(\boldsymbol{\theta}) = \frac{1}{2}\left(\log(2\pi) - \log(p) + \frac{\psi^2}{p}\right). \tag{4}$$

In the exponential family formulation the maximum log likelihood objective function has the simple form

$$L(\boldsymbol{\theta}) = \mathbf{s}^\top \boldsymbol{\theta} - \log Z(\boldsymbol{\theta}) \tag{5}$$

where $\mathbf{s}$ is the empirical expected value of the sufficient statistic. The Kullback-Leibler (KL) divergence between two one-dimensional normal distributions will be needed later. It is given by the following formulas:

$$\begin{aligned} D(f\|g) &= \int f(x)\log\left(\frac{f(x)}{g(x)}\right)\mathrm{d}x & (6)\\ &= \log\frac{Z(\boldsymbol{\theta}_g)}{Z(\boldsymbol{\theta}_f)} + (\boldsymbol{\theta}_f - \boldsymbol{\theta}_g)^\top \mathrm{E}_{\boldsymbol{\theta}_f}[\boldsymbol{\phi}(x)] & (7)\\ &= \frac{1}{2}\left(\frac{v_f}{v_g} - 1 - \log\left(\frac{v_f}{v_g}\right) + \frac{(\mu_f - \mu_g)^2}{v_g}\right) & (8) \end{aligned}$$

The expected value of the features in (7) is given by

$$\mathrm{E}_{\boldsymbol{\theta}}[\boldsymbol{\phi}(x)] = \begin{pmatrix} \mu \\ v + \mu^2 \end{pmatrix} = \begin{pmatrix} \frac{\psi}{p} \\ \frac{\psi^2 + p}{p^2} \end{pmatrix}. \tag{9}$$

II. Review of MAP adaptation

Let $\mathcal{H} = \{\boldsymbol{\pi}, \mathbf{A}, \boldsymbol{\Xi}\}$ be a Hidden Markov Model (HMM), where $\boldsymbol{\pi}$ is the initial state distribution, $\mathbf{A}$ is the transition matrix and $\boldsymbol{\Xi}$ is the acoustic model $\boldsymbol{\Xi} = \{\omega_g, \boldsymbol{\mu}_g, \mathbf{v}_g\}_{g=1}^G$, where G is the total number of gaussians. The likelihood of the training data can then be written

$$\mathrm{P}(\mathbf{X}|\mathcal{H}) = \sum_{\boldsymbol{\sigma}} \pi_{\sigma_0} \prod_t a_{\sigma_{t-1}\sigma_t} \sum_{g\in\mathcal{G}_{\sigma_t}} \omega_g \mathcal{N}(\mathbf{x}_t; \boldsymbol{\mu}_g, \mathbf{v}_g), \tag{10}$$

where the outer sum is over all possible state sequences $\boldsymbol{\sigma}$, and the inner sum is over the mixture components corresponding to the state σ_t. For the acoustic model parameters we provide the following prior distribution

$$\mathrm{P}(\boldsymbol{\Xi}) = \prod_{g=1}^{G} P(\boldsymbol{\mu}_g, \mathbf{v}_g|\boldsymbol{\mu}_g^{\mathrm{old}}, \mathbf{v}_g^{\mathrm{old}}, \tau_\mu, \tau_v)\mathrm{P}(\omega_g|\omega_g^{\mathrm{old}}, \tau_\omega), \tag{11}$$

where $\tau_\mu, \tau_v, \tau_\omega$ are hyper-parameters that control the strength of the prior, and $\boldsymbol{\mu}_g^{\mathrm{old}}, \mathbf{v}_g^{\mathrm{old}}, \omega_g^{\mathrm{old}}$ are hyper-parameters inherited from the base acoustic model. We will assume that the priors for the remaining parameters are uniform, and therefore $\mathrm{P}(\mathcal{H}) \propto \mathrm{P}(\boldsymbol{\Xi})$. To estimate the acoustic model parameters we maximize the Bayesian likelihood

$$\mathrm{P}(\mathbf{X}|\mathcal{H})\mathrm{P}(\mathcal{H}). \tag{12}$$

Thus the Bayesian log likelihood is $\log \mathrm{P}(\mathbf{X}|\mathcal{H}) + \log \mathrm{P}(\boldsymbol{\Xi}) + \mathrm{const}$. Following [4] we use the Expectation Maximization (EM) framework to formulate an auxiliary function

$$\begin{aligned} Q(\boldsymbol{\Xi}^{(k)}, \boldsymbol{\Xi}^{(k-1)}) &= \sum_{gt} \gamma_g(\mathbf{x}_t) \log \frac{\omega_g^{(k)} \mathcal{N}(\mathbf{x}_t; \boldsymbol{\mu}_g^{(k)}, \mathbf{v}_g^{(k)})}{\omega_g^{(k-1)} \mathcal{N}(\mathbf{x}_t; \boldsymbol{\mu}_g^{(k-1)}, \mathbf{v}_g^{(k-1)})} \\ &\quad + \log \frac{\mathrm{P}(\boldsymbol{\Xi}^{(k)})}{\mathrm{P}(\boldsymbol{\Xi}^{(k-1)})}. \end{aligned} \tag{13}$$

