Discussion of the paper “Riemann manifold Langevin and Hamiltonian Monte Carlo methods” by M. Girolami and B. Calderhead

Dr Maurizio Filippone†

University of Glasgow, Glasgow, UK.

Consider non-parametric logistic regression with Gaussian Process priors (Rasmussen and Williams, 2006), where a set of n covariates $\mathbf{x}_i \in \mathbb{R}^d$ are associated with response $y_i \in \{0, 1\}$:

$$p(\mathbf{f}|\boldsymbol{\theta}) \sim \mathcal{N}(\mathbf{f}|\mathbf{0}, K) \qquad p(y_i|f_i) = \sigma(f_i)^{y_i}(1-\sigma(f_i))^{(1-y_i)}$$

Let K be the covariance matrix parameterized by a vector of (hyper)parameters $\boldsymbol{\theta} = (\psi_\sigma, \psi_{\tau_1}, \ldots, \psi_{\tau_d})$:

$$k(\mathbf{x}_i, \mathbf{x}_j|\boldsymbol{\theta}) = \exp(\psi_\sigma)\exp\left[-\frac{1}{2}(\mathbf{x}_i - \mathbf{x}_j)^{\mathrm{T}} A (\mathbf{x}_i - \mathbf{x}_j)\right]$$

$$A^{-1} = \mathrm{diag}\left(\exp(\psi_{\tau_1}), \ldots, \exp(\psi_{\tau_d})\right)$$

We consider the manifold methods presented in this paper in comparison to a set of alternative algorithms to sample from the joint-posterior distribution of $\mathbf{f}$ and $\boldsymbol{\theta}$.

Efficiently sampling $\mathbf{f}$ and $\boldsymbol{\theta}$ is complex because of their strong coupling (Murray and Adams, 2010; Neal, 1999). Gibbs style samplers, as used by the Authors in Section 9, based on sampling $\mathbf{f}|\boldsymbol{\theta}, \mathbf{y}$ and $\boldsymbol{\theta}|\mathbf{f}, \mathbf{y}$ are convenient from an implementation standpoint, but extremely inefficient. This is due to the fact that fixing $\mathbf{f}$ induces a sharply peaked posterior for $\boldsymbol{\theta}$, resulting in a poor Effective Sample Size (ESS) for the length-scale parameters (Murray and Adams, 2010).

The metric tensor comprises the Fisher Information (FI) and the negative of the Hessian of the prior:

$$G_{\mathbf{f}} = -\mathrm{E}_{\mathbf{y}|\mathbf{f}}\left[\nabla_{\mathbf{f}}\nabla_{\mathbf{f}}\mathcal{L}\right] = \sigma(\mathbf{f})(1-\sigma(\mathbf{f})) + K^{-1} = \Lambda + K^{-1} \qquad \Lambda = \mathrm{diag}(\sigma(\mathbf{f})(1-\sigma(\mathbf{f})))$$

$$G_{\mathbf{f},\theta_i} = -\mathrm{E}_{\mathbf{y},\mathbf{f}|\boldsymbol{\theta}}\left[\frac{\partial \nabla_{\mathbf{f}}\mathcal{L}}{\partial \theta_i}\right] = -\mathrm{E}_{\mathbf{f}|\boldsymbol{\theta}}\left[K^{-1}\frac{\partial K}{\partial \theta_i}K^{-1}\mathbf{f}\right] = 0$$

$$G_{\theta_j,\theta_i} = -\mathrm{E}_{\mathbf{y},\mathbf{f}|\boldsymbol{\theta}}\left[\frac{\partial^2 \mathcal{L}}{\partial \theta_i \partial \theta_j}\right] = \frac{1}{2}\mathrm{Tr}\left(K^{-1}\frac{\partial K}{\partial \theta_i}K^{-1}\frac{\partial K}{\partial \theta_j}\right) - \frac{\partial^2 \log[p(\boldsymbol{\theta})]}{\partial \theta_i \partial \theta_j}$$

$$G = \begin{pmatrix} G_{\mathbf{f}} & \mathbf{0} \\ \mathbf{0} & G_{\boldsymbol{\theta}} \end{pmatrix}$$

The derivatives of G follow from standard properties of matrix derivatives. Note that taking the expectations with respect to $\mathbf{y}$ alone does not lead to a positive definite matrix G and

†Address for sending the proofs: Department of Statistical Science, University College. Gower Street, London WC1E 6BT, United Kingdom

