International Statistical Review (2014), 82, 1, 76–79 doi:10.1111/insr.12027

David R. Hunter

Department of Statistics, Pennsylvania State University, University Park, PA 16802, USA
E-mail: dhunter@stat.psu.edu

The paper by Lange, Chi and Zhou (hereafter 'the authors') serves not only as a useful reminder about the importance of optimization techniques in implementing modern statistical methods but also as a collection of various techniques along with citations for further study. The authors point out that it is useful to be able to 'mix and match' these techniques. My brief comments will focus on a few aspects of this mixing and matching.

1 Tradeoffs

In the section on algorithm acceleration, the authors state 'Many MM and block descent algorithms converge very slowly. In partial compensation, the computational work per iteration may be light.' I would add to this that the coding work can be light for MM and block descent; Hessian matrices are not always easy to calculate. Also, effectiveness of an iteration can vary widely: Near an optimum point, Newton's method often works very well (as long as the objective function is twice-differentiable), whereas it can be ineffective when the current iterate is far from an optimum. On the other hand, calculation of the Hessian matrix in Newton's method means that, in some problems, standard error estimation is easy, whereas alternative methods might have to resort to separate calculation of the Hessian matrix or approximation techniques such as those of Meng & Rubin (1991) for this purpose.

In Lange *et al.* (2000) and Hunter (2004), the tradeoff between number of iterations and complexity per iteration is illustrated using a simple model called a Bradley-Terry model. This model assumes that, in any comparison between individuals 1 and 2, where these individuals have intrinsic unobserved merit parameters $\gamma_1 > 0$ and $\gamma_2 > 0$, we have

$$P(1 \text{ beats } 2) = 1 - P(2 \text{ beats } 1) = \frac{\gamma_1}{\gamma_1 + \gamma_2} \tag{1}$$

The log-likelihood function resulting from a set of independent comparisons may be shown, under mild conditions (Hunter, 2004), to be strictly concave after the reparameterisation $\exp\{\lambda_i\} = \gamma_i$ and to admit a maximizer (which must therefore be unique). This maximum likelihood estimate may be found in multiple ways, including Newton's method and a simple MM algorithm. Newton's method definitely appears to be best in terms of total iteration count; however, total iteration count is only part of the story.

Data are simulated for a large test problem as follows: Given $\lambda_i = i + (p/10)$ for $i = 1, \ldots, p$, each of the p individuals selects $p/20$ opponents for comparison uniformly without replacement from the $p - 1$ possibilities. The resulting total $p^2/20$ comparisons are assumed independent, and the results are simulated according to the probabilities in Equation (1). Then, maximum likelihood estimates are calculated using two different methods. Both the Newton method (using the reparameterisation $\exp\{\lambda_i\} = \gamma_i$) and the MM algorithm

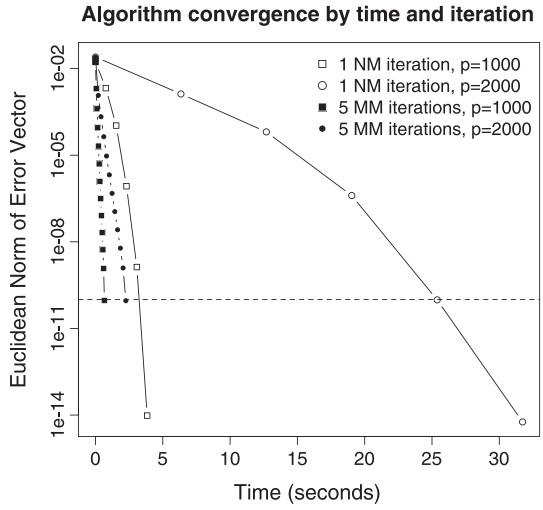

Figure 1. *Horizontal spacing between single iterations is taken to be the median time per iteration (which is multiplied by 5 for the MM points in the figure). The MM algorithms only display every five iterations for the sake of clarity. The vertical axis plots* $\|\gamma^{\text{current}} - \gamma^{\text{final}}\|$ *on a logarithmic scale.*

$$\hat{\lambda}_i^{\text{next}} = \text{wins}_i \left[\sum_{j \neq i} \frac{\text{comparisons}_{ij}}{\hat{\lambda}_i^{\text{current}} + \hat{\lambda}_j^{\text{current}}} \right]^{-1}$$

are coded in R (R Core Team, 2013); code is available at http://sites.stat.psu.edu/~dhunter/code/bt2013/. Each algorithm is started from the point $\gamma_1 = \cdots = \gamma_p$ (where the gamma vector is always normalised to sum to one), and convergence is declared when the Euclidean norm of the parameter change vector in a single iteration is less than 10^{-10}.