Here we have used $\gamma_g(\mathbf{x}_t) = \mathrm{P}(g|\boldsymbol{\Xi}^{(k-1)}, \mathbf{x}_t)$ for the gaussian posterior at time t. This auxilliary function is a lower bound on the Bayesian log likelihood ratio, which can be maximized with respect to $\boldsymbol{\Xi}^{(k)}$. Dropping terms that depend only on $\boldsymbol{\Xi}^{(k-1)}$, leads to $Q(\boldsymbol{\Xi}^{(k)}) = \sum_g L(\omega_g^{(k)}, \boldsymbol{\mu}_g^{(k)}, \mathbf{v}_g^{(k)}) + R(\omega_g^{(k)}, \boldsymbol{\mu}_g^{(k)}, \mathbf{v}_g^{(k)}) + \mathrm{const}$, where

$$L(\omega_g, \boldsymbol{\mu}_g, \mathbf{v}_g) = \sum_t \gamma_g(\mathbf{x}_t) \log\left(\omega_g \mathcal{N}(\mathbf{x}_t; \boldsymbol{\mu}_g, \mathbf{v}_g)\right) \tag{14}$$

and

$$\begin{aligned} R(\omega_g, \boldsymbol{\mu}_g, \mathbf{v}_g) &= \log P(\boldsymbol{\mu}_g, \mathbf{v}_g|\boldsymbol{\mu}_g^{\mathrm{old}}, \mathbf{v}_g^{\mathrm{old}}, \tau_\mu, \tau_v) \\ &\quad + \log \mathrm{P}(\omega_g|\omega_g^{\mathrm{old}}, \tau_\omega). \end{aligned} \tag{15}$$

In our work we use a Bayesian prior that is a special case of the more general Bayesian prior discussed in [4]. We use the I-smoothing framework, [1], with the simple regularizer, [10],

$$R(\omega_g, \boldsymbol{\mu}_g, \mathbf{v}_g) = -\tau D\left(\mathcal{N}(\boldsymbol{\mu}_g^{\mathrm{old}}, \mathbf{v}_g^{\mathrm{old}}) \| \mathcal{N}(\boldsymbol{\mu}_g, \mathbf{v}_g)\right), \tag{16}$$

where $\tau_\mu = \tau_v = \tau$ and the prior on w is uniform. The corresponding log likelihood can be written

$$\begin{aligned} L(\omega_g, \boldsymbol{\mu}_g, \mathbf{v}_g) &= T_g\Bigg(\log\omega_g + \sum_{i=1}^{d} \frac{\mu_{gi} s_{1gi}}{v_{gi}} \\ &\quad - \frac{1}{2}\Big(\frac{s_{2gi} + \mu_{gi}^2}{v_{gi}} + \log(2\pi v_{gi})\Big)\Bigg). \end{aligned}$$

Here $T_g = \sum_t \gamma_g(\mathbf{x}_t)$ is the posterior count, $s_{1gi} = \frac{1}{T_g}\sum_t \gamma_g(\mathbf{x}_t) x_{ti}$, and $s_{2gi} = \frac{1}{T_g}\sum_t \gamma_g(\mathbf{x}_t) x_{ti}^2$.

Now, that we have formulated the MAP auxiliary objective function, we will make some simplifications. Since the auxiliary objective function decouples across gaussians and dimensions we simply drop the indices g and i. Also, we are not particularly interested in sparsity on ω_g, so we will ignore this variable altogether. Thus,

$$L(\boldsymbol{\mu}, \mathbf{v}; s_1, s_2) = \frac{T\mu s_1}{v} - \frac{T}{2}\left(\frac{s_2 + \mu^2}{v} + \log(2\pi v)\right).$$

One more simplification is useful, to use the exponential family representation:

$$L(\boldsymbol{\theta}; \mathbf{s}) = \mathbf{s}^\top \boldsymbol{\theta} - \log Z(\boldsymbol{\theta}), \tag{17}$$

where $\mathbf{s} = (s_1, s_2)^\top$. In the exponential family representation the penalty term can be written

$$\begin{aligned} R(\boldsymbol{\theta}) &= -\tau D\left(\mathcal{N}(\mu^{\mathrm{old}}, v^{\mathrm{old}}) \| \mathcal{N}(\mu, v)\right) \\ &= -\tau \log Z(\boldsymbol{\theta}) + \tau \boldsymbol{\theta}^\top \mathrm{E}_{\boldsymbol{\theta}^{\mathrm{old}}}[\boldsymbol{\phi}(x)] + \mathrm{const}. \end{aligned} \tag{18}$$

