Shrinkage Model Adaptation in Automatic Speech Recognition

Jinyu Li[1], Yu Tsao[2] and Chin-Hui Lee[3]

[1]Microsoft Corporation, Redmond, WA. 98052 USA
[2]National Institute of Information and Communications Technology, Japan
[3]Georgia Institute of Technology, Atlanta, GA. 30332 USA
jinyli@microsoft.com, yu.tsao@nict.go.jp, chl@ece.gatech.edu

Abstract

We propose a parameter shrinkage adaptation framework to estimate models with only a limited set of adaptation data to improve accuracy for automatic speech recognition. By regularizing an objective function with a sum of parameter-wise power q constraint. For the first attempt, we formulate ridge maximum likelihood linear regression (MLLR) and ridge constraint MLLR (CMLLR) with an element-wise squared sum constraint to regularize the objective functions of the conventional MLLR and CMLLR, respectively. Tested on the 5k-WSJ0 task, the proposed ridge MLLR and ridge CMLLR algorithms give significant word error rate reduction from the errors obtained with standard MLLR and CMLLR in an utterance-by-utterance unsupervised adaptation scenario.

Index Terms: shrinkage model adaptation, insufficient data, ridge MLLR and ridge CMLLR

1. Introduction

Parameter adaptation is one of the most efficient techniques to address the potential mismatches between the training and testing environments in automatic speech recognition (ASR). Maximum likelihood linear regression (MLLR) [1] and maximum a posteriori (MAP) [2] adaptation are two such successful methods. If there is only a limited amount of adaptation data, MLLR and its variant, constrained MLLR (CMLLR) [3] (also known as feature space MLLR, or fMLLR [4]) are often preferred because they map the original model space into a new space by linear transformations. Matrix clusters can also be used for transformation sharing. In some application scenarios, such as voice search or voice mail transcription [5], only one utterance can be available for self-adaptation. The limited data size may not be enough for reliably estimating even one transformation matrix. There are two popular solutions to this data insufficiency problem. One is to prepare a prior distribution for the transformation, and the MAP criterion is used for matrix parameter estimation (e.g., MAPLR [6]). The other solution is to employ eigen-family methods, such as eigen-voice [7], eigen-MLLR [8], and eigen-fMLLR [9]. They online estimate the combination coefficients for a series of pre-computed basis vectors or matrices. If the number of combination coefficients is less than the total number of transformation parameters, the coefficients estimation requires less data. Both solutions need either pre-computed prior distributions or a collection of basis vectors/matrices.

In statistical learning, parameter shrinkage has been demonstrated as an effective method to handle the data sparsity problem. By adding an element-wise power regularization term to the original objective function, the shrinkage method can be very effective in controlling over-fitting because it shrinks parameter values toward zero and reduces the degree of freedom to estimate them. Successful applications include ridge regression [10], least absolute shrinkage and selection operator (LASSO) [11] in linear regression, and weight decaying [12] in neural networks, to name a few.

In this study, we apply the idea of parameter shrinkage to solve the data insufficiency problem in speech recognition. Using MLLR and CMLLR as examples, we formulate shrinkage model adaptation by adding a sum of element-wise power q constraint to the objective function. As a first attempt, an element-wise squared sum constraint is used to derive ridge MLLR and CMLLR. Tested on the 5k-WSJ0 task in an utterance-by-utterance unsupervised self-adaptation scenario the proposed ridge MLLR and ridge CMLLR algorithms significantly outperform the standard MLLR and CMLLR alternatives.

2. Shrinkage Model Adaptation

In this following the theory of parameter shrinkage in linear regression is briefly reviewed. Then MLLR and CMLLR with rotation matrix parameter shrinkage are formulated. Ridge MLLR and CMLLR are then derived by applying the element-wise square constraint for a close-form solution. Next, a comparison of ridge MLLR and MAPLR is discussed. Since parameter shrinkage is a general idea, we also extend it to formulate a novel full precision matrix estimation method in the last part of this section.

