Proposal of the vote of thanks for the paper:
Riemann manifold Langevin and Hamiltonian Monte Carlo methods
(Girolami and Calderhead)
Journal of the Royal Statistical Society, Ser. B

The Metropolis Adjusted Langevin Algorithm (MALA) and the Hybrid (or Hamiltonian) Monte Carlo Method (HMC) are successful Markov Chain Monte Carlo methods that use proposals based on knowledge of the target probability distribution and consequently outperform algorithms with random walk proposals. That improvement does not come without a price tag. Both methods include a user-determined matrix (the mass-matrix in HMC, the preconditioner in MALA) as a 'free parameter' whose tuning is a difficult art which in practice requires expensive trial runs. The stimulating paper by Girolami and Calderhead provides theoretical guidance into the choice of matrices and does so by exploiting two main ideas:

(1) The algorithms are generalized to incorporate a mass-matrix/preconditioner that is a *function* of the state of the Markov chain. At first sight this would seem to make matters worse vis-a-vis the freedom in the choice of matrix.

(2) In the case of interest in Bayesian inference, where the target is a likelihood, the authors note that, once the matrix is allowed to be state-dependent, it may be chosen to coincide with the Fisher information matrix which defines a 'natural' metric to compute distances between parameterized probability measures. Endowed with such a metric, the space of parameterized distributions is a Riemannian manifold. In the case of MALA, the resulting algorithm (MMALA) has a beautiful structure. The increment from the current state to the proposal includes a random component given by a Brownian motion on the Riemannian manifold and a deterministic component in the direction of steepest ascent in likelihood, where now 'steepest' is understood as measured by the natural metric for distributions rather than by the Euclidean distance between distribution parameter values.

The article clearly bears out the advantages of the new algorithms, MMALA and RMHMC, based on such an automatic, natural, geometric choice of the matrices and I have little doubt that it will lead to much future work. In connection with item (1) above there are obvious lines open to research. May the rather involved MMALA proposal be simplified (in ways different from that already considered in the paper)? For RMHMC, what is the most efficient way to integrate numerically the relevant canonical equations subject to preservation of geometric properties like reversibility and conservation of volume? Here the Hamiltonian function, while being separated in potential and kinetic components, possesses a variable mass-matrix, a case that has not been addressed in the, by now large, body of work on geometric numerical integration as defined by Sanz-Serna (1997). The simple leapfrog/Verlet algorithm is the integrator of choice in conjunction with standard HMC; higher-order integrators, while potentially advantageous (see the bounds in Beskos *et al.* (2010)), suffer from its more demanding computational cost per time step. For RMHMC, where the simplest integrator is implicit, would it pay to move to higher order? In more general terms, the final success of

MMALA ad RMHMC will depend on addressing a number of implementation issues, particularly so when the state space possesses large dimensionality.

Item (2) will also attract much attention, if I am not mistaken. Are there alternative useful metrics beyond that defined by the Fisher information? The authors mention in this connection the observed Fisher information matrix or the empirical information matrix. The former, the negative Hessian of the log-probability, has the appeal of making sense not only in the context of Bayesian statistics but for any target distribution and in fact may turn out to be useful in, say, sampling the canonical distribution in molecular simulations (I am currently experimenting with that possibility). Unfortunately the Hessian typically is not positive definite throughout the state space (for instance it is not in the illustrative example in Section 5.1), a difficulty that has to be addressed. By the way, in the Bayesian context of the paper, the metric tensor equals the expected information matrix plus the negative Hessian of the log-prior: the first is necessarily positive definite, the second is not in general.

The last comments lead us to explore connections with well-known ideas from the field of optimization. There is of course a clear relation between exploring a probability distribution and locating the maxima of the log-probability. The message of the paper is then akin to something that has been known for long in optimization: taking local steps in the direction of the current (standard) gradient is not the best way to reach the maximum of the objective function. The product of the inverse negative Hessian and the gradient provides a much better alternative as shown in the figure.

Something I have enjoyed when reading the article by Girolami and Calderhead is the wide variety of ideas that contribute to shape the final algorithms, from Riemannian geometry to Bayesian statistics, from Hamiltonian dynamics to numerical geometric integration. For a non-statistician like me, it has been a privilege to study a piece of work that clearly demonstrates the inherent unity of all mathematical sciences. It gives me great pleasure to propose the vote of thanks.

References

Beskos, A.; Pillai, N.; Roberts, G. O.; Sanz-Serna, J. M. and Stuart, A. M. (2010) Optimal tuning of the Hybrid Monte-Carlos algorithm. *Technical Report.* Department of Statistical Science, University College of London, London.

Sanz-Serna, J. M. (1997) Geometric integration. In *The State of the Art in Numerical Analysis* (eds. I. S. Duff and G. A. Watson), pp.121–143. Clarendon Press: Oxford.

J. M. Sanz-Serna
Universidad de Valladolid

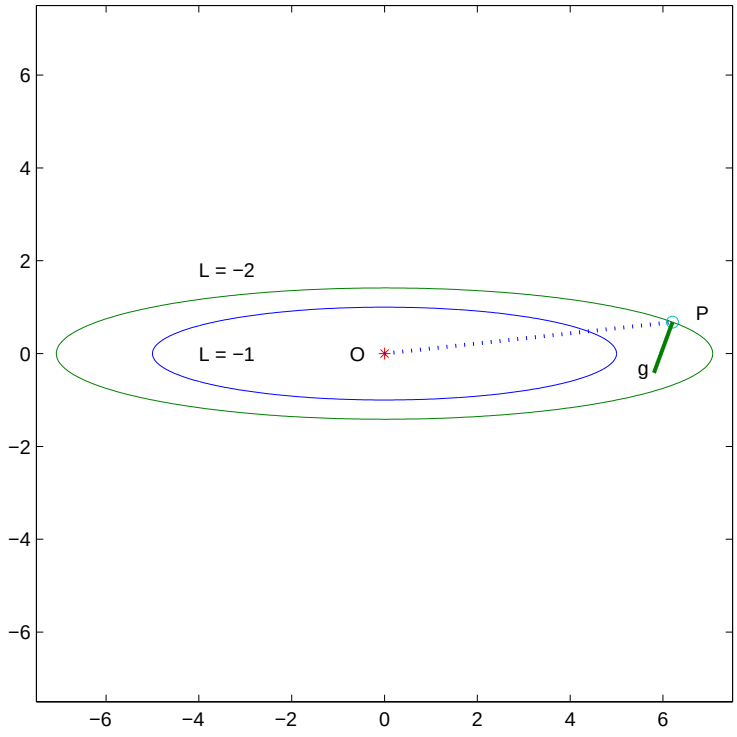

Figure 1: The ellipses represent level sets of a real quadratic function L with a maximum at the point O. The direction of the standard Euclidean gradient g of L at P is not optimal when trying to reach O from P. The best possible direction PO is given by the product of the negative inverse Hessian and g; this is the gradient of L with respect to the metric defined by the negative Hessian. In probability, L corresponds to the log-probability density. In mechanics L is the negative potential and g provides the direction of the force at P; after choosing the mass matrix appropriately the acceleration will be aligned with PO. In optics the ellipses depict wave-fronts in a non-isotropic medium, OP a ray emanating from O; the ray and the wave-front are not orthogonal in the standard sense. Hamilton found the canonical equations guided by this analogy between optics and mechanics.