Defining $\mathbf{s}^{\text{old}} = \mathrm{E}_{\boldsymbol{\theta}^{\text{old}}}[\boldsymbol{\phi}(x)]$ yields the Bayesian auxiliary objective:

$$\begin{aligned} Q(\boldsymbol{\theta}) &= L(\boldsymbol{\theta};\mathbf{s}) + R(\boldsymbol{\theta}) \\ &= \boldsymbol{\theta}^\top(T\mathbf{s}+\tau\mathbf{s}^{\text{old}}) - (T+\tau)\log Z(\boldsymbol{\theta}) + \text{const} \\ &= (T+\tau)L(\boldsymbol{\theta};\mathbf{s}^{\text{MAP}}) + \text{const}, \end{aligned} \tag{19}$$

where $\mathbf{s}^{\text{MAP}} = \frac{T\mathbf{s}+\tau\mathbf{s}^{\text{old}}}{T+\tau}$. In other words the auxiliary MAP objective function is of the same form as the maximum likelihood objective function, but with the sufficient statistics replaced by smooth statistics $\mathbf{s}^{\text{MAP}}$. Note that it is possible to apply similar regularization to discriminative objective functions, [11]. Discriminative MAP is challenging because discriminative objectives are difficult to optimize even when data is plentiful.

III. Sparse Constraints

A. Restricting parameter movement

Consider the following constrained MAP problem where we only allow N parameters to change:

$$\max_{\boldsymbol{\theta}} \qquad \sum_{gi}(T_g+\tau)L(\boldsymbol{\theta}_{gi};\mathbf{s}^{\text{MAP}}_{gi}) \tag{20}$$

$$\text{subject to:} \qquad N = \sum_{gi}\|\boldsymbol{\theta}_{gi}-\boldsymbol{\theta}^{\text{old}}_{gi}\|_0, \tag{21}$$

where $\|\boldsymbol{\theta}\|_0 = \#\{j : \theta_j \neq 0\}$ is the counting norm, and $\boldsymbol{\theta}_{gi} = (\psi_{gi}, -p_{gi}/2)^\top$ is the parameter vector for dimension i of gaussian g. This problem can be solved exactly, because the objective function and the constraint are direct sums over g and i. The same result holds for the moment variable case. For each term there are four possible scenarios for $\|\boldsymbol{\theta}_{gi}-\boldsymbol{\theta}^{\text{old}}_{gi}\|_0 \in \{0,1,2\}$: (1) $\psi_{gi}=\psi^{\text{old}}_{gi}, p_{gi}=p^{\text{old}}_{gi}$, (2) $\psi_{gi}\neq\psi^{\text{old}}_{gi}, p_{gi}=p^{\text{old}}_{gi}$, (3) $\psi_{gi}=\psi^{\text{old}}_{gi}, p_{gi}\neq p^{\text{old}}_{gi}$ or (4) $\psi_{gi}\neq\psi^{\text{old}}_{gi}, p_{gi}\neq p^{\text{old}}_{gi}$. For each scenario $L(\boldsymbol{\theta}_{gi};\mathbf{s}^{\text{MAP}}_{gi})$ can be maximized analytically. From these, it is trivial to maximize $L(\boldsymbol{\theta}_{gi};\mathbf{s}^{\text{MAP}}_{gi})$ for 0, 1, or 2 parameter changes. To solve the presented MAP constraint problem for $N=1$, it is clear that the single parameter change that causes the largest increase in the objective function is optimal. The optimal solution for general N can be built up by greedily selecting the next parameter change that most increases the objective function.
Alternatively we can consider maximizing the Lagrangian

$$\begin{aligned} \mathcal{L}(\boldsymbol{\Theta};\lambda) &= \sum_{gi}(T_g+\tau)L(\boldsymbol{\theta}_{gi};\mathbf{s}^{\text{MAP}}_{gi}) \\ &\quad -\lambda\Big(\sum_{gi}\|\boldsymbol{\theta}_{gi}-\boldsymbol{\theta}^{\text{old}}_{gi}\|_0 - N\Big) \end{aligned}$$

with respect to $\boldsymbol{\Theta}$ for each λ. This dual problem is not equivalent to the constrained MAP problem stated above: in general the duality gap may be non-zero. For fixed λ the problem *fully* decouples across g, i and we solve each sub-problem

$$\max_{\boldsymbol{\theta}_{gi}}(T_g+\tau)L(\boldsymbol{\theta}_{gi};\mathbf{s}^{\text{MAP}}_{gi}) - \lambda\|\boldsymbol{\theta}_{gi}-\boldsymbol{\theta}^{\text{old}}_{gi}\|_0 \tag{22}$$

independently. A bisection search can locate the value of λ that gives approximately N parameter changes (exactly if there is no duality gap). This second method is efficient and simple to implement. and it motivates the introduction of the sparsity promoting $\|\cdot\|_0$ penalty as well as $\|\cdot\|_1$.

