Extrinsic Prior for Simple and Efficient Bayesian Modeling with Parameter Constraints

Abstract: Parameter constraints are very common in statistical models. Examples include linear inequality, parameter ordering, monotonicity, orthogonality, etc. Bayesian approach is quite useful for probabilistic modeling and uncerntainty quantification in the constrained space. Although specific solutions have been made for different constraints, it is challenging to incorporate them in advanced applications, such as modeling with non-parametric assumption or high-dimensional data. In this paper, we propose a simple and general solution by first replacing constraints with strongly informative prior. Through this *extrinsic* prior, the parameters are relaxed to a less restrictive space, where conventional tools such as Hamiltonian Monte Carlo are utilized to obtain approximate posterior. Then these posteriors can be easily projected back to the constrained space for exact solution. This approach is very efficient and applicable to a wide range of problems with equality and inequality constraints. The generality allows more families of prior to be chosen for the constrained parameters, and simplifies the adoption of multiple constraints for desired property such as identifiability. Theory is developed and novel statistical applications under constraints are illustrated.

KEY WORDS: Constraint violation; Space embedding; Monotone Dirichlet; Orthogonal Gaussian processes; Posterior mixing; Projected Markov chain

1 Introduction

Constraints are very common in statistical modeling. In applied domain, modeling assumptions often require some constraint. For example, in functional data analysis related to degenerative disease, it is common to assume the curve need to satisfy certain shape constraint such as monotonicty (Lin and Dunson, 2014). In stastical optimization, constraints such as orthonormality are also routinely used to ensure identifiability of the model (Uschmajew, 2010). In manifold modeling, a large class of manifolds can be viewed as submanifolds of a more conventional space (e.g. Euchlean space), embedded via different constraints.

These constraints can cause substantial modeling difficulty. When constraints are applied on the data, they could generate an intractable integral with the parameter in the likelihood, leading to a "doubly intractable" problem. Several successful solutions have been proposed to address this issue (Murray et al.,

2012; Rao et al., 2016). On the other hand, there is a clear lack of general and simple solution, when the constraints are applied on the parameters. In frequentist optimization literature, the use of Lagrange and Karush-Kuhn-Tucker multipliers provides a means to obtain point estimate under the equality and inequality constraints (Boyd and Vandenberghe, 2004). But due to the space constraint, the standard asympotic approximation for variance estimation usually do not hold. Therefore, a Bayesian approach would be more approriate to quantify the uncertainty. Ideally, one would assign prior on a constraint support, then utilize standard toolbox such as Markov chain Monte carlo (MCMC) to obtain posterior sample on this space. However, this turns out to be very challenging.

To assign prior on the constraint space, the available families of distribution are often quite limited. For example, for othornormal matrices on the Stiefel manifold, the matrix von Mises-Fisher distribution (Khatri and Mardia, 1977) is one of the only few choices. For regression under linear inequality constraints, only until recently a tractable prior is proposed for the polyhedral region set by the inequailities (Danaher et al., 2012). Alternatively, Gelfand et al. (1992) proposed a truncation strategy by first considering common unrestricted distribution, then assigning zero support outside the constraint region. Accordingly, the posterior estimation proceeds in first generating unrestricted proposals using Gibbs sampling, then only accepting those inside the constraint space. Although this approach allows using a more general class of prior distribution, the drawback is that the unrestricted proposal can have significant mass outside the constraint region, resulting in a high rejection rate.

To meet the constraints, efficient computation is elusive and often demands substantial efforts to develop. And often it can be disrupted by slight complication such as hierarchical structure or additional constraint. For example, stick-breaking parameterization is commonly used in probability simplex modeling, in order to circumvent the constraint of vertices summing to 1. However, its computational efficency can be broken by additional structure constraint, such as the ordering of the simplex, which is useful in reducing the label switching problem in mixture modeling (Diebolt and Robert, 1994). As another example, in multiway tensor factorization, orthonormality is useful to induce good posterior mixing in estimating factor matrices. This largely relies on the sampler of Bingham-von Mises-Fisher distribution (Hoff et al., 2016). However, when there is symmetry in the slices of tensor (commonly in population of undirected networks), at least two factors would be the same. This disrupts the closed form of the posterior, demanding new rejection sampling algorithm to be developed. The Bayesian manifold modeling also faces the same quadmire. Hamiltonian Monte Carlo accomodating the geometric structure of the manifold have been developed (Girolami and Calderhead, 2011; Byrne and Girolami, 2013), but the computation is intensive and mixing is suboptimal. These challenges prohibit a good utilization of constraints in statistics.

