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 useful for 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 modern statistical modeling. For example, functional data anaylsis often impose certain shape constraint such as monotonicity or convexity on curves (Kelly and Rice, 1990); matrix and tensor decomposition utilize orthonormality for better identifiability (Uschmajew, 2010); many manifolds such as simplex can be considered as sub-manifolds of a Euclidean space embedded via certain constraints.

Constraints can cause substantial modeling difficulty. When data are in constrained space, parameters can enter the likelihood via an integral without closed-form, commonly known as "doubly intractable" problem. Successful solutions have been proposed to address this issue (Murray et al., 2012; Rao et al., 2016). When parameters are in constrained space, challenges often arise in the difficulty for estimation under constraint. Frequentist optimization literature often relies on Lagrange and Karush-Kuhn-Tucker multipliers for point

estimate under equality and inequality constraints (Boyd and Vandenberghe, 2004). However, the uncertainty quantification is difficult since conventional asympotic result on variance estimation often no longer hold in constrained space. Bayesian approach is more approriate for this purpose.

There have been a wide range of solutions developed for specific constraints. One strategy involves using constrained prior with posterior that can be conveniently sampled. For example, to model othornormal matrices on the Stiefel manifold, Bingham-von-Mises-Fisher distribution (Khatri and Mardia, 1977; Hoff, 2009) is a parametric family with a closed-form posterior in matrix and tensor decomposition. Lin et al. (2016) extends the flexibility of matrix von-Mises-Fisher distribution via non-parametric approach. Another strategy is to bypass the constraint via re-parameterization. The famous example is the stick-breaking construction for Dirichlet distribution and process. The re-parameterization essentially utilizes the coordinate system of the simplex, and circumvents the norm constraint on the probability vertices. As these methods directly satisfy the constraint requirement, we refer them as the intrinsic approaches. Despite the success of intrinsic approaches, the posterior can quickly become very invovled to sample under slightly more advanced model or complicated data. For example, in modeling population of undirected networks, the symmetry in each network disrupts the closed form of posterior in orthogonal tensor decomposition (Hoff et al., 2016), demanding new rejection sampling algorithm to be developed. As another example, additional strucutre (such as ordering) on the probability simplex would disrupt the simple form of stick-breaking posterior.

These drawbacks have motivated the development of extrinsic approaches. The key idea is to first sample the proposal freely in a conventional space (such as Euclidean space), then transform it back to the constrained space. One early work can be traced back to Gelfand et al. (1992), who used Gibbs sampling to first generate proposal in unrestricted region, then only accepting those inside the constraint space. One critical issue is that unrestricted proposal can have significant mass outside the constraint region, resulting in a high rejection rate. Replacing rejection sampling, Lin and Dunson (2014) and Lin et al. (2016) utilize projection to map the unconstrained posterior into the constrained space and obtain monotonicity and manifold-valued regression. These specialized cases seem to work well, but the suitable projection are often elusive and there is a clear lack of general and simple approach.

In this paper, we propose a general extrinsic approach, by parameterizing constraints as a limiting case of strongly informative prior. We refer them as extrisic priors. We then relax the effective support of the prior to a neighborhood of constraint space, obtaining posterior via efficient tools such as conventional Hamiltonian Monte Carlo (HMC). These posteriors are approximate to the canonical formulation, with approximation error bounded during the prior specification. The imperfection of approximation can be corrected with a simple projection and a Metropolis-Hastings step with high acceptance probability, leading to a Markov chain corresponding to the exact formulation. Compared to other manifold based methods such as Riemannian

and geodesic HMC (Girolami and Calderhead, 2011; Byrne and Girolami, 2013), our approach is efficient in computation and simple to implement via highly automatic software like STAN. The simplicity enables a larger spectrum of prior to be chosen and more free adaoption of constraints in modeling. Theoretic studies are conducted and original models are 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. Letting this space be equiped with a σ-field $\mathscr{B}$, the standard Bayesian approach assigns a prior for θ in $\mathcal{D}$, based on a density $\pi_{0,\mathcal{D}}(\theta)$ in a separable space $(\mathcal{D}, \mathscr{B})$. In intrinsic approaches, priors are chosen for computational conveniences so that the posterior can be easily sampled strictly inside $\mathcal{D}$. Clearly, the choices of priors and constraints one can impose are very limited. Instead, we consider extrinsic approaches by estimating θ in the larger space $\mathcal{R}$ where $\mathcal{D} \in \mathcal{R}$. We first provide a probabilistic justification.