Table 1. ESS for Gibbs Metropolis-Hastings (Gibbs MH), Gibbs Simplified MMALA (Gibbs S-MMALA), Gibbs RM-HMC (Gibbs RM-HMC), Gibbs Whitening (Gibbs Wht), HMC, Simplified MMALA (S-MMALA), and RM-HMC, all averaged over 10 runs (the standard deviation is in parenthesis). In Hamiltonian based methods, the maximum number of leapfrog steps was set to 30. Gibbs MH and HMC were tuned on the basis of posterior covariances estimated from pilot runs of Gibbs Wht. We also report the number of calls (in thousands) to the functions computing G, $G_{\boldsymbol{\theta}}$ and the Cholesky decomposition of K that are the main computational bottle-necks (along with the derivatives of $G_{\boldsymbol{\theta}}$ with respect to $\boldsymbol{\theta}$, although we are not reporting these statistics). All the methods were initialized from the true values used to generate the data; the ESS is computed over 2000 samples collected after 1000 burn-in samples. In Gibbs style samplers, the length-scale parameters have a poor ESS, whereas the latent functions are sampled quite efficiently by manifold methods, confirming that the geometric argument is effective in improving the sampling of $\mathbf{f}$.

	Gibbs	Gibbs	Gibbs	Gibbs			
	MH	S-MMALA	RM-HMC	Wht	HMC	S-MMALA	RM-HMC
min ESS $\mathbf{f}$	3(0)	27(17)	78(69)	26(24)	3(0)	17(6)	182(50)
avg ESS $\mathbf{f}$	6(0)	102(27)	404(92)	112(24)	6(0)	51(4)	531(80)
max ESS $\mathbf{f}$	18(3)	205(35)	888(60)	309(55)	21(4)	94(12)	1001(61)
ESS ψ_{σ}	30(11)	54(38)	30(13)	56(20)	5(2)	18(10)	530(250)
ESS ψ_{τ_1}	30(13)	6(2)	6(4)	203(112)	12(11)	6(2)	86(33)
ESS ψ_{τ_2}	36(23)	8(3)	7(3)	136(60)	10(6)	7(4)	111(40)
$10^3 \times G$	$--$	$--$	$--$	$--$	$--$	3(0)	257(18)
$10^3 \times G_{\boldsymbol{\theta}}$	$--$	3(0)	80(6)	$--$	$--$	$--$	$--$
$10^3 \times \mathrm{chol}(K)$	3(0)	3(0)	80(6)	3(0)	47(1)	3(0)	257(18)

it is therefore necessary to take them with respect to $\mathbf{y}$ and $\mathbf{f}$ jointly (for $G_{\mathbf{f}}$ we compute the expectation with respect to $\mathbf{y}$ to leave the dependency from $\mathbf{f}$). G is block diagonal, so the geometric based argument in favor of the decoupling of $\mathbf{f}$ and $\boldsymbol{\theta}$, when sampled jointly using manifold methods, does not hold.

Results and experimental settings for a bivariate logistic regression problem with $n = 100$ are reported in Table 1. The results confirm that Gibbs style samplers are very inefficient in sampling the length-scale parameters.

Gibbs with RM-HMC proposals seems suboptimal in this problem, as it may well be for the Log-Gaussian Cox model presented by the Authors in Section 9. A natural decoupling of $\mathbf{f}$ and $\boldsymbol{\theta}$ is offered by whitening the prior over $\mathbf{f}$. Given the decomposition $K = LL^{\mathrm{T}}$, define $\boldsymbol{\nu} = L^{-1}\mathbf{f}$; sampling $\boldsymbol{\theta}|\mathbf{f}, \mathbf{y}$ is replaced by $\boldsymbol{\theta}|\boldsymbol{\nu}, \mathbf{y}$.

Even if G is block diagonal, the results for computationally demanding runs of RM-HMC show some potential in achieving a comparable ESS to the whitening method. This motivates further investigation on less expensive (guiding) Hamiltonians for the joint update of $\mathbf{f}$ and $\boldsymbol{\theta}$ trading off some efficiency. Also, it would be particularly interesting to start off from the whitened model and study whether manifold methods can improve sampling efficiency.

References

Murray, I. and R. P. Adams (2010). Slice sampling covariance hyperparameters of latent Gaussian models. In *Advances in Neural Information Processing Systems 23 (NIPS 2010)*, to appear.

Neal, R. M. (1999). Regression and classification using Gaussian process priors (with discussion). *Bayesian Statistics 6*, 475–501.

Rasmussen, C. E. and C. Williams (2006). *Gaussian Processes for Machine Learning*. MIT Press.