DISCUSSION ARTICLE

Discussion

Andrew GELMAN

The general idea of optimization transfer is very appealing to me, especially since I have never succeeded in fully understanding the EM algorithm. I like the examples in Lange, Hunter, and Yang's article and suspect that even further generalization is possible, especially in the direction of allowing the optimization algorithms to have tuning parameters that themselves can be optimized over (or in statistical terms, "estimated"), as discussed by van Dyk and Meng (1999). For example, the suggested logistic regression computation (Example 3, p. 5) with $B = \frac{1}{4}\sum_{i=1}^{m} x_i x_i^t$ corresponds to a linearization of the logistic regression with equal variances for the m data points. Presumably a more effective approximation would use the local second derivatives at a reasonably chosen approximate estimate for θ. To put it another way, this and other optimization transfer algorithms could be made adaptive by occasionally updating the tuning parameters in the local optimizations.

My main interest in this article, however, is in its potential application to stochastic algorithms. In much of statistical computation, especially for applied Bayesian inference, optimization has been replaced by simulation, with mode-finding algorithms often reduced to the role of starting points for Monte Carlo simulation (see, e.g., Gilks, Richardson, and Spiegelhalter 1996). This change in practice is relevant to the article under discussion for three reasons. First, where optimization is being used mostly as a starting point, it is more important than ever for its computations to be fast, with speed, in fact, being more important than convergence. Second, if the optimization is over a statistical likelihood or posterior density (rather than, for example, a net profit in an operations research context), then there is no special virtue in a global mode, and in fact it is useful to have a fast algorithm that can be started at many places to find various local modes, all of which can be used as starting points in a subsequent iterative simulation algorithm. Third, the natural question arises as to whether the methods in this article can be generalized to simulation-based computation.

Just as data augmentation or Gibbs sampling can be viewed as a stochastic generalization of EM (Tanner and Wong 1987), is there a Markov chain "simulation transfer algorithm" that generalizes the optimization transfer of this article? I have no conclusive

Andrew Gelman is Associate Professor, Department of Statistics, Columbia University, New York 10027 (E-mail: gelman@stat.columbia.edu).

Journal of Computational and Graphical Statistics, Volume 9, Number 1, Pages 49–51

answer here, but various possibilities suggest themselves.

To start with, data augmentation and Gibbs sampling can be viewed as simulation transfer algorithms, where the simulation is performed iteratively on conditional distributions rather than on the target joint distribution. The Metropolis–Hastings algorithm transfers simulations to jumping distributions that depend on the current position of θ, with an accept-reject step adjusting for the differences in the distributions. Algorithms such as hybrid Monte Carlo (Duane, Kennedy, Pendleton, and Roweth 1987; Neal 1993) transfer simulations to a "dynamical algorithm" that moves efficiently through a local approximation of parameter space, with an accept-reject step again correcting for the inexactness of the simulation transfer.

It would be nice to know whether the algorithms discussed in this article can be transferred effectively to the simulation context. In particular, an efficient simulation algorithm for linear regression posterior distributions based on Example 7 (p. 7) would be useful. In a Metropolis–Hastings context, the challenge would be to correct for the approximation inherent in the updating step (see Wikle, Milliff, Nychka, and Berliner 1988 for a related algorithm).

Finally, it is possible that more can be done on convergence analysis, once again following up on work in Markov chain simulation. The authors define convergence in terms of increments of the hill-climbing algorithm, but we wonder whether for a well-behaved algorithm it might be excessive to wait on the order of 10 million iterations until successive increments differ by 10^{-8}. In addition to practical concerns, it seems possible that a measure of convergence based on increments might be problematic when making comparisons between different algorithms, in that this measure might very well favor conservative EM-type algorithms that always increase the objective function, compared to algorithms that take larger jumps.