2.1 Parameter Shrinkage in Linear Regression

Linear regression is to use a linear model to predict the output Y with an input vector, $X = (X_1, X_2, \dots \dots X_p)$, as

$$f(X) = \beta_0 + \sum_{j=1}^{p} X_j \beta_j,$$

where p is the vector dimension, and β is the regression coefficient vector. If N samples are available, we can denote $\boldsymbol{X}$ as the $N * (p+1)$ input matrix (including value 1 as the bias), and $\boldsymbol{y}$ as the N-vector outputs. Then β can be estimated by minimizing the residual sum-of-squares, defined as $RSS(\beta) = (\boldsymbol{y} - \boldsymbol{X}\beta)^T(\boldsymbol{y} - \boldsymbol{X}\beta)$.

When N is small, i.e., only limited samples are available, a regularization term (element-wise q power sum) can be

added to $RSS(\beta)$ for a reliable estimations of β, i.e.,

$$\hat{\beta} = argmin\,(\boldsymbol{y} - \boldsymbol{X}\beta)^T(\boldsymbol{y} - \boldsymbol{X}\beta) + \lambda \sum_{j=1}^{p} |\beta_j|^q .$$

If $q = 2$, it is called ridge regression [10]. In the case of $q = 1$, the method is called LASSO [11].

The element-wise power q sum constraint makes some components in β shrink toward zero and assures the available samples can reliably estimate the remaining components. This shrinkage strategy has been demonstrated to be effective in the machine learning community [13].

2.2 Formulation of Shrinkage Model Adaptation

As stated in [1], the above linear regression is a special case of MLLR. Therefore, it is straightforward to borrow the success of parameter shrinkage in linear regression to MLLR in which adaptation is performed with a linear transformation matrix W on the augmented mean vector as

$$\hat{\mu}_s = W\xi_s,$$

where W is a *p*(p+1)* matrix with $W = [b\ A]$, b is a bias vector, and A is a rotation matrix. $\hat{\mu}_s$ is the new mean of state s and ξ_s is the augmented vector of the mean vector μ_s.

$$\xi_s = [1, \mu_s']'.$$

This can be solved with the expectation maximization (EM) algorithm [14] by maximizing this auxiliary function:

$Q_{MLLR} = -\frac{1}{2}\sum_i \sum_{t=1}^{T} \gamma_i(t)\big[plog2\pi + \log|\Sigma_i| + (x_t - W\xi_i)'\Sigma_i^{-1}(x_t - W\xi_i)\big]$, or

$Q_{MLLR} \sim -\frac{1}{2}\sum_i \sum_{t=1}^{T} \gamma_i(t)[(x_t - W\xi_i)'\Sigma_i^{-1}(x_t - W\xi_i)]$,

where x_t is the observation vector at time t, $\gamma_i(t)$ is the posterior probability of state i at time t, and Σ_i is the covariance matrix of state i.

Now we can formulate the shrinkage MLLR as follows:

$$max\ -\frac{1}{2}\sum_i \sum_{t=1}^{T} \gamma_i(t)[(x_t - W\xi_i)'\Sigma_i^{-1}(x_t - W\xi_i)]$$

or

$$min\ \frac{1}{2}\sum_i \sum_{t=1}^{T} \gamma_i(t)[(x_t - W\xi_i)'\Sigma_i^{-1}(x_t - W\xi_i)]$$

with a constraint that

$$\sum_{ij} |A_{ij}|^q < c \quad ,$$

where c is a positive constant.

Re-formulating the problem, we have

$\widehat{W} = [\hat{b}\ \hat{A}] = \arg min\ \frac{1}{2}\sum_i \sum_{t=1}^{T} \gamma_i(t)\big[(x_t - W\xi_i)'\Sigma_i^{-1}(x_t - W\xi_i)\big] + \frac{\lambda}{2}\sum_{ij} |A_{ij}|^q$,

where λ is the interpolation coefficient.