B. Sparsity Promoting Regularizers

More generally, we can restrict parameter changes by imposing a sparsity promoting regularizer of the form:

$$R(\boldsymbol{\Theta}) = \sum_{gi} -\tau D(\boldsymbol{\theta}^{\text{old}}_{gi}\|\boldsymbol{\theta}_{gi}) - \lambda\|\boldsymbol{\theta}_{gi}-\boldsymbol{\theta}^{\text{old}}_{gi}\|_q$$

for $0 \leq q \leq 1$. Here we consider two cases: $q=0$ and $q=1$, which can be solved analytically. Both cases can be solved by partitioning the domain into pieces where the function is continuously differentiable. On each piece the local maxima can be found analytically. The same holds for representations in terms of moment variables.
The total auxiliary penalized Bayesian log likelihood can then be written

$$Q(\boldsymbol{\Theta}) = \sum_{gi}(T_g+\tau)L(\boldsymbol{\theta}_{gi};\mathbf{s}^{\text{MAP}}_{gi}) - \lambda\|\boldsymbol{\theta}_{gi}-\boldsymbol{\theta}^{\text{old}}_{gi}\|_q \tag{23}$$

for the exponential family representation and

$$Q(\boldsymbol{\Xi}) = \sum_{gi}(T_g+\tau)L(\boldsymbol{\xi}_{gi};\mathbf{s}^{\text{MAP}}_{gi}) - \lambda\|\boldsymbol{\xi}_{gi}-\boldsymbol{\xi}^{\text{old}}_{gi}\|_q \tag{24}$$

for the moment variable case.

C. Optimization

In considering maximizing the auxiliary objective function we find it useful to instead minimize the function

$$F(\boldsymbol{\theta}_{gi};\mathbf{s}^{\text{MAP}}_{gi}) = -2L(\boldsymbol{\theta}_{gi};\mathbf{s}^{\text{MAP}}_{gi}) + \alpha\|\boldsymbol{\theta}_{gi}-\boldsymbol{\theta}^{\text{old}}_{gi}\|_q \tag{25}$$

for each g, i for the exponential family representation and

$$F(\boldsymbol{\xi}_{gi};\mathbf{s}^{\text{MAP}}_{gi}) = -2L(\boldsymbol{\xi}_{gi};\mathbf{s}^{\text{MAP}}_{gi}) + \alpha\|\boldsymbol{\xi}_{gi}-\boldsymbol{\xi}^{\text{old}}_{gi}\|_q \tag{26}$$

for each g, i for the moment variable case. Here $\alpha = \frac{2\lambda}{T_g+\tau}$. The function F can be maximized analytically. For the ℓ_0 constraint we must consider 4 different cases, and for the ℓ_1 constraint we must consider 9 different cases. For example for $\boldsymbol{\xi}_{gi} = (\mu_{gi}, v_{gi})$, μ_{gi} can be less than, equal to and greater than μ^{old}_{gi} and similarly for v_{gi}. The ℓ_1 penalty is continuously differentiable on each of these pieces and the solution for each case can be readily computed. For the ℓ_0 constraint with moment variables the solution is given by Algorithm 1, for exponential family variables by Algorithm 2, for ℓ_1 with exponential family variables by Algorithm 4. Finally for the ℓ_1 case with moment variables we give the algorithm to locate all the local minima of F by Algorithm 3.

input : Statistics $\mathbf{s}^{\text{MAP}}$, T, model μ^{old}, v^{old} and τ, λ
output: μ, v

Compute $\alpha = 2\lambda/(T+\tau)$
Compute candidate answers: μ^{MAP}, v^{MAP} and v^*
$\mu^{\text{MAP}} \longleftarrow s_1^{\text{MAP}}$
$v^{\text{MAP}} \longleftarrow s_2^{\text{MAP}} - (s_1^{\text{MAP}})^2$
$v^* \longleftarrow v^{\text{MAP}} + (\mu^{\text{old}} - \mu^{\text{MAP}})^2$

Compute Lagrangian objectives:
$F_1 \longleftarrow \frac{(\mu^{\text{old}} - \mu^{\text{MAP}})^2 + v^{\text{MAP}}}{v^{\text{old}}} + \log(2\pi v^{\text{old}})$
$F_2 \longleftarrow \left(\frac{v^{\text{MAP}}}{v^{\text{old}}} + \log(2\pi v^{\text{old}})\right) + \alpha$
$F_3 \longleftarrow \log(2\pi e v^*) + \alpha$
$F_4 \longleftarrow \log(2\pi e v^{\text{MAP}}) + 2\alpha$

if F_1 *smallest* **then** $(\mu, v) \longleftarrow (\mu^{\text{old}}, v^{\text{old}})$
if F_2 *smallest* **then** $(\mu, v) \longleftarrow (\mu^{\text{MAP}}, v^{\text{old}})$
if F_3 *smallest* **then** $(\mu, v) \longleftarrow (\mu^{\text{MAP}}, v^*)$
if F_4 *smallest* **then** $(\mu, v) \longleftarrow (\mu^{\text{MAP}}, v^{\text{MAP}})$

Algorithm 1: Solution to sparse MAP adaptation with ℓ_0 penalty on the moment variables.

input : Statistics $\mathbf{s}^{\text{MAP}}$, T, model μ^{old}, v^{old} and τ, λ
output: ψ, p

Compute $\alpha = 2\lambda/(T+\tau)$
Compute candidate answers: μ^{MAP}, v^{MAP} and p^*
$\mu^{\text{MAP}} \longleftarrow s_1^{\text{MAP}}$
$v^{\text{MAP}} \longleftarrow s_2^{\text{MAP}} - (s_1^{\text{MAP}})^2$
$p^* \longleftarrow \frac{1+\sqrt{1+(\psi^{\text{old}})^2 s_2^{\text{MAP}}}}{2 s_2^{\text{MAP}}}$