Theoretically, Newton's method enjoys a quadratic rate of convergence, whereas many alternative algorithms such as MM exhibit a linear rate. That is, the exponent α in

$$\|\hat{\lambda}^{\text{next}} - \hat{\lambda}^{\text{final}}\| \approx C \|\hat{\lambda}^{\text{current}} - \hat{\lambda}^{\text{final}}\|^{\alpha}$$

for current iterates near the final solution may be shown to be two for Newton's method—which is very desirable—compared with one for MM algorithms. However, Figure 1 shows that the Newton method's smaller number of iterations come at a cost: When $p = 1000$, the five Newton iterations take more than three times as long to converge as the 60 MM iterations; when $p = 2000$, the five Newton iterations take more than eight times as long as the 55 MM iterations. Although there is an inherent difficulty with using computing time as a horizontal axis label as in Figure 1—namely, the fact that different computers or even implementations will result in different timings—this is still less misleading than counting iterations.

This example illustrates that where numerical algorithms are concerned, the number of iterations to convergence can be a meaningless criterion from a practical point of view; overall speed, simplicity of implementation, ability to derive standard errors and other features of the problem such as differentiability often dictate the choice of a specific algorithm. There is simply no such thing as a universal 'gold standard' when it comes to algorithms.

2 Best of Both Worlds?

An appreciation for the relative strengths of various competing algorithms will lead to better mixing and matching. In some cases, it is even possible to combine the best features of more than one algorithm.

The authors discuss quasi-Newton methods, which may be helpful in cases where the objective function is sufficiently smooth and Newton's method itself is problematic, for instance, because the Hessian matrix is difficult to calculate or too large to invert quickly. Just as the Hessian matrix satisfies the secant condition $H_{n+1}d_n = g_n$ for

$$\begin{aligned} g_n &= \nabla f(\theta_{n+1}) - \nabla f(\theta_n) \\ d_n &= \theta_{n+1} - \theta_n \end{aligned}$$

the inverse Hessian satisfies the secant condition $H_{n+1}^{-1} g_n = d_n$. Thus, the Davidon rank-one update may be reformulated, *mutatis mutandis*, to approximate H_{n+1}^{-1} directly, thus avoiding matrix inversion in cases where this is desirable.

In addition to the acceleration ideas proposed by the authors, it bears emphasising that hybrid algorithms combining MM algorithms with quasi-Newton acceleration have been shown to be highly effective for certain problems. That is, as the authors state, 'if an accelerated step fails the descent test, one can revert to the ordinary MM or block descent step.' For example, Lange *et al.* (2000) and Hunter & Lange (2002) demonstrated that acceleration may be achieved without any matrix inversions and, using an approximation due to Jamshidian & Jennrich (1997), the secant condition leading to the Davidon update is particularly simple. Thus, these examples exploit the reliability of the MM ascent property, gradually building an approximation to the inverse Hessian function over several MM iterations until this approximation proves useful, in which case the quasi-Newton algorithm takes over and speeds convergence. For problems in multi-dimensional scaling (Lange *et al.*, 2000) and proportional odds model estimation (Hunter & Lange, 2002), this mixing of algorithms can be orders of magnitude faster, measured in total floating point operations performed, than standard Newton or MM algorithms.