We propose to solve this problem by viewing the constrigent constraints as a limiting case of a strongly

informative prior, referred as extrisic prior. We then relax the effective support of the prior to the neighbor of the constraint space, allowing approximate posterior to be collected efficiently via Hamiltonian Monte Carlo directly in Euchledean space. The imperfection of approximation can be corrected by an efficient reconstruction of an exact Markov chain after projection. The proposed approach is simple to implement and can be automatically carried out via software such as STAN. Theoretic studies are conducted and substantial improvement is shown in simulations and data application.

2 Method

We consider a parameter θ in a constrained space $\mathcal{D}$. The space $\mathcal{D}$ can be high- or infinite-dimensional. The common Bayesian approach assigns a prior for θ in $\mathcal{D}$, denoted by $\pi_{0,\mathcal{D}}(\theta)$. Typically, priors are chosen for computational conveniences so that the posterior can be easily sampled, strictly from $\mathcal{D}$. We refer these strategies as intrinsic approaches. Clearly, the choices of priors and constraints one can impose are very limited. Instead, we consider extrinsic approaches, where one sample random variables in the larger space $\mathcal{R}$ where $\mathcal{D} \in \mathcal{R}$, then use them as approximate or proposal in rejection sampling. We first provide an probabilistic justification for extrinsic approaches.

Assuming $\pi_{0,\mathcal{D}}(\theta)$ is proper $\int_{\mathcal{D}} \pi_{0,\mathcal{D}}(\theta) < \infty$, then this prior can be viewed as a conditional density, based on another density $\pi_{0,\mathcal{R}}(\theta)$ associated with $\mathcal{R}$:

$$\pi_{0,\mathcal{D}}(\theta) = \pi_{0,\mathcal{R}}(\theta \mid \theta \in \mathcal{D}) = \frac{\pi_{0,\mathcal{R}}(\theta)\mathbb{1}_{\theta\in\mathcal{D}}}{\int_{\mathcal{D}} \pi_{0,\mathcal{R}}(\theta)d\theta}. \tag{1}$$

where $\mathbb{1}_{\theta\in\mathcal{D}} = 1$ when $\theta \in \mathcal{D}$, 0 otherwise. Note $\pi_{0,\mathcal{R}}(\theta)$ can be improper, as along as $\pi_{0,\mathcal{R}}(\theta)\mathbb{1}_{\theta\in\mathcal{D}}$ is proper. Then the posterior can be obtained via:

$$\pi(\theta \mid y, \theta \in \mathcal{D}) = \frac{L(\theta; y)\pi_{0,\mathcal{D}}(\theta)}{\int_{\mathcal{D}} L(\theta; y)\pi_{0,\mathcal{D}}(\theta)d\theta} = \frac{L(\theta; y)\pi_{0,\mathcal{R}}(\theta)\mathbb{1}_{\theta\in\mathcal{D}}}{\int_{\mathcal{D}} L(\theta; y)\pi_{0,\mathcal{R}}(\theta)d\theta}, \tag{2}$$

where the last equality holds because $\int_{\mathcal{D}} \pi_{0,\mathcal{R}}(\theta)d\theta$ is a finite constant. Utilizing this form, Gelfand et al. (1992) suggested generating proposal in $\mathcal{R}$ based on $L(\theta; y)\pi_{0,\mathcal{R}}(\theta)$, then accepting it when it falls in $\mathcal{D}$. When the probability $\Pr(\theta \in \mathcal{D} \mid y)/\Pr(\theta \in \mathcal{R} \setminus D \mid y) \approx 0$, this leads to significant amount of rejections.

