Exact Hamiltonian Monte Carlo for Truncated Multivariate Gaussians

Ari Pakman* and Liam Paninski†

Department of Statistics and Center for Theoretical Neuroscience
Columbia University

Abstract

We present a Hamiltonian Monte Carlo algorithm to sample from multivariate Gaussian distributions in which the target space is constrained by linear and quadratic inequalities or products thereof. The Hamiltonian equations of motion can be integrated exactly and there are no parameters to tune. The algorithm mixes fast and outperforms Gibbs sampling for constraint geometries that impose strong correlations among the variables. The runtime scales linearly with the number of constraints but the algorithm is highly parallelizable. A simple extension of the algorithm permits sampling from distributions whose log-density is piecewise quadratic, as in the "Bayesian lasso" model.

Keywords: Markov Chain Monte Carlo, Hamiltonian Monte Carlo, Truncated Multivariate Gaussians.

*ari@stat.columbia.edu

†liam@stat.columbia.edu

1 Introduction

The advent of Markov Chain Monte Carlo methods has made it possible to sample from complex multivariate probability distributions [Robert and Casella, 2004], leading to a remarkable progress in Bayesian modeling, with applications to many areas of applied statistics and machine learning [Gelman et al., 2004].

In many cases, the data or the parameter space are constrained [Gelfand et al., 1992] and the need arises for efficient sampling techniques for truncated distributions. In this paper we will focus on the Truncated Multivariate Gaussian (TMG), a d-dimensional multivariate Gaussian distribution of the form

$$\log p(\mathbf{X}) \propto -\frac{1}{2}\mathbf{X}^T\mathbf{M}\mathbf{X} + \mathbf{r}^T\mathbf{X} \tag{1.1}$$

with $\mathbf{X}$, $\mathbf{r} \in \mathbb{R}^d$ and $\mathbf{M}$ positive definite, subject to m inequalities

$$Q_j(\mathbf{X}) \geq 0 \qquad j = 1, \ldots, m\,, \tag{1.2}$$

where $Q_j(\mathbf{X})$ is a product of linear and quadratic polynomials. These distributions play a central role in models as diverse as the Probit and Tobit models [Albert and Chib, 1993, Tobin, 1958], the dichotomized Gaussian model [Emrich and Piedmonte, 1991, Cox and Wermuth, 2002], stochastic integrate-and-fire neural models [Paninski et al., 2004], Bayesian isotonic regression [Neelon and Dunson, 2004], the Bayesian bridge model expressed as a mixture of Bartlett-Fejer kernels [Polson and Scott, 2011], and many others.

The standard approach to sample from TMGs is the Gibbs sampler [Geweke, 1991], which reduces the problem to one-dimensional truncated Gaussians, for which simple and efficient sampling methods exist [Robert, 1995, Damien and Walker, 2001]. While it enjoys the benefit of having no parameters to tune, the Gibbs sampler can suffer from two problems, which make it inefficient in some situations. Firstly, its run-time scales linearly with the number of dimensions. Secondly, even though a change of variables that maps $\mathbf{M}$ in (1.1) to the identity often improves the mixing speed [Rodriguez-Yam et al., 2004], the exploration of the target space can still be very slow when the constraints (1.2) impose high correlations among the coordinates. Figure 1 illustrates this effect in a simple example. Improvement over the Gibbs run-time can be obtained with a hit-and-run algorithm [Chen and Deely, 1992], but the latter suffers from the same slow convergence problem when the constraints impose strong correlations.

In this paper we present an alternative algorithm to sample from TMG distributions for constraints $Q_j(\mathbf{X})$ in (1.2) given by linear or quadratic functions or products thereof, based on the Hamiltonian Monte Carlo (HMC) approach. The HMC method, introduced in Duane et al. [1987], considers the log of the probability distribution as minus the potential energy of a particle, and introduces a Gaussian distribution for momentum variables in order to define a Hamiltonian function. The method generally avoids random walks and mixes faster than Gibbs or Metropolis-Hastings techniques. The HMC sampling procedure alternates between sampling the Gaussian momenta and letting the position of the particle evolve by integrating its Hamiltonian equations of motion. In most models, the latter cannot be integrated exactly, so the resulting position is used as a Metropolis proposal, with an acceptance probability that depends exponentially on the energy gained due to the numerical error. The downside is that two parameters must be fine-tuned for the algorithm to work properly: the integration time-step size and the number of time-steps. In general the values selected correspond to a compromise between a high acceptance rate and a good rate of exploration of the space [Hoffman and Gelman, 2011]. More details of HMC can be found in the reviews by Kennedy [1990] and Neal [2010].

The case we consider in this work is special because the Hamiltonian equations of motion can be integrated exactly, thus leading to the best of both worlds: HMC mixes fast and, as in Gibbs, there are no parameters to tune and the Metropolis step always accepts (because the energy is conserved exactly). The truncations (1.2) are incorporated via hard walls, against which the particle bounces off elastically. The run-time depends highly on the shape and location of the truncation, as most of the computing time goes into finding the time of the next wall bounce and the direction of the reflected particle. But unlike the Gibbs sampler, these computations are parallelizable, potentially allowing fast implementations.

The discontinuity that a particle experiences when bouncing off a constraint wall is similar when the log-density is piecewise quadratic. We show that a simple extension of the algorithm allows us to sample from such distributions, focusing on the example of the "Bayesian lasso" model [Park and Casella, 2008].

Previous HMC applications that made use of exactly solvable Hamiltonian equations include sampling from non-trivial integrable Hamiltonians [Kennedy and Bitar, 1994], and importance sampling, with the target distribution approximated by a distribution with an integrable Hamiltonian [Rasmussen, 2003, Izaguirre and Hampton, 2004].

In the next section we present the new sampling algorithm for linear and quadratic constraints along with two example applications; in Section 3 we present the extension to the Bayesian lasso model. We have implemented the sampling algorithm in the R package "tmg."

2 The Sampling Algorithm

2.1 Linear Inequalities

Consider first sampling from

$$\log p(\mathbf{X}) \propto -\frac{1}{2}\mathbf{X}\cdot\mathbf{X} \tag{2.1}$$

subject to

$$\mathbf{F}_j\cdot\mathbf{X} + g_j \geq 0 \qquad j = 1,\ldots,m\,. \tag{2.2}$$

Any quadratic form for $\log p(\mathbf{X})$, as in (1.1), can be brought to the above canonical form by a linear change of variables. Let us denote the components of $\mathbf{X}$ and $\mathbf{F}_j$ as

$$\begin{aligned} \mathbf{X} &= (x_1,\ldots,x_d)\,, & (2.3)\\ \mathbf{F}_j &= (f_j^1,\ldots,f_j^d)\,. & (2.4) \end{aligned}$$

In order to apply the HMC method, we introduce momentum variables $\mathbf{\Pi}$,

$$\mathbf{\Pi} = (\pi^1,\ldots,\pi^d)\,, \tag{2.5}$$

and consider the Hamiltonian

$$H = \frac{1}{2}\mathbf{X}\cdot\mathbf{X} + \frac{1}{2}\mathbf{\Pi}\cdot\mathbf{\Pi}\,, \tag{2.6}$$