3 Beyond Lasso

Of course, many modern statistical optimisation problems involve non-differentiable objective functions, which typically makes Newton-based algorithms ineffective. As a particularly relevant example, the authors discuss lasso-penalised regression, in which a loss function such as a negative log-likelihood is penalised by the weighted sum of absolute values of the individual parameter components. Thus, the objective function takes the form

$$f(\theta) + \sum_j p_j(\theta_j) \tag{2}$$

where $f(\theta)$ is smooth, possibly even quadratic, whereas the penalty function $p_j(\theta_j)$ is typically not differentiable. The authors point out that the lasso penalty $p_j(x) = \rho w_j |x|$ leads to shrinkage of many parameter estimates to zero in a standard regression setting, making the lasso penalty useful for selecting important variables. Yet Fan & Li (2001) pointed out that lasso leads to biased estimates, proposing instead a different $p_j(\theta_j)$, called the smoothly clipped absolute deviation (SCAD) penalty, that is in some sense even worse behaved than the lasso penalty from an optimisation standpoint, because the SCAD penalty is not merely non-differentiable but also non-convex. The past decade has produced a vast literature on variable selection via non-differentiable penalty terms such as lasso or SCAD.

Zou & Li (2008) discussed a technique that they call local linear approximation, that amounts to majorisation of non-differentiable penalty functions using a lasso penalty. In doing so, they essentially develop a class of MM algorithms for solving many penalised likelihood problems

in which the majorisation step involves a non-differentiable (lasso-penalised) function that must be minimised, or at least reduced, at each MM iteration. In other words, the coordinate descent approach outlined by the authors for lasso-penalised regression has potential applicability to a much wider class of modern penalty functions via the combined local linear approximation and MM ideas. Effective use of this idea will undoubtedly require tuning of the numerical method, because there is a tradeoff between the amount of work expended in the minimisation step at each iteration and the number of iterations ultimately required for convergence.

4 Conclusion

Choosing an effective optimisation method in modern statistical problems is not always a simple matter of choosing the correct 'off-the-shelf' method. Often, an effective approach will combine more than one idea. It is thus beneficial to be well-acquainted with a variety of techniques, which is why the paper by Lange, Chi and Zhou is so useful. Practitioners adept at mixing and matching these techniques will find that they are well-equipped to tackle many statistical optimisation problems.

References

Fan, J. & Li, R. (2001). Variable selection via nonconcave penalized likelihood and its oracle properties. *J. Amer. Statist. Assoc.*, **96**(456), 1348–1360.

Hunter, D. R. (2004). MM algorithms for generalized Bradley-Terry models. *Ann. Statist.*, **32**(1), 384–406.

Hunter, D. R. & Lange, K. (2002). Computing estimates in the proportional odds model. *Ann. Inst. Statist. Math.*, **54**, 155–168.

Jamshidian, M. & Jennrich, R. I. (1997). Acceleration of the EM algorithm by using quasi-Newton methods. *J. R. Stat. Soc. Ser. B*, **59**(3), 569–587.

Lange, K., Hunter, D. R. & Yang, I. (2000). Optimization transfer using surrogate objective functions. *J. Comput. Graph. Statist.*, **9**, 1–20.

Meng, X.-L. & Rubin, D. B. (1991). Using EM to obtain asymptotic variance-covariance matrices: The SEM algorithm. *J. Amer. Statist. Assoc.*, **86**(416), 899–909.

R Core Team. (2013). *R: A Language and Environment for Statistical Computing.* Vienna, Austria: R Foundation for Statistical Computing. http://www.R-project.org/.

Zou, H. & Li, R. (2008). One-step sparse estimates in nonconcave penalized likelihood models. *Ann. Statist.*, **36**(4), 1509–1533.

[*Received June 2013, accepted June 2013*]

International Statistical Review (2014), 82, 1, 79–81 doi:10.1111/insr.12040

Christian P. Robert

Universit Paris-Dauphine, CEREMADE and CREST, Paris, France; Department of Statistics, University of Warwick, UK
E-mail: xian@ceremade.dauphine.fr

The review on optimisation methods by Lange, Chi, and Zhou is quite welcomed, if only because optimisation comes less naturally to (us) statisticians than integration! It is also quite enjoyable to be reminded of the mathematics behind the methods, and I appreciated the way the ubiquitous expectation–maximisation algorithm fits into the global picture. I however fear the (motivating) perspective defended in the paper, namely that most of modern statistics can