Similarly, we can have shrinkage CMLLR formulated as

$\widehat{W} = [\hat{b}\ \hat{A}] = \arg min\ \frac{1}{2}\sum_i \sum_{t=1}^{T} \gamma_i(t)\big[(W\zeta(t) - \mu_i)'\Sigma_i^{-1}(W\zeta(t) - \mu_i) - \log|A|^2\big] + \frac{\lambda}{2}\sum_{ij} |A_{ij}|^q$,

where $\zeta(t) = [1\ x(t)']'$.

2.3 Adaptation with a Sum of Squares Constraint

We next solve shrinkage model adaptation when $q = 2$, i.e., with a square sum constraint. We call them ridge MLLR and ridge CMLLR, in the same notion as ridge regression [9].

For ridge MLLR, we have

$\widehat{W} = [\hat{b}\ \hat{A}] = \arg min\ \frac{1}{2}\sum_i \sum_{t=1}^{T} \gamma_i(t)\big[(x_t - W\xi_i)'\Sigma_i^{-1}(x_t - W\xi_i)\big] + \frac{\lambda}{2}\sum_{ij} |A_{ij}|^2$,

with $\left(\sum |A_{ij}|^2\right)^{\frac{1}{2}} = \|A\|_F$, known as the Frobenius norm.

Take the derivative of $W = [b\ A]$, and set it as 0. Since $\partial\|A\|_F^2/\partial A = 2A$, we have

$$\frac{\partial Q}{\partial b} \sim -\sum_i \sum_{t=1}^{T} \gamma_i(t)\Sigma_i^{-1}(x_t - b - A\mu_i) = 0$$

$$\frac{\partial Q}{\partial A} \sim -\sum_i \sum_{t=1}^{T} \gamma_i(t)\Sigma_i^{-1}(x_t - b - A\mu_i)\mu_i' + \lambda A = 0$$

With some term manipulations, we can have the line-by-line solution as

$$G^l = \sum_i \frac{1}{\sigma_i^2}\xi_i\xi_i' \sum_{t=1}^{T} \gamma_i(t) + [0, \lambda, \dots\dots, \lambda]I$$

$$K^l = \sum_i \sum_{t=1}^{T} \gamma_i(t)\frac{1}{\sigma_i^2}\, x_t\xi_i'$$

$$w^l = (G^l)^{-1}K^l$$

where w^l is the lth row of W.

Similarly, ridge CMLLR has the G^l and K^l as the following, and the solution has the same format as the process in [3].

$$G^l = \sum_i \frac{1}{\sigma_i^2}\sum_{t=1}^{T} \gamma_i(t)\zeta(t)\zeta(t)' + [0, \lambda, \dots\dots, \lambda]I$$

$$K^l = \sum_i \frac{1}{\sigma_i^2}\mu_i \sum_{t=1}^{T} \gamma_i(t)\, x_t\zeta(t)'$$

2.4 Discussion

In MAPLR [5], the following prior density is used

$$p(W) \sim |\Sigma|^{-\frac{p+1}{2}}|\Phi|^{-\frac{p}{2}} exp\left\{-\frac{1}{2}tr(W - M)'\Sigma^{-1}(W - M)\Phi^{-1}\right\}$$

where $\{M, \Sigma, \Phi\}$ are the corresponding hyper-parameters, $W, M \in \Re^{p*(p+1)}$, $\Sigma \in \Re^{p*p}$, and $\Phi \in \Re^{(p+1)*(p+1)}$. In an extreme case, if we set matrix M as 0, Σ and Φ as the identity matrix, then

$$p(W) \sim \exp\{-\frac{1}{2}W'W\}.$$

With some derivation, this MAPLR will have a very similar formulation as ridge MLLR. However, the major difference is that ridge MLLR only regularizes the rotation matrix A and still gives the freedom to estimate the bias b when the amount of adaptation data is very limited. In contrast, in the above extreme case, MAPLR will simply make the estimation of $W = [b\ A]$ to be 0.