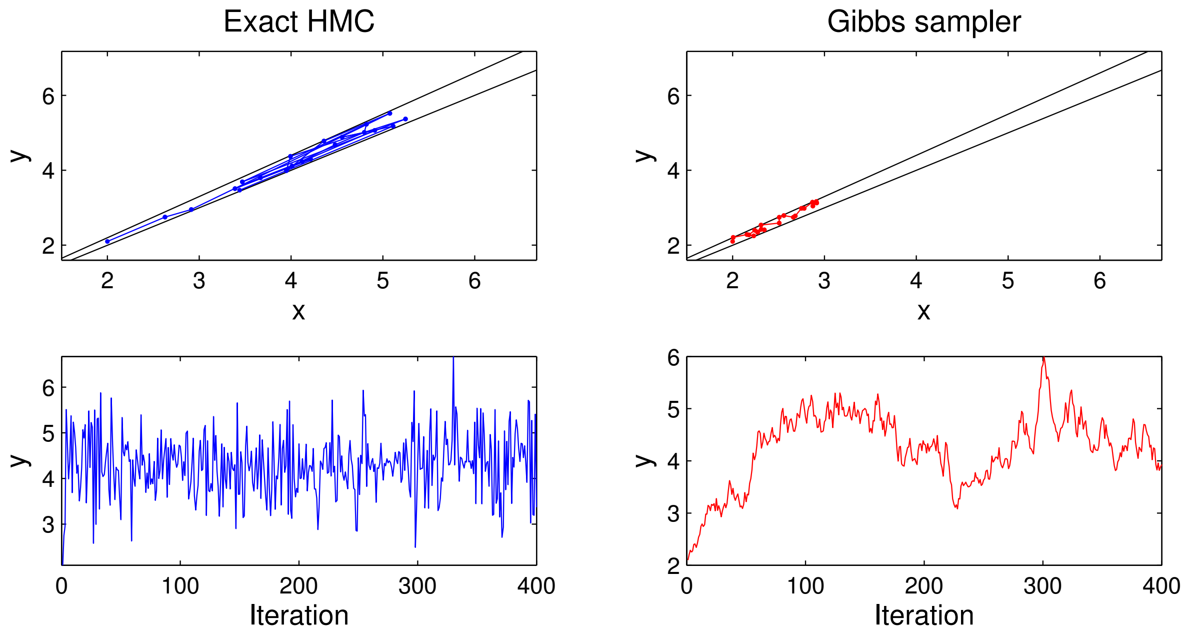

Figure 1: HMC vs Gibbs sampler. *Comparison for a two-dimensional distribution with* $\log p(x,y) \propto -\frac{1}{2}(x-4)^2 - \frac{1}{2}(y-4)^2$, *constrained to the wedge* $x \le y \le 1.1x$ *and* $x, y \ge 0$. *The initial point is* $(x,y) = (2, 2.1)$. *Upper panels: first 20 iterations. Lower panels: Second coordinate of the first 400 iterations. The exact HMC sampler moves rapidly to oscillate around* $y = 4$, *as desired, while the Gibbs sampler mixes relatively slowly.*

such that the joint distribution is $p(\mathbf{X}, \mathbf{\Pi}) = \exp(-H)$. The equations of motion following from (2.6) are

$$\dot{x}_i = \frac{\partial H}{\partial \pi^i} = \pi^i \tag{2.7}$$

$$\dot{\pi}^i = -\frac{\partial H}{\partial x_i} = -x_i \qquad i = 1, \ldots, d \tag{2.8}$$

which can be combined to

$$\ddot{x}_i = -x_i\,, \tag{2.9}$$

and have a solution

$$x_i(t) = a_i \sin(t) + b_i \cos(t)\,. \tag{2.10}$$

The constants a_i, b_i can be expressed in terms of the initial conditions as

$$b_i = x_i(0) \tag{2.11}$$

$$a_i = \dot{x}_i(0) = \pi_i(0) \tag{2.12}$$

The HMC algorithm proceeds by alternating between two steps. In the first step we sample $\mathbf{\Pi}$ from $p(\mathbf{\Pi}|\mathbf{X}) = p(\mathbf{\Pi}) = \mathcal{N}(0, \mathbb{I}_d)$. In the second step we use this $\mathbf{\Pi}$ and the last value of $\mathbf{X}$ as initial conditions, and let the particle move during a time T. The value of $\mathbf{X}$ at the end of the trajectory belongs to a Markov chain with equilibrium distribution $p(\mathbf{X})$. The value of T is arbitrary; one reasonable approach would be to choose the T that leads to the farthest final distance from the starting point, and thus to a fast exploration of the target space [Hoffman and Gelman, 2011]. However, this value is not easy to compute when the particle is being reflected off of many walls. Therefore we simply sample T uniformly from $[0, \pi]$ at each iteration (recall that the unconstrained trajectories are sinusoidal with period 2π). Another option would be to just set $T = \pi$.

The trajectory of the particle is given by (2.10) until it hits a wall, and this occurs whenever any of the inequalities (2.2) is saturated. To find the time at which this occurs, it is convenient to define

$$\begin{aligned} K_j(t) &= \sum_{i=1}^{d} f_j^i x_i(t) + g_j \qquad j = 1, \ldots, m\,. && (2.13)\\ &= \sum_{i=1}^{d} f_j^i a_i \sin(t) + \sum_{i=1}^{d} f_j^i b_i \cos(t) + g_j && (2.14)\\ &= u_j \cos(t + \varphi_j) + g_j && (2.15)\end{aligned}$$

where

$$\begin{aligned} u_j &= \sqrt{(\sum_{i=1}^{d} f_j^i a_i)^2 + (\sum_{i=1}^{d} f_j^i b_i)^2}\,, && (2.16)\\ \tan\varphi_j &= -\frac{\sum_{i=1}^{d} f_j^i a_i}{\sum_{i=1}^{d} f_j^i b_i}\,. && (2.17)\end{aligned}$$

Along the trajectory we have $K_j(t) \geq 0$ for all j and a wall hit corresponds to $K_j(t) = 0$, so from (2.15) it follows that the particle can only reach those walls satisfying $u_j > |g_j|$. Each one of those reachable walls has associated two times $t_j > 0$ such that

$$K_j(t_j) = 0\,, \qquad (2.18)$$

and the actual wall hit corresponds to the smallest of all these times. Suppose that the latter occurs for $j = h$. At the hitting point, the particle bounces off the wall and the trajectory continues with a reflected velocity. The latter can be obtained by noting that the vector $\mathbf{F}_h$ is perpendicular to the reflecting plane. Let us decompose the velocity as

$$\dot{\mathbf{X}}(t_h) = \dot{\mathbf{X}}_\perp(t_h) + \alpha_h \mathbf{F}_h\,, \qquad (2.19)$$

where $\mathbf{F}_h \cdot \dot{\mathbf{X}}_\perp(t_h) = 0$ and

$$\alpha_h = \frac{\mathbf{F}_h \cdot \dot{\mathbf{X}}(t_h)}{||\mathbf{F}_h||^2}\,. \qquad (2.20)$$

The reflected velocity, $\dot{\mathbf{X}}_R(t_h)$, is obtained by inverting the component perpendicular to the reflecting plane

$$\begin{aligned} \dot{\mathbf{X}}_R(t_h) &= \dot{\mathbf{X}}_\perp(t_h) - \alpha_h \mathbf{F}_h\,, && (2.21)\\ &= \dot{\mathbf{X}}(t_h) - 2\alpha_h \mathbf{F}_h\,. && (2.22)\end{aligned}$$

It is easy to verify that this transformation leaves the Hamiltonian (2.6) invariant. Once the reflected velocity is computed, we use it as an initial condition in (2.12) to continue the particle trajectory.