2.1 Extrinsic Prior for Constraints

We propose a different strategy. We first assume the embedding of $\mathcal{D}$ in $\mathcal{R}$ is done via equality and inequality constraints, although other types of constraints can be incorporated similarly. There are m equalities and l inequalities, leading to $\mathcal{D} = \{\theta \in \mathcal{R} : E_k(\theta) = 0 \text{ for } k = 1, \ldots, m, \quad G_{k'}(\theta) \leq 0 \text{ for } k' = 1, \ldots, l\}$, where

$E_k(.)$ and $G_{k'}(.)$ are functions that map from $\mathcal{R}$ to real line $\mathbb{R}$. Then the indicator function is a product of $(m+l)$ functions $\mathbb{1}_{\theta\in\mathcal{D}} = \prod_k \mathbb{1}_{E_k(\theta)=0} \cdot \prod_{k'} \mathbb{1}_{G'_k(\theta)\leq 0}$.

We now view the constringent embedding $\prod_k \mathbb{1}_{E_k(\theta)=0} \cdot \prod_{k'} \mathbb{1}_{G'_k(\theta)\leq 0}$ as a limiting case of a set of strongly informative priors, represented by $(m+l)$ kernel functions $K_{\cdot}(.)$. This first to an alternative posterior:

$$\pi_K(\theta \mid y) \propto L(\theta; y)\pi_{0,\mathcal{R}}(\theta) \cdot \prod_{k=1}^{m} K_{1,k}\Big(|E_k(\theta)|\Big) \cdot \prod_{k'=1}^{l} K_{2,k'}\Big(\big(G_{k'}(\theta)\big)_+\Big) \tag{3}$$

where $(x)_+ = x$ if $x > 0$, 0 if $x \leq 0$. The posterior $\pi_K(\theta \mid y)$ is an approximation to $\pi(\theta \mid y)$ in (2). The random variable sampled from the kernel is the amount of contraint violation $|E_k(\theta)| \in [0, \infty)$ or $(G_{k'}(\theta))_+ \in [0, \infty)$. Each kernel $K_{i,k}$ satisfies $K_{i,k}(0) = 1$ and is controled by a hyper-parameter $\lambda_{i,k}$, when $\lambda_{i,k} \to \infty$, the kernel becomes a point mass at 0. Then (3) becomes the same as (2). For example, one simple and useful kernel is the truncated Gaussian $K_{i,k}(x) = \exp(-\lambda_{i,k}x^2)\mathbb{1}_{x<\epsilon}$.

Instead of taking infinity, we assign large but finite value for each $\lambda_{i,k}$. This gives rise to a continuous relaxation of the sharp boundary of the embedding. The relaxation allows the posterior θ to be easily sampled in $\mathcal{R}$ under the guidance of the constraints. For example, one can carry out Hamiltonian Monte Carlo directly in Euchledean space, aided by highly automatic software such as STAN. At the same time, since posteriors are generated in a tight neighborhood of $\mathcal{D}$, they can be easily projected back to $\mathcal{D}$ to make them good proposal in a correcting Metropolis-Hastings step.

Let $\mathcal{E}_{i,k}(x) = K_{i,k}(x)/\big(\int_{\mathcal{R}} K_{\cdot}(x)dx\big), \quad x \geq 0$ be the density for $K_{i,k}(x)$. As they induce the constraints *extrinsically*, we refer them as extrinsic priors. We first describe the kernel specification, the posterior sampling for $\pi_K(\theta \mid y)$ and then the correcting step to convert $\pi_K(\theta \mid y)$ into the exact posterior.

2.2 Kernel Specification

In the approximate posterior (3), when $\theta \in \mathcal{D}$, each kernel is 1 and the density becomes the same as (2). However, since we induce positive support in $\mathcal{R} \setminus \mathcal{D}$, it is worth studying how the approximate posterior are distributed relatively to the space $\mathcal{D}$. This is reflected in the posterior distribution of the constraint violation $|E_k(\theta)|$ and $(G_{k'}(\theta))_+$.