Compute Lagrangian objectives:
$F_1 \longleftarrow \frac{(\mu^{\text{old}} - \mu^{\text{MAP}})^2 + v^{\text{MAP}}}{v^{\text{old}}} + \log(2\pi v^{\text{old}})$
$F_2 \longleftarrow \left(\frac{v^{\text{MAP}}}{v^{\text{old}}} + \log(2\pi v^{\text{old}})\right) + \alpha$
$F_3 \longleftarrow -2\psi^{\text{old}}\mu^{\text{MAP}} + p^* s_2^{\text{MAP}} + \frac{(\psi^{\text{old}})^2}{p^*} + \log\left(\frac{2\pi}{p^*}\right) + \alpha$
$F_4 \longleftarrow \log(2\pi e v^{\text{MAP}}) + 2\alpha$

if F_1 *smallest* **then** $(\psi, p) \longleftarrow (\mu^{\text{old}}/v^{\text{old}}, 1/v^{\text{old}})$
if F_2 *smallest* **then** $(\psi, p) \longleftarrow (\mu^{\text{MAP}}/v^{\text{old}}, 1/v^{\text{old}})$
if F_3 *smallest* **then** $(\psi, p) \longleftarrow (\mu^{\text{old}}/v^{\text{old}}, p^*)$
if F_4 *smallest* **then** $(\psi, p) \longleftarrow (\mu^{\text{MAP}}/v^{\text{MAP}}, 1/v^{\text{MAP}})$

Algorithm 2: Solution to sparse MAP adaptation with ℓ_0 penalty on the exponential family variables.

IV. Experiments

A. Task Description

We used an internal US English speech recognition task for all experiments. The training set consists of 2000 hours of recordings. The test and adaptation sets were collected from the same set of 26 speakers. The enrollment data used for adaptation consists of 2.8 ± 1.5 hours of data per speaker with known transcripts. The test data has 7-20 minutes per speaker (52K words in total). Acoustic features were constructed from 12 dimensional Mel-frequency Cepstra coefficients and

input : Statistics $\mathbf{s}^{\text{MAP}}$, T, model μ^{old}, v^{old} and τ, λ
output: Candidate list C

Initialize: $C \leftarrow \emptyset$, $i \leftarrow 0$.
Compute $\alpha = 2\lambda/(T+\tau)$ and $v^* = (\mu^{\text{old}} - \mu^{\text{MAP}})^2 + v^{\text{MAP}}$.

if $|\mu^{\text{old}} - \mu^{\text{MAP}}| \leq \frac{\alpha}{2} v^{\text{old}}$ **and** $\left|1 - \frac{v^*}{v^{\text{old}}}\right| \leq \alpha v^{\text{old}}$ **then**
  $C[i] = (\mu^{\text{old}}, v^{\text{old}}, F(\mu^{\text{old}}, v^{\text{old}}))$, $i \leftarrow i+1$
end
if $4\alpha v^* \leq 1$ **then**
  $v_{el} = \frac{1-\sqrt{1-4\alpha v^*}}{2\alpha}$, $\mu_{el} = \mu^{\text{old}}$
  if $v_{el} < v^{\text{old}}$ **and** $|\mu^{\text{old}} - \mu^{\text{MAP}}| \leq \frac{\alpha}{2} v_{el}$ **then**
    $C[i] = (\mu_{el}, v_{el}, F(\mu_{el}, v_{el}))$, $i \leftarrow i+1$
  end
end