Assuming $\pi_{0,\mathcal{D}}(\theta)$ is proper $\int_{\mathcal{D}} \pi_{0,\mathcal{D}}(\theta)d\theta < \infty$, then this prior can be viewed as a conditional density, based on another density $\pi_{0,\mathcal{R}}(\theta)$ in $(\mathcal{R}, \mathscr{C})$ with $\mathscr{C}$ as the σ-field of $\mathcal{R}$:

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

where $\mathbb{1}_{\theta\in\mathcal{D}} = 1$ when $\theta \in \mathcal{D}$, 0 otherwise. Note as along as $\pi_{0,\mathcal{R}}(\theta)\mathbb{1}_{\theta\in\mathcal{D}}$ is proper, $\pi_{0,\mathcal{R}}(\theta)$ can be improper. Letting $L(\theta; y)$ be the likelihood function and y be the observed data, 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.

2.1 Extrinsic Prior for Constraints

One obvious extrinsic approach utilizing (2) is to first generate proposal in $\mathcal{R}$ based on $L(\theta; y)\pi_{0,\mathcal{R}}(\theta)$ (assuming it is proper), then accepting it when it falls in $\mathcal{D}$ (Gelfand et al., 1992). However, when the probability $\pi(\theta \in \mathcal{D} \mid y)/\pi(\theta \in \mathcal{R} \setminus D \mid y) \approx 0$, this leads to significant amount of rejections. This is very common in equality constraint.

We propose a different strategy. Instead of ignoring $\mathbb{1}_{\theta\in\mathcal{D}}$ in the first step, we approximate it with an additional strongly informative prior $\mathcal{E}(\theta)$. The prior has its support in $\mathcal{R}$, but concentrated around $\mathcal{D}$. When $\theta \in \mathcal{D}$, $\mathcal{E}(\theta)$ is constant; $\theta \notin D$, $\mathcal{E}(\theta)$ quickly drops to 0. Then one can first obtain approximate posterior based on density proportional to $L(\theta; y)\pi_{0,\mathcal{R}}(\theta)\mathcal{E}(\theta)$ (the conditions for posterior properity is postponed to the theory section).

In this paper, we focus on the embedding of $\mathcal{D}$ in $\mathcal{R}$ 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 $\mathbb{1}_{\theta\in\mathcal{D}} = \prod_k \mathbb{1}_{E_k(\theta)=0} \cdot \prod_{k'} \mathbb{1}_{G'_k(\theta)\leq 0}$.

We now replace the indicator functions with $\mathcal{E}(\theta)$, represented as a product of $(m + l)$ kernel functions $K_{\cdot}(.)$, leading to approximate posterior:

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

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). We will now refer $\pi_K(\theta \mid y)$ as "extrinsic posterior". The functions $|E_k(\theta)| \in [0, \infty)$ or $(G_{k'}(\theta))_+ \in [0, \infty)$ represent the amount of relaxation or violation of each constraint, where 0 represents no violation. Each kernel $K_{i,k}$ satisfies $K_{i,k}(0) = 1$; the tolerable amount of violation is controlled by a hyper-parameter $\lambda_{i,k}$. When $\lambda_{i,k} \to \infty$, the kernel becomes a point mass at 0. Therefore, (2) is a limiting case of (3). 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 infinite values for λ's, we assign large but finite ones. This gives rise to a continuous relaxation of the sharp boundary of the indicator function. 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 conventional HMC for constrained parameters directly in Euchledean space. At the same time, since posteriors are generated in a tight neighborhood of $\mathcal{D}$, they can be easily projected back to $\mathcal{D}$ as proposals in a correcting Metropolis-Hastings step.