We control the amount of constraint violation via a tight prior support near 0 in each $\mathcal{E}_{i,k}(x)$. That is $\int_{x<\epsilon} \mathcal{E}_{i,k}(x) = 1$. The pre-specified constant ϵ represents the element-wise tolerance for violating each constraint. The bounded prior support allows us to theoretically control the posterior approximation error. Letting $\pi_K(\theta) = \prod_{i,k} \mathcal{E}_{i,k}(x)$ be the joint extrinsic prior density. Since $\pi_K(\theta \mid y) \ll \pi_K(\theta)$, the posterior for each constraint violation is bounded in $[0, \epsilon)$ with probability 1.

In practice, one may wish to use kernel $K^*_{i,k}(x)$ with unbounded support on $[0,\infty)$ for computing conveniences. To adopt them, one can first choose $\lambda_{i,k}$ to have $\int_{x<\epsilon} \mathcal{K}^*_{i,k}(x)/\big(\int_{\mathcal{R}} K^*_{\cdot}(x)dx\big) = 1-\eta$ with η small, then apply truncation $K_{i,k}(x) = K^*_{i,k}(x)\mathbb{1}_{x<\epsilon}$ to ensure $x<\epsilon \quad a.s.$. In most cases, the truncation is simply nominal for a theoretic guarantee. For example, in Gaussian kernel $\exp(-\lambda x^2)$ setting $\lambda = \frac{1}{2(\epsilon/4)^2}$ ensures $x<\epsilon$ with probability 0.99993 apriori; the truncation at $x<\epsilon$ is mostly not needed for posterior sampling.

To illustrate, we consider a simple example of generating a truncated Gaussian distribution $\theta \mid y \sim \mathrm{No}_{(\alpha,\beta)}(0,1)$, with mean 0 and variance 1 and truncation $\theta \in (\alpha,\beta)$. The exact and the approximate densities are:

$$\pi(\theta \mid y) \propto \exp(-\frac{\theta^2}{2})\mathbb{1}_{\theta\in(\alpha,\beta)}, \quad \pi_K(\theta \mid y) \propto \exp(-\frac{\theta^2}{2})K\left((\alpha-\theta)_+\right)K\left((\theta-\beta)_+\right).$$

with $K(x) = \exp(-\lambda x^2)\mathbb{1}_{x<4/\sqrt{2\lambda}}$. We set $(\alpha,\beta) = (1,2)$. Figure 1 plots the unnormalized densities under the exact posterior and approximation with different λ's. The approximate densities inside $\mathcal{D} = (1,2)$ are the same as the exact one, up to a constant difference due to normalization. Outside $\mathcal{D}$, the larger λ is associated with more rapid decline of density and therefore smaller tolerance for constraint violation.

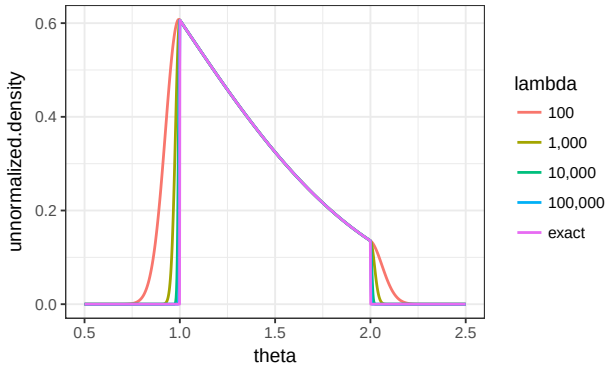

Figure 1: Unnormalized densities for truncated normal $\mathrm{No}_{(1,2)}(0,1)$, under exact and approximating densities. The exact density abruptly drops to 0 on the two boundaries, while the approximating ones drop continuously. In the approximation, larger λ is associated with lower tolerance for constraint violation ($(1-\theta)_+$ and $(\theta-2)_+$) in the posterior. All densities inside $(1,2)$ are the same up to a constant difference.

The kernel specification allows us to obtain strict control of the constraint violation in the posterior. It is attempting to use large λ for any model, however, we discover that for model with narrow support on $\mathcal{D}$ would have a small step size due to the large λ. Therefore, in those cases, it is more useful to obtain approximate posterior with moderately small error, then project back to $\mathcal{D}$ for exact posterior. More details will be illustrated in the example of next section.