$v_{eg} = \frac{\sqrt{1+4\alpha v^*}-1}{2\alpha}$, $\mu_{eg} = \mu^{\text{old}}$
if $v_{eg} > v^{\text{old}}$ **and** $|\mu^{\text{old}} - \mu^{\text{MAP}}| \leq \frac{\alpha}{2} v_{eg}$ **then**
  $C[i] = (\mu_{eg}, v_{eg}, F(\mu_{eg}, v_{eg}))$, $i \leftarrow i+1$
end
if $\left|\frac{1}{v^{\text{old}}}\left(1 - \frac{v^{\text{MAP}}}{v^{\text{old}}}\right) - \frac{\alpha^2}{4}\right| \leq \alpha$ **then**
  $\mu_{le} = \mu^{\text{MAP}} + \frac{\alpha}{2} v^{\text{old}}$, $v_{le} = v^{\text{old}}$
  if $\mu_{le} < \mu^{\text{old}}$ **then**
    $C[i] = (\mu_{le}, v_{le}, F(\mu_{le}, v_{le}))$, $i \leftarrow i+1$
  end
  $\mu_{ge} = \mu^{\text{MAP}} - \frac{\alpha}{2} v^{\text{old}}$, $v_{ge} = v^{\text{old}}$
  if $\mu_{ge} > \mu^{\text{old}}$ **then**
    $C[i] = (\mu_{ge}, v_{ge}, F(\mu_{ge}, v_{ge}))$, $i \leftarrow i+1$
  end
end
if $1 \geq \alpha(4+\alpha) v^{\text{MAP}}$ **then**
  $v_{ll} = v_{gl} = \frac{1-\sqrt{1-\alpha(4+\alpha)v^{\text{MAP}}}}{\alpha(4+\alpha)/2}$
  $\mu_{ll} = \mu^{\text{MAP}} + \frac{\alpha}{2} v_{ll}$, $\mu_{gl} = \mu^{\text{MAP}} - \frac{\alpha}{2} v_{gl}$
  if $\mu_{ll} < \mu^{\text{old}}$ **and** $v_{ll} < v^{\text{old}}$ **then**
    $C[i] = (\mu_{ll}, v_{ll}, F(\mu_{ll}, v_{ll}))$, $i \leftarrow i+1$
  end
  if $\mu_{gl} > \mu^{\text{old}}$ **and** $v_{gl} < v^{\text{old}}$ **then**
    $C[i] = (\mu_{gl}, v_{gl}, F(\mu_{gl}, v_{gl}))$, $i \leftarrow i+1$
  end
end
if $\alpha \leq 4$ **or** *(*$\alpha > 4$ **and** $v^{\text{MAP}}\alpha(\alpha-4) \leq 1$*)* **then**
  $$v_{lg} = v_{gg} = \begin{cases} \frac{\sqrt{1+v^{\text{MAP}}\alpha(4-\alpha)}-1}{\alpha/2(4-\alpha)} & \text{if} \quad \alpha < 4 \\ v^{\text{MAP}} & \text{if} \quad \alpha = 4 \\ \frac{1-\sqrt{1-v^{\text{MAP}}\alpha(\alpha-4)}}{\alpha/2(\alpha-4)} & \text{if} \quad \alpha > 4. \end{cases}$$
  $\mu_{gg} = \mu^{\text{MAP}} - \frac{\alpha}{2} v_{gg}$, $\mu_{lg} = \mu^{\text{MAP}} + \frac{\alpha}{2} v_{1g}$
  if $\mu_{lg} < \mu^{\text{old}}$ **and** $v_{lg} > v^{\text{old}}$ **then**
    $C[i] = (\mu_{lg}, v_{lg}, F(\mu_{lg}, v_{lg}))$, $i \leftarrow i+1$
  end
  if $\mu_{gg} > \mu^{\text{old}}$ **and** $v_{gg} > v^{\text{old}}$ **then**
    $C[i] = (\mu_{gg}, v_{gg}, F(\mu_{gg}, v_{gg}))$, $i \leftarrow i+1$
  end
end

Algorithm 3: Candidate local minima for sparse ℓ_1 penalty in the moment variable case.

input : Statistics $\mathbf{s}^{\text{MAP}}$, T, model μ^{old}, v^{old} and τ, λ
output: Parameters (ψ, p) attaining global minimum

Compute $\alpha = 2\lambda/(T+\tau)$ and $\psi^{\text{old}} = \mu^{\text{old}}/v^{\text{old}}$, $p^{\text{old}} = 1/v^{\text{old}}$
Compute MAP solution $\mu^{\text{MAP}} = s_1^{\text{MAP}}$,$v^{\text{MAP}} = s_2^{\text{MAP}} - (s_1^{\text{MAP}})^2$, $\psi^{\text{MAP}} = \mu^{\text{MAP}}/v^{\text{MAP}}$, $p^{\text{MAP}} = 1/v^{\text{MAP}}$

Compute $p_a = \alpha p^{\text{MAP}}/2$, $p_{aa} = \alpha p_a/2$, $\psi_a = \alpha\psi^{\text{MAP}}/2$
for $\epsilon \in \{-1, 1\}$ **do**
  for $\delta \in \{-1, 1\}$ **do**
    $\Delta = 1 + 2\epsilon\psi_a + \delta p_a - p_{aa}$
    $\psi = \frac{\psi^{\text{MAP}} - \epsilon p_a}{\Delta}$, $p = \frac{p^{\text{MAP}}}{\Delta}$
    if *(*$\epsilon\psi > \epsilon\psi^{\text{old}}$ **and** $\delta p < \delta p^{\text{old}}$ **and** $p > 0$*)* **then**
      return (ψ, p)
    end
  end
end

$p = p^{\text{old}}$
for $\epsilon \in \{-1, 1\}$ **do**
  $\psi = \frac{p^{\text{old}}}{p^{\text{MAP}}}\left(\psi^{\text{MAP}} - \epsilon p_a\right)$
  $\eta = (-s_2^{\text{MAP}} + 1/p + \psi^2/p^2)/2$
  $\nabla_{\text{left}} = \eta - \alpha/4$, $\nabla_{\text{right}} = \eta + \alpha/4$
  if *(*$\epsilon\psi > \epsilon\psi^{\text{old}}$ **and** $\nabla_{\text{left}} \leq 0$ **and** $\nabla_{\text{right}} \geq 0$*)* **then**
    return (ψ, p)
  end
end