Also, in MAPLR the hyper-parameters $\{M, \Sigma, \Phi\}$ are always pre-computed from the training set. Therefore, the above extreme case should not exist. In another word, MAPLR estimates the transformation parameters by incorporating regularization from prior knowledge while ridge MLLR uses a constraint that aims at improving the estimation reliability without any prior knowledge.

2.5 Extension

The idea of shrinkage model adaptation can also be extended to other data insufficiency problems in ASR. For example, because of limited available data, the model

covariance matrix in ASR is usually assumed to be diagonal. If we intend to use the full covariance matrix, model tying (such as semi-tied covariance matrices [15]) or precision matrix combinations [16] are usually used.

We believe the parameter shrinkage method can also be utilized by applying regularization to the model precision matrix P_s with the following formulation:

$$\widehat{P_s} = \arg min \frac{1}{2}\sum_{t=1}^{T}\gamma_i(t)[(x_t-\mu_s)'P_s(x_t-\mu_s) + \log|P_s^{-1}|] + \frac{\lambda}{2}\sum_{i\neq j}|P_{sij}|^q$$

With the element-wise q power sum constraint on the off-diagonal precision matrix terms, we should have reliable estimation of the model precision matrix in the data sparsity scenario. We will study this issue elsewhere.

3. Experiment

We used the 5k-WSJ0 task to evaluate the effectiveness of shrinkage MLLR and shrinkage CMLLR. The training set is the SI-84 set with 7077 utterances. All testing is conducted on the Nov92 evaluation set with 330 utterances. Baseline models used cross-word triphones obtained with maximum likelihood estimation. There were 2818 shared states resulted from a decision tree state clustering. Each state observation density is characterized by an 8-mixture Gaussian mixture model. The input features were 12 MFCCs + energy, and their first and second order time derivatives. A trigram language model was used for decoding. The baseline word error rate (WER) was 5.08%.

To evaluate the unsupervised adaptation performance, every test utterance is first decoded to get its transcription hypothesis. Then this decoded transcription is used to adapt models for this utterance. The adapted model is used to get the final decoding transcription.

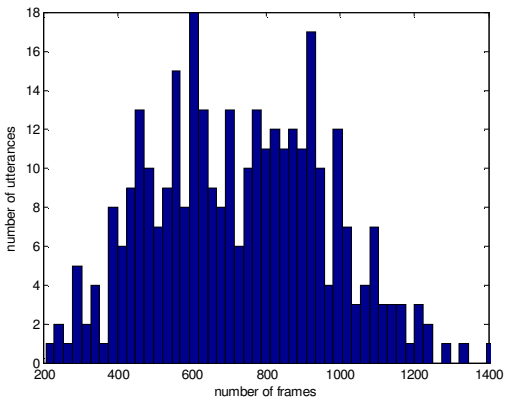

Figure 1: *Histogram of the number of frames in the WSJ0 test utterances*

As shown in Figure 1, the test set of WSJ0 is quite suitable to evaluate the proposed approach on the insufficient data problem. If we consider that roughly 39*39+39=1560 frames are need to reliably estimate the MLLR or CMLLR transform matrix, no test utterance can reach that criterion.

Table 1 lists the WERs of the baseline, standard MLLR, and ridge MLLR with different setups. The standard MLLR gets a slightly better WER than the baseline. Most likely, it is because of the data insufficiency problem. Within a broad range of λ (from 100 to 500), ridge MLLR is much better than the standard MLLR. If λ is too small, then ridge MLLR behaviors similarly as standard MLLR. In contrast, if λ is too large (e.g., 800), ridge MLLR shrinks toward bias estimation with small gain. The best ridge MLLR obtains 4.58% WER, corresponding to about 7.10% relative WER reduction from the 4.93% WER of standard MLLR.

