SPSA vs Simplex in statistical machine translation optimization

Patrik Lambert^{*1} and **Rafael E. Banchs**^{**1,2}

[1] Universitat Politecnica de Catalunya, modul D5-119, Jordi Girona 1-3, 08034 Barcelona, Spain.
[2] Fundacion Barcelona Media, Ocata 1, 08003 Barcelona, Spain.

Most statistical machine translation systems are combinations of various models and tuning scaling factors is an important step. However, this optimisation problem is hard because the objective function has many local minima and the available algorithms cannot achieve a global optimum. Consequently, optimisations starting from different initial settings can converge to fairly different solutions. We present tuning experiments with the Simultaneous Perturbation Stochastic Approximation (SPSA) algorithm, and compare them with the widely used downhill simplex method. With IWSLT 2005 Chinese-English data, both methods showed similar performance, but SPSA was more robust to the choice of initial settings.

1 Introduction

In present SMT systems, the noisy channel approach has been expanded to a more general maximum entropy approach in which a log-linear combination of multiple feature functions is implemented. Translation quality can be improved by adjusting the weight of each feature function in the log-linear combination. This can be effectively performed by minimising translation error over a development corpus for which manually translated references are available [1].

In this paper we compare tuning with the Downhill Simplex method and with the Simultaneous Perturbation Stochastic Approximation [2] (SPSA) method. SPSA has been successfully applied in areas including statistical parameter estimation, simulation-based optimisation, signal and image processing. The SPSA method is based on a gradient approximation which requires only two evaluations of the objective function (or one in one-sided gradient approximation), regardless of the dimension of the optimisation problem. This feature makes it especially powerful when the number of dimensions is increased. The SPSA procedure is in the general recursive stochastic approximation form:

$$\hat{\lambda}_{k+1} = \hat{\lambda}_k - \mathbf{a}_k \hat{\mathbf{g}}_k(\hat{\lambda}_k) \tag{1}$$

where $\hat{\mathbf{g}}_k(\hat{\lambda}_k)$ is the estimate of the gradient $\mathbf{g}(\lambda) \equiv \partial E / \partial \lambda$ at the iterate $\hat{\lambda}_k$ based on the previous mentioned evaluations of the objective function. a_k denotes a positive number that usually gets smaller as k gets larger. One-sided gradient approximations involve evaluations of $E(\hat{\lambda}_k)$ and $E(\hat{\lambda}_k + \text{perturbation})$. In the simultaneous perturbation approximation, all elements of $\hat{\lambda}_k$ are randomly perturbed together and the approximated gradient vector is:

$$\frac{E(\hat{\lambda}_k + c_k \mathbf{\Delta_k}) - E(\hat{\lambda}_k)}{2c_k} \begin{bmatrix} 1/\Delta_{k1} \\ 1/\Delta_{k2} \\ \vdots \\ 1/\Delta_{kN} \end{bmatrix} \tag{2}$$

In equation 2, $\mathbf{\Delta}_k$ is a perturbation vector of same dimension N as λ, whose values Δ_i are computed randomly. c_k denotes a small positive number that usually gets smaller as k gets larger. Notice that in this case only one objective function evaluation is needed in order to approximate the gradient.

2 Translation system and data set

The SMT approach used here considers a translation model which is based on a 4-grams language model of bilingual units which are referred to as tuples. Tuples are extracted from Viterbi alignments and can be formally defined as the set of shortest phrases that provides a monotonic segmentation of the bilingual corpus. In addition to the bilingual 4-gram translation model, the translation system implements a log linear combination of five additional feature functions. For a more complete description see [3].

The translation system was trained with the Chinese-English data provided for IWSLT'06 evaluation campaign, and the parameters were tuned over the development set provided for the same evaluation (http://www.slc.atr.jp/IWSLT2006/).

[*] Corresponding author: e-mail: lambert@gps.tsc.upc.edu, Phone: +34 934 011 066, Fax: +34 934 016 447
[**] e-mail: rafael.banchs@barcelonamedia.org

3 Experimental results

We report in table 1 the average BLEU score and standard deviation obtained after running 20, 40, 60 and 80 iterations of the simplex and SPSA algorithms. When an optimisation converged before 80 iterations, the optimum value was taken. In each cell of table 1, the upper number refers to the simplex, and the lower refers to SPSA. For each initial set of parameters, average and standard deviation are calculated over 10 slightly different realisations controlled with the random seeds.