The run-time of each iteration scales linearly with m, since we have to compute the m values u_j, and, when $u_j > |g_j|$, we need to compute φ_j and t_j, defined in (2.17) and (2.18). (Of course, the total runtime is also proportional to the number of times the particle hits the wall per iteration, which varies according to the shape and location of the walls.) The dominant cost is in the computation of the sums in expressions (2.16) and (2.17): these can be interpreted as matrix-vector multiplications, with cost $O(md)$ for general constraint matrices $\mathbf{F} = (\mathbf{F}_1^T \mathbf{F}_2^T \dots \mathbf{F}_m^T)^T$. Note that these matrix-vector multiplications are highly parallelizable. In addition, in many cases there may be some special structure that can be exploited to speed computation further; for example, if $\mathbf{F}$ can be expressed as a sparse matrix in a convenient basis, this cost can be reduced to $O(d)$. Note that the transformation of a general quadratic form for $\log p(\mathbf{X})$, as in (1.1), to the canonical form (2.1) is not always computationally efficient, because the constraints also change under the transformation and a sparse constraint in the original frame may became dense in the whitened frame. This situation occurs in the examples below and in Section 3. For these cases, it can be convenient to keep the original distribution in the form (1.1) and consider the Hamiltonian

$$H = \frac{1}{2}\mathbf{X}^T \mathbf{M} \mathbf{X} - \mathbf{r}^T \mathbf{X} + \frac{1}{2}\mathbf{\Pi}^T \mathbf{M}^{-1} \mathbf{\Pi} \,. \tag{2.23}$$

As we show in Section 3, such a mass matrix for the momenta also leads to independent trigonometric solutions for each coordinate. But the time saved in the fast evaluation of the constraints has a trade-off in that now, at each iteration, the momenta $\mathbf{\Pi}$ should be sampled from the distribution $\mathcal{N}(0, \mathbf{M})$, with a non-trivial covariance. Again, it is often possible to exploit structure in $\mathbf{M}$ to speed up computation (e.g., via specialized Cholesky decomposition approaches).

2.2 Quadratic and Higher Order Inequalities

The sampling algorithm can be extended in principle to polynomial constraints of the form

$$Q_j(\mathbf{X}) \geq 0 \qquad j = 1, \dots, m \,. \tag{2.24}$$

Evaluating $Q_j(\mathbf{X})$ at the solution (2.10) leads to a polynomial in $\sin(t)$ and $\cos(t)$, whose zeros must be found in order to find the hitting times. When a wall is hit, we reflect the velocity by inverting the sign of the component perpendicular to the wall, given by the gradient $\nabla Q_j(\mathbf{X})$. This vector plays a role similar to $\mathbf{F}_j$ in (2.19)-(2.22). Of course, for general polynomials $Q_j(\mathbf{X})$ computing the hitting times might be numerically challenging.

One family of solvable constraints involves quadratic inequalities of the form

$$Q_j(\mathbf{X}) = \mathbf{X}^T \mathbf{A}_j \mathbf{X} + \mathbf{X} \cdot \mathbf{B}_j + C_j \geq 0 \qquad j = 1, \dots, m \,, \tag{2.25}$$

where $\mathbf{A}_j \in \mathbb{R}^{d,d}$, $\mathbf{B}_j \in \mathbb{R}^d$, $C_j \in \mathbb{R}$. For statistics applications where these constraints are important, see e.g. Ellis and Maitra [2007]. Inserting (2.10) in the equality for (2.25) leads, for each j, to the following equation for the hitting time:

$$q_1 \cos^2(t) + q_2 \cos(t) + q_3 = -\sin(t)(q_4 \cos(t) + q_5) \,, \tag{2.26}$$

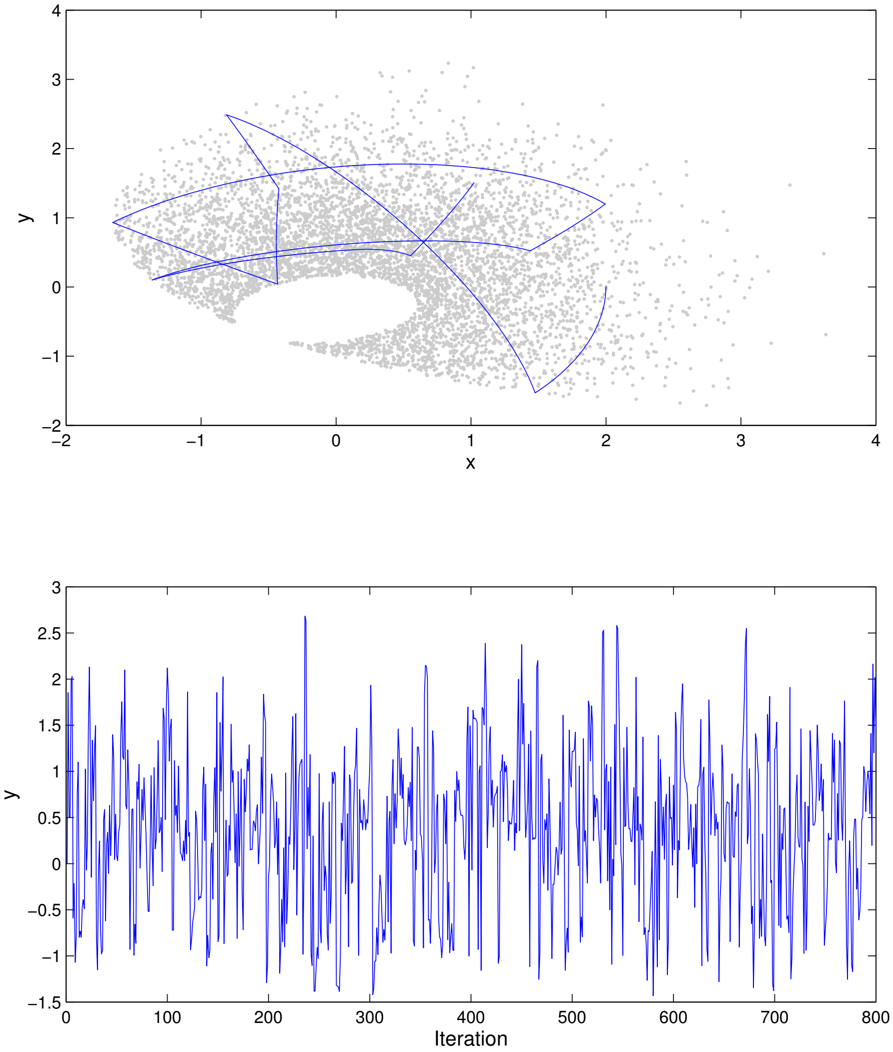

Figure 2: Truncation by quadratic inequalities. *Above: 6000 samples of a two-dimensional canonical normal distribution, constrained by the quadratic inequalities (2.41) -(2.42). The piecewise elliptic curve shows the trajectory of the particle in the first iterations, with starting point* $(x, y) = (2, 0)$*. Below: first* 800 *iterations of the vertical coordinate. For the algebraic solution of (2.35), we used the C++ code from the DynamO package [Bannerman et al., 2011].*

with

$$q_1 = \sum_{i,k} A^{ik} b_i b_k - \sum_{ik} A^{ik} a_i a_k \,, \tag{2.27}$$

$$q_2 = \sum_i B^i b_i \,, \tag{2.28}$$

$$q_3 = C + \sum_{ik} A^{ik} a_i a_k \,, \tag{2.29}$$

$$q_4 = 2 \sum_{i,k} A^{ik} a_i b_k \,, \tag{2.30}$$

$$q_5 = \sum_i B^i a_i \tag{2.31}$$

and we omitted the j dependence to simplify the notation. If the ellipse in (2.25) is centered at the origin, we have $\mathbf{B}_j = q_2 = q_5 = 0$, and equation (2.26) simplifies to