The specification of λ's not only controls the approximation error in terms of the constrain violation, but could also potentially impact the computing efficiency in sampling $\pi_K(\theta \mid y)$. It is useful to control constraint violation within acceptance range, while optimizing for sampling efficiency.

2.2 Control of Constraint Violation

In extrinsic posterior (3), when $\theta \in \mathcal{D}$, the density is the same as (2), up to a constant difference. 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 $(G_{k'}(\theta))_+$.

The constraint violation can be controlled via a tight prior support near 0 for each kernel. That is $\int_{x<\epsilon} \mathcal{C}_{i,k}(x)dx = 1$, with $\mathcal{C}_{i,k}(x) = K_{i,k}(x)/\int_{\mathcal{R}} K_{i,k}(x)dx$. 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. As $\mathcal{E}(\theta) \propto \prod_{i,k} \mathcal{C}_{i,k}(x)$ is the joint extrinsic prior density, since $\pi_K(\theta \mid y) \ll \mathcal{E}(\theta)$, the posterior for each constraint violation is bounded in $[0, \epsilon)$ with probability 1.

In practice, one may wish to utilize a 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 $\int_{x<\epsilon} K^*_{i,k}(x) / \left(\int_{\mathcal{R}} K^*_{\cdot}(x) dx \right) = 1 - \eta$ with η small, then apply truncation $K_{i,k}(x) = K^*_{i,k}(x) \mathbb{1}_{x<\epsilon}$ to ensure $x < \epsilon$ almost surely. In most cases, the truncation is just nominal for a theoretic guarantee; in computation it can be satisfied automatically with high probability. 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; for posterior sampling, one can first do an untruncated sampling, then reject those $x < \epsilon$, which is likely rare due to the small prior probability.

To illustrate the approximation of extrinsic prior and control of constraint violation, 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 extrinsic posterior 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.

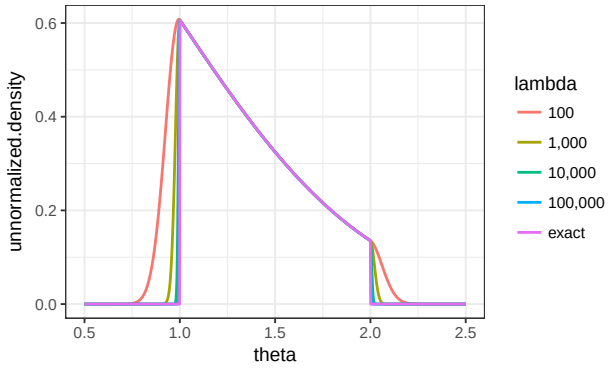

Figure 1: Unnormalized densities for truncated normal $\mathrm{No}_{(1,2)}(0, 1)$, under exact and approximating densities. The exact density abruptly drops to 0 on the two boundaries, while the approximating ones drop continuously. In the approximation, larger λ is associated with lower tolerance for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

It is tempting to use very large λ for all models. However, when the constrained space $\mathcal{D}$ has a geometry not suitable for discrete step update in $\mathcal{R}$ (as commonly used in HMC), large λ would cause slow posterior

mixing. Since the approximation can be eventually corrected in an additional step, it is rather useful to first use a smaller λ to induce a larger space expansion for efficient posterior sampling. We will illustrate this via an example in next section.

2.3 Posterior Sampling for Extrinsic Posterior

Extrinsic posterior is the approximation to the ones under exact formulation. As it is defined on a less restrictive space $\mathcal{R}$, it can be sampled easily. The tradiational sampling tools such as slice sampling, adaptive Metrohepolis-Hastings can be utilized. 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 convergence and posterior mixing. Various adaptive algorithms such as Hoffman and Gelman (2014) have been developed for making the new state less correlated with the current state.

