| | Number of Dimensions | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | 10 | 25 | 50 | 75 | 100 | 150 | 200 | 300 | 500 | 1000 | opt. |
| Affine, $SS$ | 119 | 90 | 79 | 74 | 71 | 68 | 67 | 67 | 69 | - | **67** |
| | *115* | *84* | *71* | *64* | *59* | *54* | *51* | *47* | *43* | | |
| Affine, $\chi^2$ | 186 | 106 | 85 | 78 | 73 | 69 | 68 | 66 | 67 | - | **66** |
| | *185* | *102* | *79* | *71* | *64* | *57* | *52* | *48* | *44* | | |
| Exponential | 527 | 345 | 267 | 219 | 194 | 158 | 131 | 106 | 92 | 89 | **89** |
| | *511* | *317* | *230* | *175* | *145* | *106* | *80* | *57* | *45* | *41* | |
| Spherical | 141 | 123 | 122 | 122 | 123 | 123 | 125 | 128 | 135 | - | **122** |
| | *134* | *111* | *104* | *100* | *97* | *92* | *90* | *89* | *88* | | |
| Affine, NMF | 104 | 83 | 75 | 71 | 69 | 66 | 64 | 63 | 62 | 61 | **61** |
| | 99 | 79 | 66 | 62 | 58 | 57 | 56 | 55 | 55 | 54 | |
| EM (tempered) | 103 | 81 | 74 | 69 | 67 | 64 | 62 | 61 | 61 | 60 | **60** |
| | 98 | 77 | 68 | 62 | 59 | 55 | 53 | 51 | 50 | 50 | |

*Table 3.* Perplexity results on the Bigram1000 data. Numbers in italics are results obtained on the training data. The last column reports the optimal result obtained by picking the best number of dimension $K$.

verb-phrase pair and morphologically reduced these words using the WordNet Morphy system. The data set Verb1000 models the occurrence of the verbs given the preceding subject.

### 4.3 Series of Experiments

We have empirically investigated the following models: (i) affine model with the sum of squared error measure in (4), (ii) affine model with $\chi^2$ fitting criterion in (3), (iii) exponential subfamily model with one-step orthogonal projection, (iv) spherical model with one-step geodesic projection, (v) affine (convex) model trained with the multiplicative matrix update rule and early stopping, and (vi) affine (convex) model trained by tempered EM.

We have used the log-likelihood as an evaluation criterion. Following common practice in language modeling, we report perplexity results:

$$PERP \equiv \exp\left[-\sum_{x\in\bar{X}}\sum_{j=1}^{M} n_j^x \log\theta_j^x / L\right]. \qquad (12)$$

The data has been split into three sets: (i) a training set consisting of 80% of the observations, (ii) a (hold-out) validation set to determine optimal parameters (such as dimensions, smoothing parameters, number of iterations, optimal $\beta$), (iii) a test set on which we report perplexity results. For both, validation and test set, 10% of the data was used.

**Perplexity Minimization by Dimension Reduction** In the first series of experiment, we have computed (approximately) optimal families $F_K(\phi)$ for various choices of dimensions. Results are reported for $K = 10, 25, 50, 100, 150, 200, 500, 1000, 2000$. Tables 1, 2, and 3 show the results on Medline1033, Verb1000, and Bigram1000, respectively.

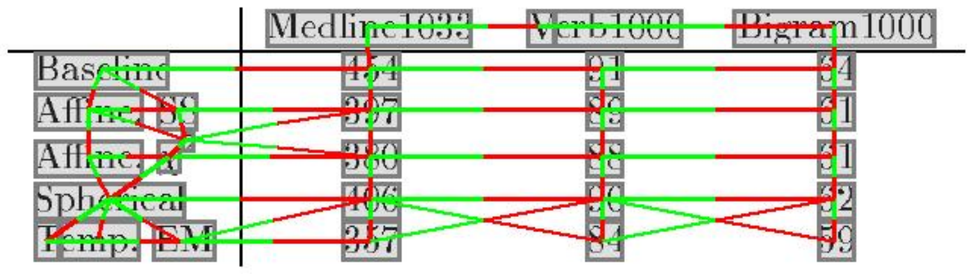


*Table 4.* Perplexity results for smoothing the MLE.

Let us compare the SVD–based class of methods first. (i) In accordance with results reported in (Gous, 1998; Gous, 1999), the spherical model shows the best fit on the training data for an intermediate range of dimensions, while the affine model trained with weighted sum of squared error (4) is typically doing best for lower dimensionalities. The exponential model does not fit the training data well. (ii) In term of generalization performance, however, the performance of the spherical model is very poor as can be seen from the large discrepancy between training vs. test set performance. Overall, the affine model based on (3) shows the best generalization performance among the SVD-based approaches.

Comparing SVD–based methods with the iterative optimization techniques based on the mixture model formulation, the latter show substantial and consistent performance gains on all data sets. The regularization in tempered EM proves to be more effective than early stopping. Tempered EM achieves the best results for all data sets and all dimensions.

**Low-Dimensional Families as Back-off Models** In a second series of experiments, we have evaluated the use of low-dimensional families for the purpose of smoothing the MLE. More precisely, we have investigated a linear interpolation scheme of the type $\bar{\theta}^x = \lambda_1\hat{\theta}^x + \lambda_2\theta(\tau^x,\phi) + \lambda_3\hat{\theta}^0$, where $\lambda_1+\lambda_2+\lambda_3 = 1$.