2.3 Posterior Sampling Under Extrinsic Prior

We first collect posterior $\pi_K(\theta \mid y)$ under the $\pi_{0,\mathcal{R}}(\theta)\pi_K(\theta)$. Since the posterior is defined on a less restrictive $\mathcal{R}$, it can be sampled easily. The tradiational sampling approach such as slice sampling, adaptive Metrohepolis-Hastings can be utilized in computation. In this section, we present the sampling algorithm using Hamiltonian Monte Carlo (HMC), due to its high-level automation aided by software and excellent performance in terms of posterior mixing.

In using the conventional HMC, we assume $\mathcal{R}$ is an Euchledean space and the constraint functions $E_k(\theta)$'s and $G_k(\theta)$'s are differentiable with respect to θ. We focus on the case where θ is continuous, although discrete extension is possible (Zhang et al., 2012).

HMC is essentially a data augmentation based Monte Carlo. Using a latent variable "veolicty" p that has the same dimension as θ, the total negative log-likelihood-prior function based on (3) is

$$\begin{aligned} &H(\theta, p) = U(\theta) + M(p), \\ &\text{where } U(\theta) = -\log\left\{ L(\theta; y)\pi_{0,\mathcal{R}}(\theta) \cdot \prod_{k=1}^{m} K_{1,k}\Big(|E_k(\theta)|\Big) \cdot \prod_{k'=1}^{l} K_{2,k'}\Big(\big(G_{k'}(\theta)\big)_+\Big)\right\}, \\ &M(p) = \frac{p'\Sigma^{-1}p}{2}, \end{aligned} \tag{4}$$

with Σ^{-1} a positive definite matrix and pre-specified as a tuning parameter. Then instead of using conventional updates such as random walk or Gibbs sampling, HMC utilizes the Hamiltonian dynamics that satisfy:

$$\begin{aligned} \frac{\partial\theta(t)}{\partial t} &= \frac{\partial H(\theta, p)}{\partial p} = \Sigma^{-1}p, \\ \frac{\partial p(t)}{\partial t} &= -\frac{\partial H(\theta, p)}{\partial \theta} = -\frac{\partial U(\theta)}{\partial \theta}. \end{aligned} \tag{5}$$

At each step, the current state of θ is viewed as $\theta(0)$ with $p(0)$ randomly generated from $\mathrm{No}(0, \Sigma)$. Then they enter the Hamiltonian dynamics to generate $\theta(t)$ and $p(t)$. During this process, either exact solution or numerical approximation is obtained. When the solution is exact, the total $H(\theta(0), p(0)) = H(\theta(t), p(t))$ and $\theta(t)$ is accepted as the new state for θ in the Markov chain; when the solution is approximate (commonly via leap-frog method), an additional Metropolis-Hastings step is taken to accept $\theta(t)$ with probability $1 \wedge \exp\left(-H(\theta(t), p(t)) + H(\theta(0), p(0))\right)$. Various adaptive approaches (Neal et al., 2011; Hoffman and Gelman, 2014) have been developed for making the new state less correlated with the current state. Due to the existence of $\frac{\partial U(\theta)}{\partial \theta}$, the sampling can be carried out easily in common HMC software such as STAN. HMC is geometrically ergodic under very general conditions (Livingstone et al., 2016).

Based on (5), we provide some further intuition on the effect of extrinsic prior in HMC, and guide on selecting the hyper-parameter λ from a computing perspective. When $\theta \in \mathcal{D}$, extrinsic prior is constant and does not influence the HMC; when $\theta \notin \mathcal{D}$, it provides a space expansion by assigning positive support. The amount of expansion is reversely controlled by λ. Since the Hamiltonian dynamics (5) is commonly approximated by discrete movement in $\mathcal{R}$ (e.g. leap-frog), it is important to ensure the expanded posterior support can be efficently explored in the direction of $\frac{\partial U(\theta)}{\partial \theta}$. For example, when $\mathcal{D}$ is a truncated Euchledean space, the gradient $\frac{\partial U(\theta)}{\partial \theta}$ is still efficent for exploring its internal space, so very large λ can be used to induce almost no expanded space. However, when $\mathcal{D}$ is on a circle, updating along any Euchledean direction will move away from this space. In this case, a smaller λ would be better for larger space expansion, allowing HMC to explore more efficently.