Table 1 Average BLEU score and standard deviation obtained with the simplex method (above) and SPSA method (below) in the development set, after 20, 40, 60, and 80 iterations, for different initial points in parameter space. In the last column the average and standard deviation of the averages are displayed.

Iterations	Initial Points							Average
	1	2	3	4	5	6	7	
20	17.8±0.45 17.6±0.84	18.8±0.44 18.9±0.48	19.1±0.24 19.3±0.40	18.9±0.45 18.7±0.63	19.4±0.23 19.4±0.19	19.5±0.19 19.3±0.20	19.5±0.14 19.4±0.23	19.0±0.59 18.9±0.64
40	18.4±0.50 18.8±0.62	19.1±0.40 19.3±0.15	19.4±0.31 19.5±0.23	19.5±0.33 19.5±0.28	19.6±0.22 19.5±0.17	19.7±0.21 19.5±0.20	19.7±0.17 19.6±0.14	19.3±0.47 19.4±0.25
60	18.7±0.43 19.5±0.25	19.2±0.40 19.5±0.11	19.5±0.36 19.5±0.24	19.6±0.35 19.6±0.18	19.6±0.25 19.5±0.17	19.8±0.20 19.6±0.16	19.8±0.17 19.7±0.12	19.5±0.39 19.6±0.08
80	18.8±0.46 19.6±0.14	19.3±0.43 19.5±0.17	19.5±0.37 19.6±0.14	19.7±0.35 19.7±0.11	19.6±0.25 19.6±0.15	19.8±0.20 19.7±0.10	19.8±0.17 19.8±0.10	19.5±0.37 19.6±0.08

From table 1 it seems that both algorithms have very similar performance in terms of the optimum value achieved after a given number of iterations. Nevertheless, it is remarkable that from 60 iterations on, the standard deviation is always smaller for the SPSA algorithm, which suggests that this is a more stable method. Of course, this may be due to the fact that our implementation of the different realisations was unfair for the simplex. A change of seed to generate the simultaneous perturbation for the SPSA may be less significant as a change of initial simplex. To verify this, we need to fix the seed and see how the algorithm behaves across several initial points. We decided to explore this point in greater detail. Since for a given realisation the only varying factor is the starting point, we computed the average BLEU score and standard deviation over all 7 starting points for each of the 10 random seeds considered in the previous experiment, after 20, 40, 60, and 80 iterations. It was observed that, with the exception of one case, the optimum value obtained with SPSA was much less sensitive to the choice of initial parameters. For SPSA, the highest standard deviation after 80 iterations was below 0.2, while for the simplex, it reached 0.67. Thus doing two successive optimisations, one can expect in average up to 0.4 percent BLEU difference with SPSA and up to more than 1.3 percent BLEU difference with the simplex.

4 Conclusions and further work

We have presented experiments in which the SPSA algorithm has been used to tune SMT parameters. These experiments have been repeated with the downhill simplex method for comparison. According to the results obtained in this task, both methods seem to have similar performance. However, SPSA was more robust than the simplex with respect to the choice of initial parameters and with respect to slightly different realisations of the algorithm. This study was only performed on the development set. The first future work to be done is to see if the results obtained are reflected when the optimum parameters are used to translate a test set. We should also investigate if at some point over-fitting appears. Finally, this study should be confirmed with more experiments on various different tasks.

Acknowledgements This work has been partially funded by the European Union under the integrated project TC-STAR - Technology and Corpora for Speech to Speech Translation -(IST-2002-FP6-506738, http://www.tc-star.org).

References

[1] F. Och, 2003, “Minimum error rate training in statistical machine translation”, in *Proc. of the 41th Annual Meeting of the Association for Computational Linguistics*, pp. 160167.

[2] J. C. Spall, 1998, “Implementation of the simultaneous perturbation algorithm of stochastic optimization”, *IEEE Trans. Aerospace and Electronic Systems*, vol. 34, no. 3, pp. 817823.

[3] J. Mariño, R. Banchs, J. M. Crego, A. de Gispert, P. Lambert, J. Fonollosa, and M. Ruiz, 2005, “Bilingual ngram statistical machine translation”, in *Proc. of Machine Translation Summit X*, Phuket, Thailand, pp. 27582.