Table 1: *Detailed WERs of baseline, standard MLLR, and ridge MLLR*

system	WER
Baseline	5.08
MLLR	4.93
ridge MLLR ($\lambda = 20$)	4.84
ridge MLLR ($\lambda = 100$)	4.69
ridge MLLR ($\lambda = 200$)	4.58
ridge MLLR ($\lambda = 500$)	4.58
ridge MLLR ($\lambda = 800$)	4.86

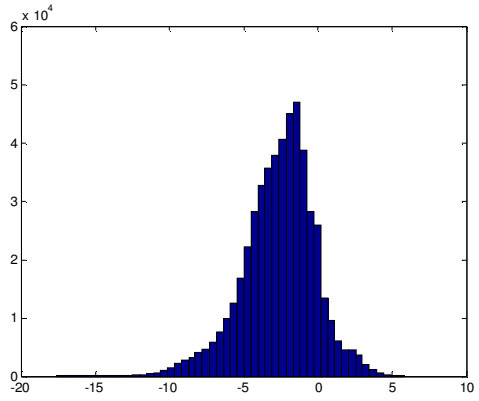

Figure 2: *Histogram of the log values of MLLR (rotation matrix A) coefficients*

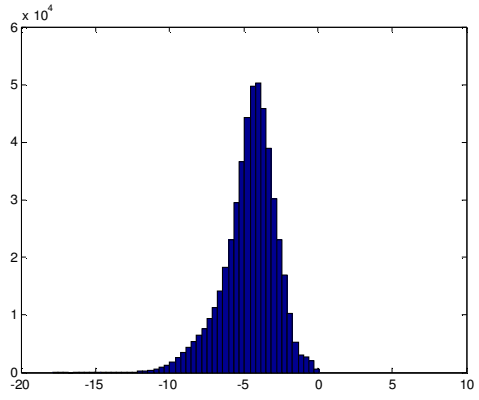

Figure 3: *Histogram of the log values of ridge MLLR (rotation matrix A) coefficients ($\lambda = 500$)*

Figures 2 and 3 compare the histograms of the log values of standard MLLR and ridge MLLR coefficients in the rotation matrix *A*. It is clear that the log values of the ridge MLLR coefficients in Figure 3 are shifted left from those in Figure 2, which means more ridge MLLR coefficients approach 0. This is exactly the effects of parameter shrinkage. As discussed above, by the parameters shrinking toward 0, we can have more reliable parameter estimation, as demonstrated in Table 1.

Table 2 lists the WERs of baseline, standard CMLLR, and ridge CMLLR with different setups. It is noted that standard CMLLR works better than standard MLLR in Table 1. A possible reason is that standard CMLLR adaptation works better in ill-condition with limited data because of its mean-covariance constraint.

Ridge CMLLR still outperforms standard CMLLR in Table 2. The best case achieves 6.7% relative WER reduction from standard CMLLR. It is also noted that the WERs of ridge CMLLR are slightly lower than those of ridge MLLR. The behaviors of different λ values for ridge CMLLR are similar to those for ridge MLLR.

Table 2: *Detailed WERs of baseline, standard CMLLR, and ridge CMLLR*

system	WER
Baseline	5.08
CMLLR	4.78
ridge CMLLR ($\lambda = 20$)	4.76
ridge CMLLR ($\lambda = 100$)	4.56
ridge CMLLR ($\lambda = 200$)	4.46
ridge CMLLR ($\lambda = 500$)	4.52
ridge CMLLR ($\lambda = 800$)	4.69

4. Conclusion

We have formulated a general form of shrinkage model adaptation to address the data insufficiency problem. Because the element-wise q power sum constraint automatically shrinks the rotation matrix coefficients toward zero, the adaptation parameters can have a reliable estimation. The shrinkage method does not require the use of pre-computed basis vectors/matrices or priors for fast adaptation. Using the element-wise square sum constraint, ridge MLLR and ridge CMLLR are derived as the first attempt to show the effectiveness of shrinkage model adaptation. We used the 5k-WSJ0 task for the unsupervised self adaptation test. Ridge MLLR and ridge CMLLR achieved about 7.1% and 6.1% relative WER reductions from their standard counterparts, respectively.