$$q_1 + 2q_3 + u\sin(2t + \varphi) = 0 \tag{2.32}$$

where

$$\begin{aligned} u^2 &= q_1^2 + q_4^2\,, & (2.33)\\ \tan\varphi &= \frac{q_1}{q_4}\,, & (2.34) \end{aligned}$$

and the hit time can be found from (2.32) as in the linear case. In the general $\mathbf{B}_j \neq 0$ case, the square of (2.26) gives the quartic equation

$$r_4\cos^4(t) + r_3\cos^3(t) + r_2\cos^2(t) + r_1\cos(t) + r_0 = 0\,, \tag{2.35}$$

where

$$\begin{aligned} r_4 &= q_1^2 + q_4^2\,, & (2.36)\\ r_3 &= 2q_1q_2 + 2q_4q_5\,, & (2.37)\\ r_2 &= q_2^2 + 2q_1q_3 + q_5^2 - q_4^2\,, & (2.38)\\ r_1 &= 2q_2q_3 - 2q_4q_5\,, & (2.39)\\ r_0 &= q_3^2 - q_5^2\,. & (2.40) \end{aligned}$$

Equation (2.35) can be solved exactly using standard algebraic methods [Herbison-Evans, 1994]. A wall hit corresponds, among all the constraints j, to the solution for $\cos(t)$ with smallest $t > 0$ and $|\cos(t)| \leq 1$, which also solves (2.26). As an example, Figure 2 shows samples from a two-dimensional canonical normal distribution, constrained by

$$\begin{aligned} \frac{(x-4)^2}{32} + \frac{(y-1)^2}{8} &\leq 1\,, & (2.41)\\ 4x^2 + 8y^2 - 2xy + 5y &\geq 1\,. & (2.42) \end{aligned}$$

Equipped with the results for linear and quadratic constraints, we can also find the hitting times for constraints of the form

$$Q(\mathbf{X}) = \prod_j Q_j(\mathbf{X}) \geq 0 \tag{2.43}$$

where each $Q_j(\mathbf{X})$ is a linear or a quadratic function. Each factor defines an equation as (2.18) or (2.35), and the hitting time is the smallest at which any factor becomes zero. For other polynomials, one has to resort to numerical methods to find the hitting times.

2.3 Example: Probit and Tobit Models

The Probit model is a popular discriminative probabilistic model for binary classification with continuous inputs [Albert and Chib, 1993]. The conditional probabilities for the binary labels

$y = \pm 1$ are given by

$$\begin{aligned} p(y=-1|\mathbf{z},\boldsymbol{\beta}) &= \Phi(\mathbf{z}\cdot\boldsymbol{\beta}) = \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{\mathbf{z}\cdot\boldsymbol{\beta}} dw\, e^{-\frac{w^2}{2}} && (2.44)\\ &= \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{0} dw\, e^{-\frac{(w+\mathbf{z}\cdot\boldsymbol{\beta})^2}{2}} && (2.45)\\ p(y=+1|\mathbf{z},\boldsymbol{\beta}) &= 1-\Phi(\mathbf{z}\cdot\boldsymbol{\beta}) && (2.46)\\ &= \frac{1}{\sqrt{2\pi}}\int_{0}^{+\infty} dw\, e^{-\frac{(w+\mathbf{z}\cdot\boldsymbol{\beta})^2}{2}} && (2.47)\end{aligned}$$

where $\mathbf{z} \in \mathbb{R}^p$ is a vector of regressors and $\boldsymbol{\beta} \in \mathbb{R}^p$ are the parameters of the model. Given N pairs of labels and regressors

$$\begin{aligned} \mathbf{Y} &= (y_1,\dots,y_N)\,, && (2.48)\\ \mathbf{Z} &= (\mathbf{z}_1,\dots,\mathbf{z}_N)\,, && (2.49)\end{aligned}$$

the posterior distribution of the parameters $\boldsymbol{\beta}$ is

$$\begin{aligned} p(\boldsymbol{\beta}|\mathbf{Y},\mathbf{Z}) &\propto p(\boldsymbol{\beta})\prod_{i=1}^{N} p(y_i|\mathbf{z}_i,\boldsymbol{\beta}) && (2.50)\\ &\propto p(\boldsymbol{\beta})\int_{y_i w_i \geq 0} dw_1 \dots dw_N\; e^{-\frac{1}{2}\sum_{i=1}^{N}(w_i+\mathbf{z}_i\cdot\boldsymbol{\beta})^2} \qquad i=1\dots N && (2.51)\end{aligned}$$

where $p(\boldsymbol{\beta})$ is the prior distribution. The likelihood $p(y_i|\mathbf{z}_i,\boldsymbol{\beta})$ corresponds to a model

$$\begin{aligned} y_i &= sign(w_i) && (2.52)\\ w_i &= -\mathbf{z}_i\cdot\boldsymbol{\beta}+\varepsilon_i && (2.53)\\ \varepsilon_i &\sim \mathcal{N}(0,1) && (2.54)\end{aligned}$$

in which only the sign of w_i is observed, but not its value. Assuming a Gaussian prior with zero mean and covariance $\sigma^2\mathbb{I}_p$, expression (2.51) is the marginal distribution of a multivariate Gaussian on $(\boldsymbol{\beta}, w_1, \dots w_N)$, truncated to $y_i w_i \geq 0$ for $i = 1,\dots,N$. The untruncated Gaussian has zero mean and precision matrix

$$M = \begin{pmatrix} M_{ww} & M_{w\boldsymbol{\beta}} \\ M_{\boldsymbol{\beta}w} & M_{\boldsymbol{\beta}\boldsymbol{\beta}} \end{pmatrix} \qquad \in \mathbb{R}^{N+p,N+p} \qquad (2.55)$$

where

$$\begin{aligned} M_{ww} &= \mathbb{I}_N \qquad \in \mathbb{R}^{N,N} && (2.56)\\ M_{w\boldsymbol{\beta}} &= M^T_{\boldsymbol{\beta}w} = \begin{pmatrix}\mathbf{z}_1\\ \vdots \\ \mathbf{z}_N\end{pmatrix} \qquad \in \mathbb{R}^{N,p} && (2.57)\\ M_{\boldsymbol{\beta}\boldsymbol{\beta}} &= \sigma^{-2}\mathbb{I}_p + \sum_{i=1}^{N}\mathbf{z}_i\mathbf{z}_i^T \qquad \in \mathbb{R}^{p,p} && (2.58)\end{aligned}$$

We can sample from the posterior in (2.51) by sampling from the truncated Gaussian for $(\boldsymbol{\beta}, w_1, \dots w_N)$ and keeping only the $\boldsymbol{\beta}$ values. It is easy to show that without the first term in (2.58), coming from