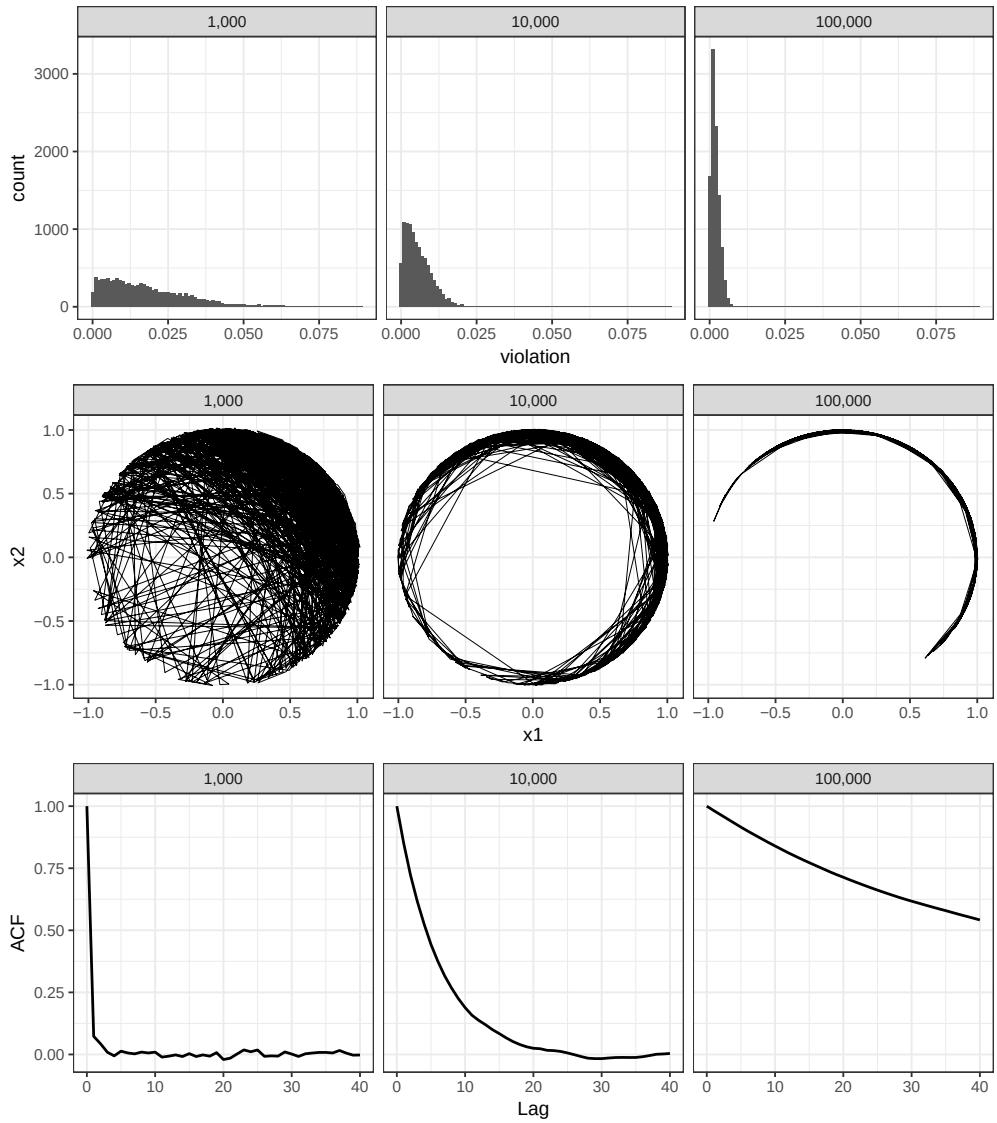

Figure 2: Sampling posterior from a von Mises–Fisher distribution on a unit circle, using HMC with extrinc prior under $\lambda = 10^3, 10^4, 10^5$. Row 1 shows the posterior distribution of the constraint violation $|\theta'\theta - 1|$; Row 2 shows the sampling path of the Markov chain; Row 3 shows the autocorrelation plot (ACF). Large λ results in small constraint violation, but suffers from slow mixing due to inefficient local update; smaller λ increases the approximation error but results in excellent mixing.

