Discussion on the paper: Catching up faster by switching sooner..., by Erven, Grünwald and Rooij

Iain Murray
School of Informatics, University of Edinburgh, UK

I'd like to thank the authors for their *catch-up* description. In particular, the example illustrated in Fig. 1 has useful tutorial value, and I regular refer people to it. The models used in the example are deliberately crude, to construct a clear example. Nevertheless, I think it is worth explicitly reviewing why the more powerful model suffers from the catch-up phenomenon, and how it might be avoided through hierarchical modelling.

The 2nd order Markov model in the example (a 'trigram model') performs worse than the 1st order model for small datasets. This result still holds for text actually generated from a 2nd order Markov model, when the trigram statistics are matched to English characters. The subjective Bayesian demands an explanation: we should use the model we believe, regardless of how much data we have.

The trigram model does poorly whenever the two characters providing context have rarely been seen before: its uniform prior does not allow generalization from past experience with other contexts. In real-world language modelling applications, prediction are 'smoothed' with statistics from shorter contexts (Chen and Goodman, 1998). I ran a 'Witten–Bell' smoothed trigram model on the Alice text: it outperformed both the other Markov models across the range (after the first few characters). The catch-up phenomenon disappeared. Moreover, the smoothed model *vastly* outperformed the switch distribution (by about $50,000$ bits by the end of Fig. 1).

In other settings I believe the catch-up phenomenon will also indicate priors that make inefficient use of data, such as structureless priors over many variables. As is well known, we should use hierarchical models, where relationships between parameters can be learned (e.g. Gelman et al., 2003). Indeed, one of the best language models, interpolated Kneser–Ney, can be derived as approximate inference for a hierarchical model, and full Bayesian inference provides state-of-the art performance (Goldwater et al., 2006; Teh, 2006).

In applications less well explored than language modelling, the catch-up phenomenon may be hard to avoid. It is certainly worth checking for, and I hope the switch distribution is useful. However, even better results might come from making large models that work at least as well as small ones (as in Rasmussen and Ghahramani, 2001). There is some work on transferring prior knowledge from simple models to more powerful ones in general situations (Neal, 2001), but this area deserves more attention. The catch-up phenomenon provides good motivation.

References

Chen, S. F. and J. Goodman (1998, August). An empirical study of smoothing techniques for language modeling. Technical Report TR-10-98, Computer Science Group, Harvard University.

Gelman, A., J. B. Carlin, H. S. Stern, and D. B. Rubin (2003). *Bayesian data analysis, second edition*. Chapman & Hall/CRC.

Goldwater, S., T. L. Griffiths, and M. Johnson (2006). Interpolating between types and tokens by estimating power law generators. In Y. Weiss, B. Schölkopf, and J. Platt (Eds.), *Advances in Neural Information Processing Systems 18*, pp. 459–466. MIT Press.

Neal, R. M. (2001). Transferring prior information between models using imaginary data. Technical Report 0108, Dept. of Statistics, University of Toronto.

Rasmussen, C. E. and Z. Ghahramani (2001). Occam's razor. In T. K. Leen, T. G. Dietterich, and V. Tresp (Eds.), *Advances in Neural Information Processing Systems 13*, pp. 294–300. MIT Press.

Teh, Y. W. (2006). A hierarchical Bayesian language model based on Pitman-Yor processes. In *Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics*, pp. 985–992.