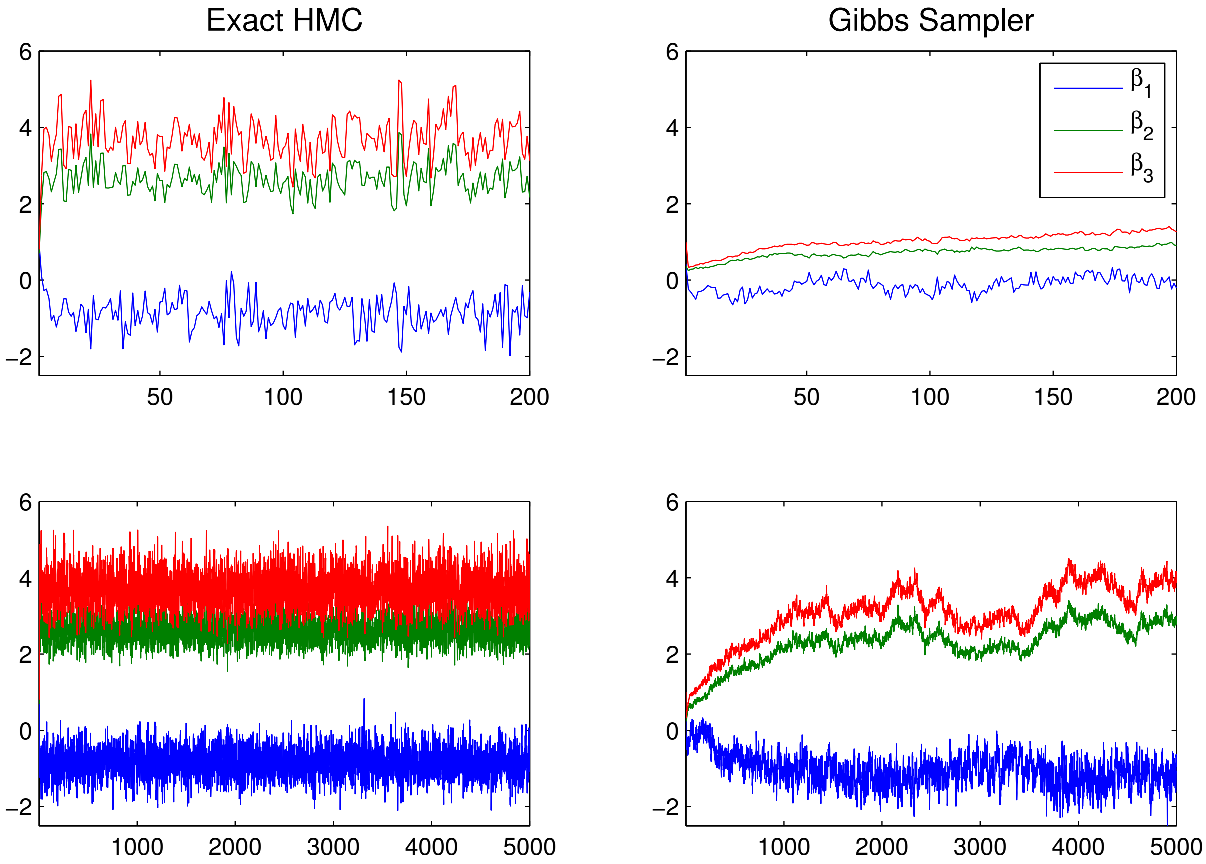

Figure 3: Bayesian Probit model. *First 200 and 5000 samples from the posterior (2.51) of a model with $p = 3$. $N = 800$ pairs (y_i, z_i) were generated with $\beta_1 = -9, \beta_2 = 20, \beta_3 = 27$ and we assumed a Gaussian prior with zero mean and $\sigma^2 = 10^5$. Note that the means of the sampled values are different from the values used to generate the data, due to the zero-mean prior. Left: Exact HMC sampler. Right: Gibbs sampler, with whitened covariance to improve mixing [Rodriguez-Yam et al., 2004].*

the prior $p(\boldsymbol{\beta})$, the precision matrix would have p null directions and our method would not be applicable, since we assume the precision matrix to be positive definite. Note that the dimension of the TMG grows linearly with the number N of data points. As an illustration, Figure 3 shows the values of $\boldsymbol{\beta}$, sampled using Gibbs and exact HMC, from the posterior of a model with $p = 3$ where $N = 800$ data points were generated. We used $z_i^1 = 1$, $z_i^2 \sim Unif[-5, +5]$ and $z_i^3 \sim \mathcal{N}(-4, \sigma = 4)$. The values of y_i were generated with $\beta_1 = -9, \beta_2 = 20, \beta_3 = 27$ and we assumed a Gaussian prior with $\sigma^2 = 10^5$. Note that the means of the sampled β_i's are different from the β_i's used to generate the data, due to the prior which pulls the β_i's towards zero. For both samplers, we made a coordinate rotation to the canonical frame in which the unconstrained Gaussian has unit covariance. This transformation changes the constraint surface and imposes correlations that make Gibbs mix slowly, as shown in Figure 3. One can similarly consider the multivariate Probit model [Ashford and Sowden, 1970], where the Bayesian approach has been shown to be superior to Maximum Likelihood [Geweke et al., 1994].

A related model is the Tobit model for censored data [Tobin, 1958], which is a linear regression model where negative values are not observed:

$$y_i = \begin{cases} y_i^* & \text{for}\, y_i^* > 0\,, \\ 0 & \text{for}\, y_i^* \leq 0\,, \end{cases} \tag{2.59}$$

where

$$y_i^* = \mathbf{z}_i \cdot \boldsymbol{\beta} + \varepsilon_i\,, \qquad \varepsilon_i \sim \mathcal{N}(0,\sigma). \tag{2.60}$$

The likelihood of a pair $(y_i, \mathbf{z}_i)$ is

$$p(y_i|\mathbf{z}_i, \boldsymbol{\beta}, \sigma) = \begin{cases} \dfrac{e^{-\frac{(y_i - \mathbf{z}_i\cdot\boldsymbol{\beta})^2}{2\sigma^2}}}{\sqrt{2\pi\sigma^2}} & \text{for}\, y_i > 0\,, \\ \frac{1}{\sqrt{2\pi\sigma^2}} \int_{-\infty}^{0} dw_i e^{-\frac{(w_i - \mathbf{z}_i\cdot\boldsymbol{\beta})^2}{2\sigma^2}} & \text{for}\, y_i = 0\,, \end{cases} \tag{2.61}$$

and the posterior probability for $\boldsymbol{\beta}$ is

$$\begin{aligned} p(\boldsymbol{\beta}|\mathbf{Y}, \mathbf{Z}, \sigma) &\propto p(\boldsymbol{\beta}|\sigma) \prod_{i=1}^{N} p(y_i|\mathbf{z}_i, \boldsymbol{\beta}, \sigma) & (2.62) \\ &\propto p(\boldsymbol{\beta}|\sigma) \prod_{i, y_i > 0} e^{-\frac{(y_i - \mathbf{z}_i\cdot\boldsymbol{\beta})^2}{2\sigma^2}} \prod_{i, y_i = 0} \int_{-\infty}^{0} dw_i\, e^{-\frac{(w_i - \mathbf{z}_i\cdot\boldsymbol{\beta})^2}{2\sigma^2}} & (2.63) \end{aligned}$$

As in (2.51), this can be treated as a marginal distribution over the variables w_i, with the joint distribution for $(\boldsymbol{\beta}, w_i)$ a truncated multivariate Gaussian.