To illustrate the latter case, consider generating a random variable $\theta = (x_1, x_2)$ on a unit circle using von Mises–Fisher distribution, $\pi(\theta \mid y) \propto \exp(F'\theta)$ with $\theta'\theta = 1$. This is a simple example of a random variable constraint on a $(2, 1)$-Stiefel manifold $\mathcal{D} = \mathcal{V}(2, 1)$. We set $F = (1, 1)$ to induce a distribution widely spreaded over the manifold, allowing great amount of uncertainty. We use extrinsic prior proportional to $K(\theta) = \exp(-\lambda(\theta'\theta - 1)^2)\mathbb{1}_{|\theta'\theta-1|<0.1}$. Geometrically, the extrinsic prior expands the posterior support from a circle to a ring, with its radius defined by the maximally tolerable constraint violation.

We tested three different values of $\lambda = 10^3, 10^4, 10^5$ associated with different radii. For each λ, we ran HMC for $10,000$ iterations, with 100 leap-frog steps in each iteration. We use $\Sigma = \text{diag}(1, 1)$ to generate

velocity. During the initial $2,000$ iterations, we automatically tune the leap-frog step size to obtain an acceptance rate close to 0.6 in each iteration, then fixed the value for the remaining part of Markov chain. The last $5,000$ iterations are used as posterior samples. Figure 2 plots the posterior distribution of constraint violation $|\theta'\theta - 1|$, the sampling path and the autocorrelation function (ACF) for each Markov chain. Very large $\lambda = 10^5$ has much less constraint violation (hence lower approximation error); however, due to the narrow radius, the associated HMC can only explore local space and results in slow mixing (large autocorrelation even at 40 lags). On the other hand, smaller $\lambda = 10^3$ has slightly larger violation (but it is still relatively small), while allows efficient exploration of the space and excellent mixing performance.

2.4 Correcting Projection to Constraint Space

The posterior sample collected under $\pi_K(\theta \mid y)$ in $\mathcal{R}$ is an approximation to $\pi(\theta \mid y)$ in constraint space $\mathcal{D}$. One may be interested in further obtaining exact posterior in $\mathcal{D}$, likely for two reasons: (1) to strictly uphold the constraints; (2) to ease the error control in approximate posterior sampling for more efficient computation.

Letting θ^* be a random sample collected based on $\pi_K(\theta \mid y)$, there exists deterministic projection $M : \mathcal{R} \to \mathcal{D}$ and obtain $\theta^*_{\mathcal{D}} = M(\theta^*)$. Using this as proposal machineary, one can construct another Markov chain based on $\pi(\theta_{\mathcal{D}} \mid y)$. Letting the current state be $\theta_{\mathcal{D}} = M(\theta)$, we generate proposal $\theta^*_{\mathcal{D}} = M(\theta^*)$ and accept it with probability:

$$1 \wedge \frac{\pi(\theta^*_{\mathcal{D}} \mid y)\pi_K(\theta \mid y)}{\pi(\theta_{\mathcal{D}} \mid y)\pi_K(\theta^* \mid y)}. \tag{6}$$

This criterion holds since the transformation is deterministic and the posterior collected based on $\pi_K(\theta \mid y)$ is assumed to be independent. Rigoriously speaking, the second condition relies on the geometric convergence of the Markov chain from sampling $\pi_K(\theta \mid y)$, which can be achieved efficiently via algorithm such as HMC.