$\psi = \psi^{\text{old}}$
for $\delta \in \{-1, 1\}$ **do**
  $\zeta = 2\left((\psi^{\text{MAP}})^2 + p^{\text{MAP}} + \delta p_a p^{\text{MAP}}\right)$
  if $\zeta > 0$ **then**
    $p = \frac{(p^{\text{MAP}})^2 + p^{\text{MAP}}\sqrt{(p^{\text{MAP}})^2 + 2\zeta\psi^2}}{\zeta}$
    $\eta = \mu^{\text{MAP}} - \psi/p$
    $\nabla_{\text{left}} = \eta - \alpha/2$, $\nabla_{\text{right}} = \eta + \alpha/2$
    if *(*$\delta p > \delta p^{\text{old}}$ **and** $\nabla_{\text{left}} \leq 0$ **and** $\nabla_{\text{right}} \geq 0$ **and** $p > 0$*)* **then**
      return (ψ, p)
    end
  end
end

return $(\psi^{\text{old}}, p^{\text{old}})$

Algorithm 4: Global minimum for sparse ℓ_1 penalty for exponential family variables.

their first, second and third derivative, followed by a Linear Discriminant Analysis (LDA) projection.

The acoustic model had 5000 HMM states and 200,000 gaussian components and was trained using feature space minimum phone error rate (fMPE) and discriminative Minimum Phone Error (MPE), as described in [12], [13]. A Constrained Maximum Likelihood Linear Regression (CMLLR) transform, [3], was learned for each speaker. In the transformed feature space, we then trained a "canonical acoustic model" using speaker adaptive training (SAT), [14].

B. Baseline MAP results

Even MAP adaptation can suffer from lack of data, and our default MAP adaptation code contains a count threshold (`cnt`) to determine whether the mean, variance and mixture weights should be updated. The default posterior count threshold in the MAP training code was `cnt` $= 10$, and although this choice had been seen to be effective on other tasks with smaller data amounts, it was apparently not effective here. Table I shows the baseline word error rates (WER) with and without MAP adaptation and with maximum likelihood (ML) estimation on only the adaptation data. ML estimation suffers greatly from lack of data and this hurts the performance significantly.

We experimented with different values of τ to verify that the best performance that could be reached was 10.9%. Table II shows that the best performance is attained for a wide range of the parameter τ (100-500).

Finally, to complete the baseline experiments, we investigated directly thresholding parameter changes to obtain sparsity. In other words, we only keep parameters for which $|\mu_i - \mu_i^{\text{old}}| \geq \epsilon$ or $|v_i - v_i^{\text{old}}| \geq \epsilon$, where ϵ is a parameter that controls the amount of sparsity. Table III shows results for a number of choices of ϵ used to threshold the moment variables. The results show that the direct method is a poor control of sparsity. At 74% sparsity, the performance is 0.7% worse than the baseline. The performance of direct thresholding could be improved by using separate thresholds for means and variances, and by making the thresholds dimension dependent. However, the sparse regularization methods presented here are more flexible, principled, and directly maximize the likelihood of the data subject to sparseness constraints.

TABLE I
WORD ERROR RATES FOR BASELINE SYSTEMS

System	WER
FMPE+CMLLR+SAT	13.2%
MAP, `cnt` $= 10$, $\tau = 500$	11.1%
MAP, `cnt` $= 0$**,** $\tau = 500$	**10.9%**
ML (adaptation data only)	23.8%

TABLE II
MAP PERFORMANCE FOR VARIOUS CHOICES OF τ

τ	0	10	50	**100**	**250**	**500**	1000
WER(%)	23.8	12.4	11.3	**10.9**	**10.9**	**10.9**	11.2

C. Sparse Penalties as Bayesian Priors

Let's consider the sparse penalties on their own as a regularizer, i.e. we choose $\tau = 0$. The value of λ that minimizes

TABLE III
WORD ERROR RATES FOR VARIOUS DIRECT SPARSITY THRESHOLDS.

ϵ	WER	Sparsity
0	10.9%	0%
1	11.0%	43%
10	11.6%	74%
100	12.8%	95%
10^6	13.2%	100%

TABLE IV
MINIMUM WORD ERROR RATES FOR ℓ_0 AND ℓ_1 SPARSITY PENALTIES WITH THE MOMENT VARIABLE (MV) AND EXPONENTIAL FAMILY VARIABLE (EFV) REPRESENTATIONS.

Penalty	λ	WER	sparsity
ℓ_0, MV	6	11.5%	97.7%
ℓ_0, EFV	6	11.6%	97.7%
ℓ_1, MV	0.1	11.0%	86.9%
ℓ_1, EFV	8500	**11.0%**	**88.0%**

TABLE V
MINIMUM WORD ERROR RATES FOR ℓ_0 AND ℓ_1 SPARSITY PENALTIES COMBINED WITH MAP.