2.4 Example: Bayesian splines for positive functions

Suppose we have noisy samples (y_i, x_i), $i = 1 \ldots N$, from an unknown smooth positive function $f(x) > 0$, with $x \in [0, h]$. We can estimate $f(x)$ using cubic splines with knots at the x_i's, plus 0 and h [Green and Silverman, 1994]. The dimension of the vector space of cubic splines with N inner knots is $N + 4$. Our model is thus

$$y_i = \sum_{s=1}^{N+4} a_s \phi_s(x_i) + \varepsilon_i \quad \varepsilon_i \sim \mathcal{N}(0,\sigma) \qquad i = 1 \ldots N\,, \tag{2.64}$$

where the functions $\phi_s(x)$ are a spline basis. Suppose we are interested in the value of $f(x)$ at the points $x = z_j$ with $j = 1 \ldots m$. To enforce $f(x) > 0$ at those points, we impose the constraints

$$\boldsymbol{\phi}(z_j) \cdot \mathbf{a} \geq 0\,, \qquad j = 1 \ldots m\,, \tag{2.65}$$

where

$$\begin{aligned} \boldsymbol{\phi}(x) &= (\phi_1(x), \ldots, \phi_{N+4}(x))\,, & (2.66) \\ \mathbf{a} &= (a_1, \ldots, a_{N+4})\,. & (2.67) \end{aligned}$$

To obtain a sparse constraint matrix, it is convenient to use the B-spline basis, in which only four elements in the vector $\boldsymbol{\phi}(z_j)$ are non-zero for any j (see, e.g. [De Boor, 2001] for details). In a Bayesian approach, we are interested in sampling from the posterior distribution

$$p(\mathbf{a}, \sigma^2|\mathbf{Y}, \mathbf{X}, \lambda) \propto p(\mathbf{Y}|\mathbf{X}, \mathbf{a}, \sigma^2) p(\mathbf{a}|\lambda, \sigma^2) p(\sigma^2)\,, \tag{2.68}$$

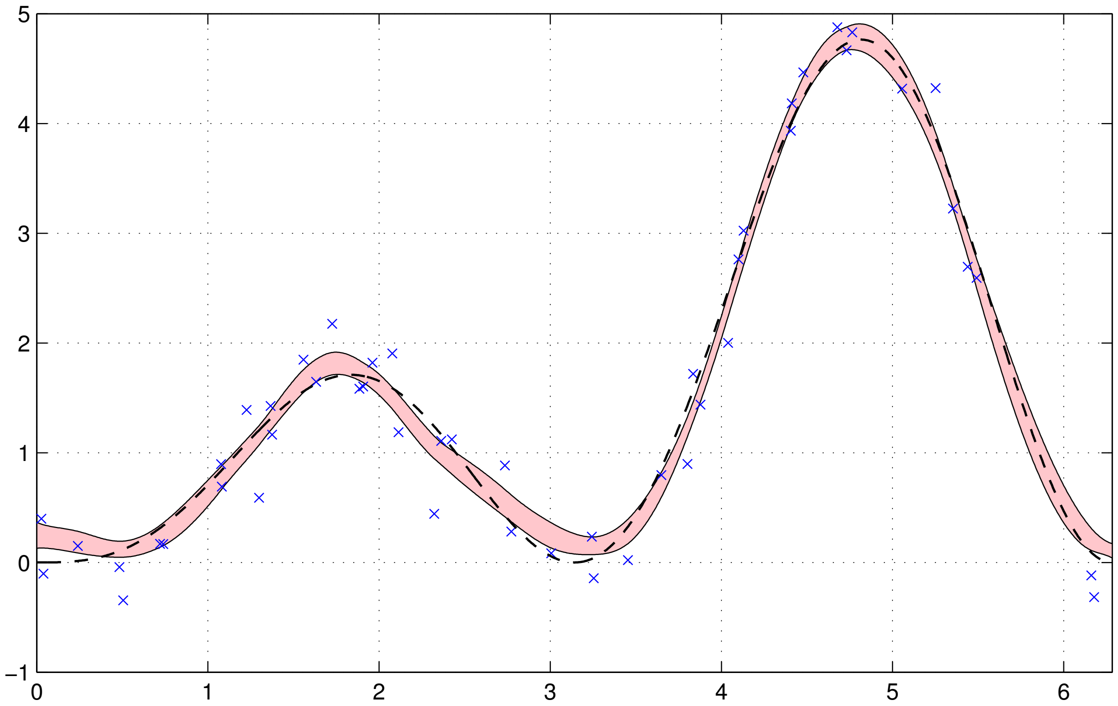

Figure 4: Bayesian splines for positive functions. *The crosses show* 50 *samples from* $y_i = x_i \sin^2(x_i) + \varepsilon_i$*, where* $\varepsilon_i \sim \mathcal{N}(0,\sigma)$ *with* $\sigma^2 = .09$*. The values of* x_i *were sampled uniformly from* $[0, 2\pi]$*. The curve* $f(x) = x \sin^2(x)$ *is shown as a dashed line. The shaded band shows the splines built with coefficients from the* .25 *and* .75 *quantiles of samples from the posterior distribution of* $\boldsymbol{a}$ *in (2.68). We used a Jeffreys prior for* σ^2 *[Jeffreys, 1946] and imposed the positivity constraints (2.65) at* 100 *points spread uniformly in* $[0, 2\pi]$*. The smoothness parameter* λ *was estimated as* $\hat{\lambda} = 0.0067$ *by maximizing the marginal likelihood (empirical Bayes criterion), using a Monte Carlo EM algorithm. The mean of the samples of* σ^2 *was* $\hat{\sigma^2} = 0.091$*. The spline computations were performed with the “fda” MATLAB package [Ramsay et al., 2009].*

where we defined

$$\begin{aligned} \mathbf{Y} &= (y_1, \ldots, y_N)\,, & (2.69)\\ \mathbf{X} &= (x_1, \ldots, x_N)\,. & (2.70) \end{aligned}$$

The likelihood is

$$p(\mathbf{Y}|\mathbf{X}, \mathbf{a}, \sigma^2) = \frac{1}{(2\pi\sigma^2)^{N/2}} \exp\left(-\frac{1}{2\sigma^2} \sum_{i=1}^{N} (y_i - \mathbf{a} \cdot \boldsymbol{\phi}(x_i))^2 \right), \qquad (2.71)$$

and for the prior on $\mathbf{a}$ we consider

$$\begin{aligned} p(\mathbf{a}|\lambda, \sigma^2) &\propto \left(\frac{\lambda}{\sigma^2}\right)^{\frac{N+4}{2}} \exp\left(-\frac{\lambda}{2\sigma^2} \int_0^h dx \left(\mathbf{a} \cdot \boldsymbol{\phi}''(x) \right)^2 \right), & (2.72)\\ &\propto \left(\frac{\lambda}{\sigma^2}\right)^{\frac{N+4}{2}} \exp\left(-\frac{\lambda}{2\sigma^2} \mathbf{a}^T \mathbf{K} \mathbf{a} \right), & (2.73) \end{aligned}$$

where $\mathbf{K} \in \mathbb{R}^{N+4,N+4}$ has entries

$$K_{sr} = \int_0^h dx\, \phi_s''(x)\phi_r''(x)\,. \tag{2.74}$$

The prior (2.72)-(2.73) is standard in the spline literature and imposes a λ-dependent penalty on the roughness of the estimated polynomial, with a bigger λ corresponding to a smoother solution. This penalty allows to avoid overfitting the data [Green and Silverman, 1994].