The next task is then to optimize the projection. Noting that

$$|\log(\frac{\pi(\theta^*_{\mathcal{D}} \mid y)\pi_K(\theta \mid y)}{\pi(\theta_{\mathcal{D}} \mid y)\pi_K(\theta^* \mid y)})| \leq \big||\log(\pi(\theta^*_{\mathcal{D}} \mid y)) - \log(\pi_K(\theta^* \mid y))| + |\log(\pi(\theta_{\mathcal{D}} \mid y)) - \log(\pi_K(\theta \mid y))|\big|, \tag{7}$$

it is sensible to use the projection $\theta_{\mathcal{D}} = M(\theta)$ minimizing $Q(\theta_{\mathcal{D}}) = |\log(\pi(\theta_{\mathcal{D}} \mid y)) - \log(\pi_K(\theta \mid y))|$ towards 0. And the acceptance rate will be close to 1. This is possible since the extrinsic prior ensures that the approximate θ is close to $\mathcal{D}$ and $\pi(\theta_{\mathcal{D}} \mid y) = \pi_K(\theta_{\mathcal{D}} \mid y)$. Obviously, when the approximate $\theta \in \mathcal{D}$, the optimal projection would be the identity function. When $\theta \notin \mathcal{D}$, standard optimization technique such as KKT multipliers can be used.

3 Illustration

Example 1: Ordered Simplex

Example 2: Monotone Spline

Example 3: Orthonormal Gaussian Processes

4 Theory

posterior propriety

5 Application

6 Discussion

References

Boyd, S. and L. Vandenberghe (2004). *Convex optimization*. Cambridge university press.

Byrne, S. and M. Girolami (2013). Geodesic monte carlo on embedded manifolds. *Scandinavian Journal of Statistics 40*(4), 825–845.

Danaher, M. R., A. Roy, Z. Chen, S. L. Mumford, and E. F. Schisterman (2012). Minkowski–weyl priors for models with parameter constraints: an analysis of the biocycle study. *Journal of the American Statistical Association 107*(500), 1395–1409.

Diebolt, J. and C. P. Robert (1994). Estimation of finite mixture distributions through bayesian sampling. *Journal of the Royal Statistical Society. Series B (Methodological)*, 363–375.

Gelfand, A. E., A. F. Smith, and T.-M. Lee (1992). Bayesian analysis of constrained parameter and truncated data problems using gibbs sampling. *Journal of the American Statistical Association 87*(418), 523–532.

Girolami, M. and B. Calderhead (2011). Riemann manifold langevin and hamiltonian monte carlo methods. *Journal of the Royal Statistical Society: Series B (Statistical Methodology) 73*(2), 123–214.

Hoff, P. D. et al. (2016). Equivariant and scale-free tucker decomposition models. *Bayesian Analysis 11*(3), 627–648.

Hoffman, M. D. and A. Gelman (2014). The no-u-turn sampler: adaptively setting path lengths in hamiltonian monte carlo. *Journal of Machine Learning Research 15*(1), 1593–1623.

Khatri, C. and K. Mardia (1977). The von mises-fisher matrix distribution in orientation statistics. *Journal of the Royal Statistical Society. Series B (Methodological)*, 95–106.

Lin, L. and D. B. Dunson (2014). Bayesian monotone regression using gaussian process projection. *Biometrika*.

Livingstone, S., M. Betancourt, S. Byrne, and M. Girolami (2016). On the geometric ergodicity of hamiltonian monte carlo. *arXiv preprint arXiv:1601.08057*.

Murray, I., Z. Ghahramani, and D. MacKay (2012). Mcmc for doubly-intractable distributions. *arXiv preprint arXiv:1206.6848*.

Neal, R. M. et al. (2011). Mcmc using hamiltonian dynamics. *Handbook of Markov Chain Monte Carlo 2*, 113–162.

Rao, V., L. Lin, and D. B. Dunson (2016). Data augmentation for models based on rejection sampling. *Biometrika 103*(2), 319–335.

Uschmajew, A. (2010). Well-posedness of convex maximization problems on stiefel manifolds and orthogonal tensor product approximations. *Numerische Mathematik 115*(2), 309–331.

Zhang, Y., Z. Ghahramani, A. J. Storkey, and C. A. Sutton (2012). Continuous relaxations for discrete hamiltonian monte carlo. In *Advances in Neural Information Processing Systems*, pp. 3194–3202.