Penalty	τ	λ	WER	sparsity
ℓ_0 MV	100	0.5	10.9%	86.0%
ℓ_0 EFV	100	0.5	10.9%	86.1%
ℓ_1 MV	50	0.1	10.8%	86.9%
ℓ_1 EFV	50	10000	**10.7%**	**89.4%**

the word error rate is shown in Table IV for the four cases discussed in this paper. The penalty with best performance was the ℓ_1 penalty with exponential family variables. The choice of parameterization seemed to have little effect. For the ℓ_0 penalty, the minimum word error rate (11.5%) was reached at 97.7% sparsity, which speaks to its power to induce sparseness. Although, the word error rate was higher than the ℓ_1 penalty results (11.0%), it is remarkable considering that this penalty essentially chooses between the parameters of the updated ML model (WER 23.8%) and the baseline model (WER 13.2%). Neither penalty was, on its own, competitive with the MAP penalty.

D. Sparse MAP Adaptation

If sparse regularizers are used together with MAP, we see only small gains in word error rate. Table V shows the minimum word error rate attained for each penalty. Only the ℓ_1 constrained penalties yielded lower word error rates than MAP adaptation.

Finally, we considered minimizing the word error rate at a 95% sparsity level. Table VI shows the results for the different penalties and the associated values for τ and λ. At this level of sparsity three of the penalties achieved an 11.0% word error rate. A 95% sparsity level implies that the speaker dependent acoustic models could potentially be compressed by a factor of close to 20, if the actual locations of the changed parameters can be stored efficiently. An examination of the sparsity structure revealed that 75% of the gaussians remained unchanged. The entropy of the binary mask for the remaining locations suggests that a Huffman coding, [15], would obtain a compression factor close to 20.

V. DISCUSSION

We have shown that through the use of sparse regularization, it is possible to obtain competitive adaptation performance by changing only a small fraction of the parameters of an acoustic model. This allows for the compression of speaker-dependent models: a capability that has important implications for systems with millions of users. We also observed some improvements in word error rate, but these gains are too small to be considered statistically significant. The work, as presented, has three main shortcomings: 1) it does not allow different rates of smoothing for means and variances (not updating the variance is known to be good for MAP adaptation with small data amounts), 2) it is not invariant to simultaneous scaling of the training and adaptation data, and 3), it has not combined the ℓ_0 and ℓ_1 regularizers. We plan to address these issues in future research.

TABLE VI
MINIMUM WORD ERROR RATES FOR ℓ_0 AND ℓ_1 SPARSITY PENALTIES COMBINED WITH MAP AT A 95% SPARSITY LEVEL.

Penalty	τ	λ	WER	sparsity
ℓ_0 MV	50	2.0	11.0%	95.1%
ℓ_0 EFV	50	2.0	11.0%	95.2%
ℓ_1 MV	50	0.325	11.3%	95.2%
ℓ_1 EFV	50	25000	**11.0%**	**95.2%**

REFERENCES

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

[2] M. Gales and P. Woodland, "Mean and variance adaptation within the MLLR framework," *Computer Speech and Language*, vol. 10, no. 4, pp. 249–264, 1996.

[3] M. Gales, "Maximum likelihood linear transformations for HMM-based speech recognition," *Computer Speech and Language*, vol. 12, pp. 75–98, 1998.

[4] J. L. Gauvain and C. H. Lee, "Maximum a posteriori estimation for multivariate gaussian mixture observations of Markov chains," *IEEE Transactions on Speech and Audio Processing*, vol. 2, no. 2, pp. 291–298, 1994.

[5] M. Zibulevsky and M. Elad, "L1-L2 optimization in signal and image processing," *IEEE Signal Processing Magazine*, vol. 27, no. 3, pp. 76–88, 2010.

[6] B. Natarajan, "Sparse approximate solutions to linear systems," *SIAM journal on computing*, vol. 24, p. 227, 1995.

[7] S. Chen, "Performance prediction for exponential language models," in *Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics*. Association for Computational Linguistics, 2009, pp. 450–458.

[8] H. Zou and T. Hastie, "Regularization and variable selection via the elastic net," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, vol. 67, no. 2, pp. 301–320, 2005.

[9] S. F. Chen and S. M. Chu, "Enhanced word classing for model M," in *Interspeech*, 2010.

[10] P. A. Olsen, V. Goel, and S. J. Rennie, "Discriminative training for full covariance models," in *ICASSP*, 2011, pp. 5312–5315.

[11] D. Povey, P. C. Woodland, and M. J. F. Gales, "Discriminative MAP for acoustic model adaptation," in *ICASSP*, vol. I, 2003, pp. 312–315.

[12] D. Povey, B. Kingsbury, L. Mangu, G. Saon, H. Soltau, and G. Zweig, "fMPE: Discriminatively trained features for speech recognition," in *Proc. ICASSP*, vol. 1, 2005, pp. 961–964.

[13] D. Povey and P. C. Woodland, "Minimum phone error and I-smoothing for improved discriminative training," in *ICASSP*, vol. I, 2002, pp. 105–108.

[14] T. Anastasakos, J. McDonough, R. Schwartz, and J. Makhoul, "A compact model for speaker-adaptive training," in *ICSLP*, vol. 2, 1996, pp. 1137–1140.

[15] D. Huffman, "A method for the construction of minimum-redundancy codes," *Proceedings of the IRE*, vol. 40, no. 9, pp. 1098–1101, 1952.