We can Gibbs sample from the posterior (2.68) by alternating between the conditional distributions of σ^2 and $\mathbf{a}$. The latter is a TMG with

$$\log p(\mathbf{a}|\sigma^2, \mathbf{X}, \mathbf{Y}, \lambda) \propto -\frac{1}{2\sigma^2}\mathbf{a}^T(\mathbf{M} + \lambda\mathbf{K})\mathbf{a} + \frac{1}{\sigma^2}\mathbf{a}^T \cdot \mathbf{r}\,, \qquad s = 1 \dots N+4\,, \tag{2.75}$$

constrained by (2.65), and we defined

$$\mathbf{M} = \sum_{i=1}^N \boldsymbol{\phi}(x_i)\boldsymbol{\phi}(x_i)^T \qquad \in \mathbb{R}^{N+4,N+4}\,, \tag{2.76}$$

$$\mathbf{r} = \sum_{i=1}^N y_i \boldsymbol{\phi}(x_i) \qquad \in \mathbb{R}^{N+4}\,. \tag{2.77}$$

Figure 4 shows an example for the function $f(x) = x \sin^2(x)$, with $N = 50$ points sampled as

$$y_i = x_i \sin^2(x_i) + \varepsilon_i \quad \varepsilon_i \sim \mathcal{N}(0,\sigma) \qquad \sigma^2 = .09\,, \tag{2.78}$$

and with the x_i sampled uniformly from $[0, 2\pi]$.

3 The Bayesian Lasso

The techniques introduced above can also be used to sample from multivariate distributions whose log density is piecewise quadratic, with linear or elliptical boundaries between the piecewise regions. Instead of presenting the most general case, let us elaborate the details for the example of the Bayesian lasso [Park and Casella, 2008, Hans, 2009, Polson and Scott, 2011].

We are interested in the posterior distribution of the coefficients $\boldsymbol{\beta} \in \mathbb{R}^d$ and σ^2 of a linear regression model

$$y_n = \mathbf{z}_n \cdot \boldsymbol{\beta} + \varepsilon_n \qquad \varepsilon_n \sim \mathcal{N}(0,\sigma) \qquad n = 1, \dots, N\,, \tag{3.1}$$

Defining

$$\mathbf{Y} = (y_1, \dots, y_N) \tag{3.2}$$

$$\mathbf{Z} = (\mathbf{z}_1, \dots, \mathbf{z}_N)\,, \tag{3.3}$$

we want to sample from the posterior distribution

$$p(\boldsymbol{\beta}, \sigma^2|\mathbf{Y}, \mathbf{Z}, \lambda) \propto p(\mathbf{Y}|\mathbf{Z}, \boldsymbol{\beta}, \sigma^2)p(\boldsymbol{\beta}|\lambda, \sigma^2)p(\sigma^2)\,, \tag{3.4}$$

with prior density for the coefficients

$$p(\boldsymbol{\beta}|\lambda, \sigma^2) = \left(\frac{\lambda}{2\sigma^2}\right)^d \exp\left(-\frac{\lambda}{\sigma^2}\sum_{i=1}^d |\beta_i|\right)\,. \tag{3.5}$$

This prior is called the *lasso* (for 'least absolute shrinkage and selection operator') and imposes a λ-dependent sparsening penalty in the maximum likelihood solutions for $\boldsymbol{\beta}$ [Tibshirani, 1996].

We can Gibbs sample from the posterior (3.4) by alternating between the conditional distributions of σ^2 and $\boldsymbol{\beta}$. The latter is given by

$$\begin{aligned} -\log p(\boldsymbol{\beta}|\mathbf{Y},\mathbf{Z},\sigma^2,\lambda) &= \frac{1}{2\sigma^2}\sum_{n=1}^{N}(\mathbf{z}_n\cdot\boldsymbol{\beta}-y_n)^2+\frac{\lambda}{\sigma^2}\sum_{i=1}^{d}|\beta_i| && (3.6)\\ &\propto \frac{1}{2\sigma^2}\boldsymbol{\beta}^T\mathbf{M}\boldsymbol{\beta}-\frac{1}{\sigma^2}\sum_{i=1}^{d}L^i(s_i)\beta_i && (3.7)\end{aligned}$$

where we defined

$$\begin{aligned} \mathbf{M} &= \sum_{n=1}^{N}\mathbf{z}_n\mathbf{z}_n^T \qquad \in\mathbb{R}^{d\times d} && (3.8)\\ L^i(s_i) &= \sum_{n=1}^{N}y_n(z_i)_n-\lambda s_i \qquad i=1,\dots,d. && (3.9)\end{aligned}$$

with

$$s_i=\operatorname{sign}(\beta_i)\,. \tag{3.10}$$

Sampling $\boldsymbol{\beta}$ from (3.7) was considered previously via Gibbs sampling, either expressing the Laplace prior (3.5) as mixtures of Gaussians [Park and Casella, 2008] or Bartlett-Fejer kernels [Polson and Scott, 2011], or directly from (3.7) [Hans, 2009].

In order to apply Hamiltonian Monte Carlo we consider the Hamiltonian

$$H=\frac{1}{2\sigma^2}\boldsymbol{\beta}^T\mathbf{M}\boldsymbol{\beta}-\frac{1}{\sigma^2}\sum_{i=1}^{d}L^i(s_i)\beta_i+\frac{\sigma^2}{2}\boldsymbol{\Pi}^T\mathbf{M}^{-1}\boldsymbol{\Pi}\,. \tag{3.11}$$

Note that we did not map the coordinates to a canonical frame, as in Section 2.1. Instead, we chose a momenta mass matrix $\sigma^{-2}\mathbf{M}$, which is equal to the precision matrix of the coordinates. This choice leads to the simple equations

$$\ddot{\beta}_i=-\beta_i+\mu_i(\mathbf{s})\,, \tag{3.12}$$

where

$$\mu_i(\mathbf{s})=\sum_{j=1}^{d}M_{ij}^{-1}L^j(s_j)\,. \tag{3.13}$$

The solution to (3.12) is

$$\begin{aligned} \beta_i(t) &= \mu_i(\mathbf{s})+a_i\sin(t)+b_i\cos(t)\,, && (3.14)\\ &= \mu_i(\mathbf{s})+A_i\cos(t+\varphi_i)\,, && (3.15)\end{aligned}$$

where

$$\begin{aligned} A_i &= \sqrt{a_i^2+b_i^2}\,, && (3.16)\\ \tan\varphi_i &= -\frac{a_i}{b_i}\,. && (3.17)\end{aligned}$$

The constants a_i, b_i in (3.14) can be expressed in terms of the initial conditions as

$$\begin{aligned} b_i &= \beta_i(0) - \mu_i(\mathbf{s}) & (3.18) \\ a_i &= \dot{\beta}_i(0) & (3.19) \\ &= M^{-1}_{ij} p^j(0). & (3.20) \end{aligned}$$

As in Section 2.1, we start by sampling $\mathbf{\Pi}$ from $p(\mathbf{\Pi}|\boldsymbol{\beta}) = p(\mathbf{\Pi}) = \mathcal{N}(0, \sigma^{-2}\mathbf{M})$ and let the particle move during a time T sampled uniformly from $[0, \pi]$. The trajectory of the particle is given by (3.15) until a coordinate crosses any of the $\beta_i = 0$ planes, which happens at the smallest time $t > 0$ such that

$$0 = \mu_i(\mathbf{s}) + A_i \cos(t + \varphi_i)\,, \qquad\qquad i = 1, \ldots, d. \tag{3.21}$$

(Note that had we transformed the coordinates $\boldsymbol{\beta}$ to a canonical frame, each condition here would have involved a sum of d terms; thus the parameterization we use here leads to sparser, and therefore faster, computations.) Suppose the constraint is met for $i = j$ at time $t = t_j$. At this point β_j changes sign, so the Hamiltonian (3.11) changes by replacing

$$L^j(s_j) \longrightarrow L^j(-s_j) = L^j(s_j) + 2s_j\lambda\,, \tag{3.22}$$

which in turn changes the values of $\mu_i(\mathbf{s})$'s in (3.13). Note from (3.12) that this causes a jump in $\ddot{\boldsymbol{\beta}}(t_j)$. Using the continuity of $\boldsymbol{\beta}(t_j)$, $\dot{\boldsymbol{\beta}}(t_j)$ and the updated $\mu_i(\mathbf{s})$'s, we can compute new values for a_i and b_i as in (3.18) and (3.19) to continue the trajectory for times $t > t_j$.