In using the conventional HMC, we assume the space $\mathcal{R}$ 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 MCMC. Using a latent variable named "veolicty" p with the same dimension as θ, the negative log-posterior function based on (3) is

$$\begin{aligned} &H(\theta, p) = U(\theta) + M(p), \\ &\text{where } U(\theta) = -\log\left\{L(\theta; y)\pi_{0,\mathcal{R}}(\theta)\mathcal{E}(\theta)\right\}, \\ &M(p) = \frac{p'\Sigma^{-1}p}{2}, \end{aligned} \tag{4}$$

with Σ^{-1} a pre-specified positive definite matrix. Unlike conventional MCMC, HMC utilizes the Hamiltonian dynamics based on the solution to the differential equations:

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

At each step, the current state of θ is viewed as $\theta(0)$ with $p(0)$ randomly generated from $\text{No}(0, \Sigma)$. Then they enter the Hamiltonian dynamics to generate $\theta(t)$ and $p(t)$. An Metropolis-Hastings step is taken at the end to accept $\theta(t)$ with probability $1 \wedge \exp\left(-H(\theta(t), p(t)) + H(\theta(0), p(0))\right)$. When solution to (5) has closed-form, the acceptance rate is always 1. However, often numerical approximation with discrete movement is commonly needed, such as the leap-frog algorithm (Neal et al., 2011):

$$
\begin{aligned}
p(T+\epsilon/2) &= p(T) - \epsilon/2 \frac{\partial U}{\partial \theta}(\theta(T)), \\
\theta(T+\epsilon) &= \theta(T) + \epsilon \Sigma^{-1} p(T+\epsilon/2), \\
p(T+\epsilon) &= p(T+\epsilon/2) - \epsilon/2 \frac{\partial U}{\partial \theta}(\theta(T+\epsilon)),
\end{aligned} \tag{6}
$$

for $T = 0, \epsilon, 2\epsilon, \ldots, L\epsilon$, with L as the total steps within one iteration and $t = L\epsilon$.

Due to the cost of multiple steps, one would hope to utilize relatively large ϵ with relatively small L. This could bring some further implication on choosing λ in the extrinsic prior $\mathcal{E}(\theta)$. Specifically, one needs to ensure a decent-sized discrete move along $\frac{\partial U(\theta)}{\partial \theta}$ does not end in a low posterior density region; otherwise, only a small step size ϵ can be used.

For example, when $\mathcal{D}$ is a truncated Euchledean space defined by simple space truncation, random move along certain directions of $\frac{\partial U(\theta)}{\partial \theta}$ can still lead to high posterior region its internal space of $\mathcal{D}$, therefore large λ generally do not impact the effectiveness of using large λ. On the other hand, when $\mathcal{D}$ is on a unit circle, random move along any Euchledean direction will move away from this space. With very large λ, the high posterior density region would be very narrow, demanding a small ϵ. In this case, a smaller λ inducing a greater support expansion would be more useful, since it allows more efficient Hamiltonian dynamics.

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, generating great amount of uncertainty to assess the mixing performance. 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.