In this study, we also propose a novel full covariance matrix estimation method by using the regularization term to handle the data insufficiency problem. We believe the parameter shrinkage methodology is powerful and can be generalized to the common data sparsity problems in ASR.

This paper only presents our initial study. We are now working on shrinkage model adaptation with the L_1 norm constraint. The problem of regression with the sum of square constraint is that all the regularized coefficients are shrunk in the same time. In contrast, regression with the L_1 norm constraint can have some of the regularized coefficients exactly to be zero. This brings more benefits to the data insufficiency problem [11]. Since ridge MLLR and ridge CMLLR already work very well in this study, we expect shrinkage MLLR and CMLLR with the L_1 norm constraint should have even better performance.

5. References

[1] Leggetter, C. J. and Woodland, P. C., "Maximum likelihood linear regression for speaker adaptation of continuous density HMMs," *Computer Speech and Language*, vol. 9, no. 2, pp. 171-185, 1995.

[2] Gauvain, J.-L. and Lee, C.-H., "Maximum a posteriori estimation for multivariate Gaussian mixture observations of Markov chains," *IEEE Trans. on Speech and Audio Proc.*, Vol. 2, No. 2, pp. 291-298, 1994.

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

[4] Li, Y., Erdogan, H., Gao, Y., Marcheret, E., "Incremental online feature space MLLR adaptation for telephony speech recognition," *Proc* Interspeech, pp. 1417-1420, 2002.

[5] Wang, Y.Y., Yu, D., Ju, Y.-C., and Acero, A., "An Introduction to voice search," in *IEEE Signal Processing Magazine (Special Issue on Spoken Language Technology)*, pp. 29-38, May 2008.

[6] Chesta, C., Siohan, O., and Lee, C. -H., "Maximum a posteriori linear regression for hidden Markov model adaptation," in *Proc. Eurospeech*, pp. 211-214, 1999.

[7] Kuhn, R., Junqua, J.-C., Nguyen, P., and Niedzielski, N., "Rapid speaker adaptation in eigenvoice space," *IEEE Trans. on Speech and Audio Proc.*, Vol. 8, No. 6, pp. 695-707, 2000.

[8] Chen, K.-T., Liau, W.-W., Wang, H.-M., and Lee, L.-S., "Fast speaker adaptation using eigenspace-based maximum likelihood linear regression," in *Proc. ICSLP*, vol. 3, pp. 742-745, 2000.

[9] Cui, X., Xue, J., and Zhou, B., "Improving online incremental speaker adaptation with eigen feature space MLLR," in *Proc. ASRU*, pp. 136 – 140.

[10] Hoerl, A. and Kennard, R., "Ridge regression: Biased estimation for nonorthogonal problems," *Technometrics*, vol. 12, 55-67, 1970.

[11] Tibshirani, R., "Regression shrinkage and selection via the LASSO". *Journal of the Royal Statistical Society*, Series B, vol. 58: 267–288, 1996.

[12] Christopher M. B., *Neural Networks for Pattern Recognition*, Oxford University Press, 1995.

[13] Christopher M. B., *Pattern Recognition and Machine Learning*, Springer Press, 2007.

[14] Dempster, A., Laird, N., and Rubin, D., "Maximum likelihood from incomplete data via the EM algorithm," *Journal of the Royal Statistical Society*, Series B, 39(1):1–38, 1977.

[15] Gales, M.J.F., "Semi-tied covariance matrices for hidden Markov models," *IEEE Trans. on Speech and Audio Proc.*, Vol. 7, No. 3, pp. 272-281, 1999.

[16] Vanhoucke, V., Sankar, A.; "Mixtures of inverse covariance", *IEEE Trans. Speech Audio Processing*, vol. 12, no.3, pp. 250 - 264, 2004.