The piecewise linear log-density (3.6) is continuous with discontinuous derivative, but we can also consider discontinuous log-densities defined piecewise. In these cases, the velocity is not continuous across the boundary of two regions, but jumps in such a way that the total energy is conserved. The extension of the basic method to this case is straightforward. Also, combining this algorithm with the imposition of constraints of the previous section, the HMC technique can be used to sample the posterior of a lasso model with additional constraints on the β_i's, such as the tree shrinkage model [LeBlanc and Tibshirani, 1998] or the hierarchical lasso [Bien et al., 2012].

Acknowledgements

This work was supported by an NSF CAREER grant, a McKnight Scholar award, NSF grant IIS-0904353 and by the Defense Advanced Research Projects Agency (DARPA) MTO under the auspices of Dr. Jack Judy, through the Space and Naval Warfare Systems Center, Pacific Grant/Contract No. N66001-11-1-4205. AP is supported by the Swartz Foundation. We thank Matt Hoffman, Alexandro Ramirez and Eftychios Pnevmatikakis for helpful discussions.

References

J.H. Albert and S. Chib. Bayesian analysis of binary and polychotomous response data. *Journal of the American statistical Association*, pages 669–679, 1993.

JR Ashford and RR Sowden. Multi-variate probit analysis. *Biometrics*, pages 535–546, 1970.

M. N. Bannerman, R. Sargant, and L. Lue. Dynamo: a free o(n) general event-driven molecular dynamics simulator. *Journal of Computational Chemistry*, 32(15):3329–3338, 2011.

J. Bien, J. Taylor, and R. Tibshirani. A Lasso for Hierarchical Interactions. *Arxiv preprint arXiv:1205.5050*, 2012.

M.H. Chen and J. Deely. Application of a new Gibbs Hit-and-Run sampler to a constrained linear multiple regression problem. Technical report, Technical Report 92-21, Purdue University, Center for Statistical Decision Sciences and Department of Statistics, 1992.

DR Cox and N. Wermuth. On some models for multivariate binary variables parallel in complexity with the multivariate Gaussian distribution. *Biometrika*, 89(2):462–469, 2002.

P. Damien and S.G. Walker. Sampling truncated normal, beta, and gamma densities. *Journal of Computational and Graphical Statistics*, 10(2):206–215, 2001.

C. De Boor. *A practical guide to splines.* Springer Verlag, 2001.

S. Duane, A.D. Kennedy, B.J. Pendleton, and D. Roweth. Hybrid monte carlo. *Physics letters B*, 195(2):216–222, 1987.

N. Ellis and R. Maitra. Multivariate Gaussian simulation outside arbitrary ellipsoids. *Journal of Computational and Graphical Statistics*, 16(3):692–708, 2007.

L.J. Emrich and M.R. Piedmonte. A method for generating high-dimensional multivariate binary variates. *American Statistician*, pages 302–304, 1991.

A.E. Gelfand, A.F.M. Smith, and T.M. Lee. Bayesian analysis of constrained parameter and truncated data problems using Gibbs sampling. *Journal of the American Statistical Association*, pages 523–532, 1992.

A. Gelman, J.B. Carlin, H.S. Stern, and D.B. Rubin. *Bayesian data analysis.* CRC press, 2004.

J. Geweke. Efficient simulation from the multivariate normal and student-t distributions subject to linear constraints and the evaluation of constraint probabilities. In *Computing Science and Statistics: Proceedings of the 23rd Symposium on the Interface*, pages 571–578, 1991.

J. Geweke, M. Keane, and D. Runkle. Alternative computational approaches to inference in the multinomial probit model. *The review of economics and statistics*, pages 609–632, 1994.

P.J. Green and B.W. Silverman. *Nonparametric regression and generalized linear models: a roughness penalty approach*, volume 58. Chapman & Hall/CRC, 1994.

C. Hans. Bayesian lasso regression. *Biometrika*, 96(4):835–845, 2009.

D. Herbison-Evans. Solving quartics and cubics for graphics. Technical report, Technical Report TR-94-487, Basser Department of Computer Science, University of Sidney, Sidney, Australia, 1994.

M.D. Hoffman and A. Gelman. The No-U-Turn sampler: adaptively setting path lengths in Hamiltonian Monte Carlo. *Arxiv preprint arXiv:1111.4246*, 2011.

J.A. Izaguirre and S.S. Hampton. Shadow hybrid Monte Carlo: an efficient propagator in phase space of macromolecules. *Journal of Computational Physics*, 200(2):581–604, 2004.

H. Jeffreys. An invariant form for the prior probability in estimation problems. *Proceedings of the Royal Society of London. Series A, Mathematical and Physical Sciences*, 186(1007):pp. 453–461, 1946.

AD Kennedy. The theory of hybrid stochastic algorithms. In *NATO ASIB Proc. 224: Probabilistic Methods in Quantum Field Theory and Quantum Gravity*, volume 1, page 209, 1990.

AD Kennedy and KM Bitar. An exact Local Hybrid Monte Carlo algorithm for gauge theories. *Nuclear Physics B-Proceedings Supplements*, 34:786–788, 1994.

M. LeBlanc and R. Tibshirani. Monotone shrinkage of trees. *Journal of Computational and Graphical Statistics*, pages 417–433, 1998.

R.M. Neal. MCMC using Hamiltonian dynamics. *Handbook of Markov Chain Monte Carlo*, 54: 113–162, 2010.

B. Neelon and D.B. Dunson. Bayesian isotonic regression and trend analysis. *Biometrics*, 60(2): 398–406, 2004.

L. Paninski, J.W. Pillow, and E.P. Simoncelli. Maximum likelihood estimation of a stochastic integrate-and-fire neural encoding model. *Neural computation*, 16(12):2533–2561, 2004.

T. Park and G. Casella. The Bayesian lasso. *Journal of the American Statistical Association*, 103 (482):681–686, 2008.

N.G. Polson and J.G. Scott. The Bayesian Bridge. *Arxiv preprint arXiv:1109.2279*, 2011.

J.O. Ramsay, G. Hooker, and S. Graves. *Functional data analysis with R and MATLAB*. Springer Verlag, 2009.

C.E. Rasmussen. Gaussian processes to speed up Hybrid Monte Carlo for expensive Bayesian integrals. In *Bayesian Statistics 7: Proceedings of the 7th Valencia International Meeting*, pages 651–659. Oxford University Press, 2003.

C.P. Robert. Simulation of truncated normal variables. *Statistics and computing*, 5(2):121–125, 1995.

C.P. Robert and G. Casella. *Monte Carlo statistical methods*. Springer Verlag, 2004.

G. Rodriguez-Yam, R.A. Davis, and L.L. Scharf. Efficient Gibbs sampling of truncated multivariate normal with application to constrained linear regression. *Unpublished Manuscript*, 2004. `http://www.stat.columbia.edu/~rdavis/papers/CLR.pdf`.

R. Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B*, 58:267–288, 1996.

J. Tobin. Estimation of relationships for limited dependent variables. *Econometrica: Journal of the Econometric Society*, pages 24–36, 1958.