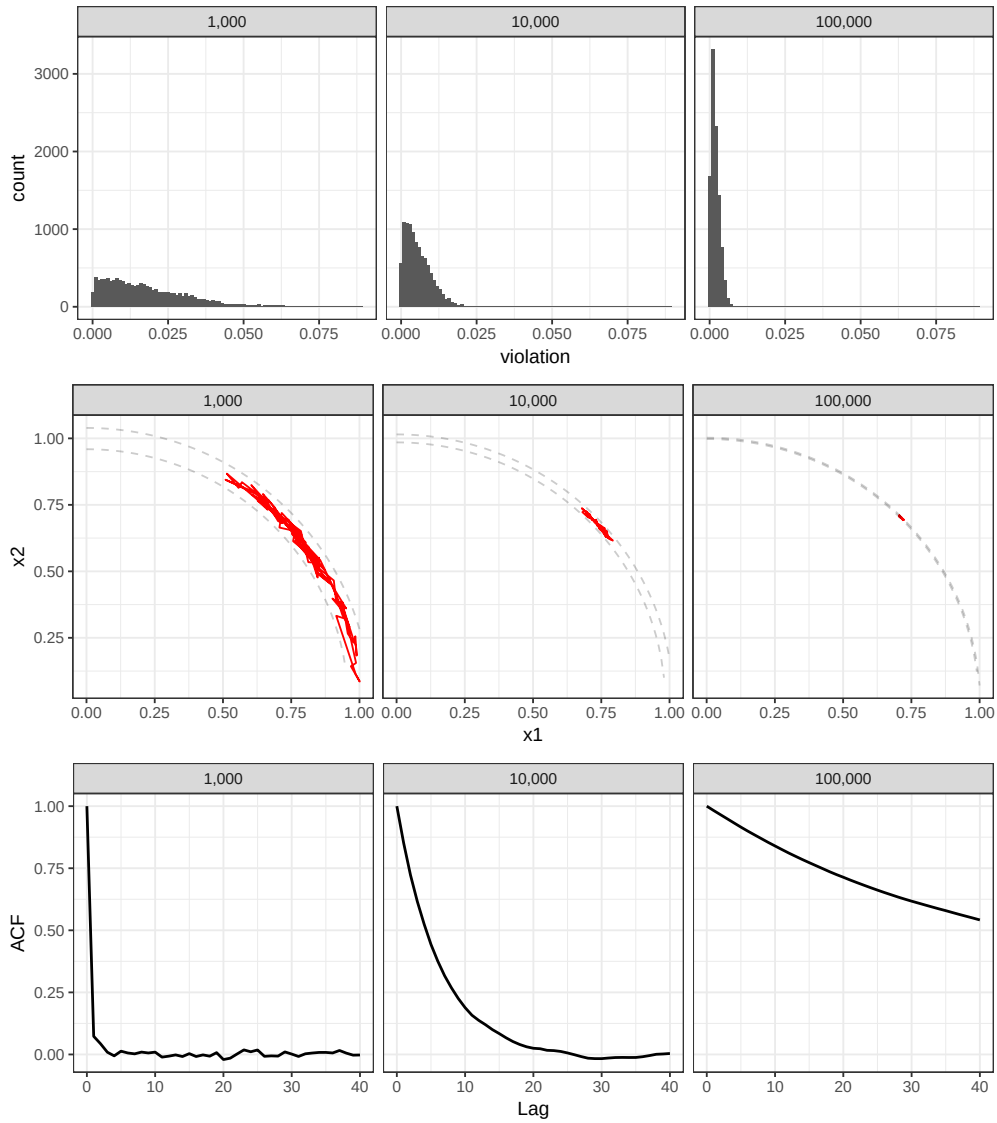

Figure 2: Sampling posterior from a von Mises–Fisher distribution on a unit circle, using HMC with extrinc prior under $\lambda = 10^3, 10^4, 10^5$. Row 1 shows the posterior distribution of the constraint violation $|\theta'\theta - 1|$; Row 2 shows the path of $1,000$ leap-frog steps; Row 3 shows the 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.

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 $L = 100$ leap-frog steps in each iteration. We use $\Sigma = \text{diag}(1, 1)$ to generate velocity. During the initial $2,000$ iterations, the leap-frog step size ϵ is automatically tuned for an acceptance rate close to 0.6, then it is fixed during 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; however, due to the narrow radius of the ring, the associated HMC has small ϵ and can only explore local space for each 100 steps. This results in a very slow mixing (large autocorrelation even

at 40 lags). On the other hand, smaller $\lambda = 10^3$ has slightly larger constraint violation, but allows much more efficient exploration of the space and excellent mixing performance.