A more relevant measure of convergence might use the distance between the current value and the actual optimum. Such a direct measure could be used in theoretical and simulation studies of the sort performed in this article. In practice, convergence could be measured using multiple optimization algorithms from different starting points (as in Gelman and Rubin 1992, for iterative simulations), with approximate convergence declared when the range or standard deviation between the current estimates is below some small value. Obviously, more would have to be done if the separate sequences were converging to multiple modes, but in that case one would not want to trust a single sequence anyway.

An estimate of closeness to the actual optimum would also be useful in judging how long to run an iterative optimization whose ultimate purpose is to construct starting points for a stochastic algorithm, as is common for Bayesian inference.

In any case, the interesting patterns revealed in the simulations suggest the possibility of future theoretical understanding of the relation between complexity and speed of convergence. Various tools of exploratory graphics might be useful for this purpose, as illustrated by Figure 1 of this discussion. Figure 1 replots Figure 2 from the article (p. 15), on the log-log scale.

In conclusion, I find this article fascinating, and I hope that researchers in computational statistics can follow up on its ideas to bring some order to the dizzying array of iterative optimization and simulation algorithms.

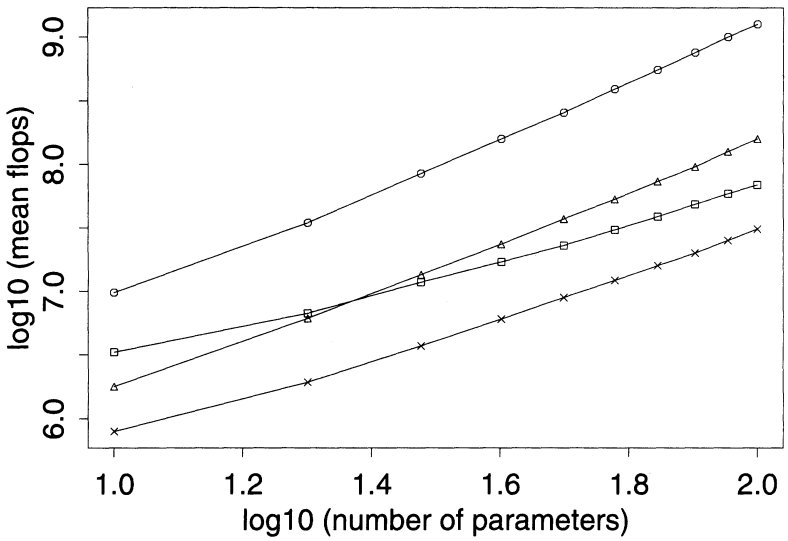

Figure 1. Figure 2 from the Lange, Hunter, and Yang article (p. 15), reexpressed on the log-log scale to more clearly show patterns in the simulation results.

ACKNOWLEDGMENTS

We thank Xiao-Li Meng for helpful comments and the U. S. National Science Foundation for grant SBR-9708424 and Young Investigator Award DMS-9796129.

[Received January 2000.]

REFERENCES

Duane, S., Kennedy, A. D., Pendleton, B. J., and Roweth, D. (1987), *Physics Letters B*, 195, 216–222.

Gelman, A., and Rubin, D. B. (1992), "Inference From Iterative Simulation Using Multiple Sequences" (with discussion), *Statistical Science*, 7, 457–511.

Gilks, W. R., Richardson, S., and Spiegelhalter, D., eds. (1996), *Practical Markov Chain Monte Carlo*, London: Chapman and Hall.

Neal, R. M. (1994), "An Improved Acceptance Procedure for the Hybrid Monte Carlo Algorithm," *Journal of Computational Physics*, 111, 194–203.

Tanner, M. A., and Wong, W. H. (1987), "The Calculation of Posterior Distributions by Data Augmentation" (with discussion), *Journal of the American Statistical Association*, 82, 528–550.

van Dyk, D. A., and Meng, X. L. (1999), "The Art of Data Augmentation," Technical Report, Department of Statistics, Harvard University.

Wikle, C. K., Milliff, R. F., Nychka, D., and Berliner, L. M. (1998), "Spatio-Temporal Hierarchical Bayesian Blending of Tropical Ocean Surface Wind Data," submitted to *Journal of the American Statistical Association*.