2.4 Correcting Projection to Constraint Space

The Markov chain produced by HMC is geometrically ergodic under very general conditions (Livingstone et al., 2016). With the extrinsic posterior $\pi_K(\theta \mid y)$ as approximation to (2), one may be interested in further obtaining exact posterior in $\mathcal{D}$, likely for two reasons: (1) to strictly uphold the constraints; (2) to erase the loose error control during extrisinc posterior sampling.

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:

$$1 \wedge \frac{\pi(\theta^*_{\mathcal{D}} \mid y)\pi_K(\theta \mid y)}{\pi(\theta_{\mathcal{D}} \mid y)\pi_K(\theta^* \mid y)} = 1 \wedge \frac{L(\theta^*_{\mathcal{D}}; y)\pi_{0,\mathcal{R}}(\theta^*_{\mathcal{D}}) \cdot L(\theta; y)\pi_{0,\mathcal{R}}(\theta)\mathcal{E}(\theta)}{L(\theta_{\mathcal{D}}; y)\pi_{0,\mathcal{R}}(\theta_{\mathcal{D}}) \cdot L(\theta^*; y)\pi_{0,\mathcal{R}}(\theta^*)\mathcal{E}(\theta^*)}. \tag{7}$$

The remaining task is then to optimize the projection with respect to the acceptance rate. Noting

$$|\log(\frac{\pi(\theta^*_{\mathcal{D}} \mid y)\pi_K(\theta \mid y)}{\pi(\theta_{\mathcal{D}} \mid y)\pi_K(\theta^* \mid y)})| \leq |\log(\pi(\theta^*_{\mathcal{D}} \mid y)) - \log(\pi_K(\theta^* \mid y))| + |\log(\pi(\theta_{\mathcal{D}} \mid y)) - \log(\pi_K(\theta \mid y))|, \tag{8}$$

it is sensible choose $\theta_{\mathcal{D}} = M(\theta)$ to minimize the difference $Q(\theta_{\mathcal{D}}) = |\log(\pi(\theta_{\mathcal{D}} \mid y)) - \log(\pi_K(\theta \mid y))| = |\log L(\theta_{\mathcal{D}}; y)\pi_{0,\mathcal{R}}(\theta_{\mathcal{D}}) - \log L(\theta; y)\pi_{0,\mathcal{R}}(\theta)\mathcal{E}(\theta)|$ towards 0 for each sample in the extrinsic posterior. Obviously, when the approximate $\theta \in \mathcal{D}$ exactly, the optimal projection would be the identity function; when $\theta \notin \mathcal{D}$, standard optimization technique can be used.

Continuing the unit circle example, we first obtain $\hat{\theta}_{\mathcal{D}} = \underset{\theta_{\mathcal{D}}:\theta'_{\mathcal{D}}\theta_{\mathcal{D}}=1}{\operatorname{argmin}} |F'\theta_{\mathcal{D}} - \{F'\theta - \lambda(\theta'\theta - 1)^2\}|$ based on the extrinsic posterior sample collected with $\lambda = 10^3$. Then we construct the exact Markov chain and obtained acceptance rate 96.9%.

2.5 Specification of Hyper-parameter

leo: Need some formalization here:.

Two possible ways to formalize the choice for hyper-parameter λ:

1. Compute some similarity measure between the local behavior of the constrained space $\mathcal{D}$ and the embedded space $\mathcal{R}$. Parameterize λ as the function of this similarity. When similarity is low, use smaller λ to expand the posterior support.

2. Use something similar to the compminimax idea. First use minimum acceptance rate to derive a maximally tolerable violation, then optimize the mixing rate.

3 Illustration

Example 1: Ordered Simplex

Example 2: Monotone Spline

Example 3: Orthonormal Gaussian Processes

4 Theory

Posterior properiety of $\pi_{0,\mathcal{R}}(\theta)\mathcal{E}(\theta)$, when $\pi_{0,\mathcal{D}}(\theta)$ is proper but $\pi_{0,\mathcal{R}}(\theta)$ is improper.

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.

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. (2009). Simulation of the matrix bingham–von mises–fisher distribution, with applications to multivariate and relational data. *Journal of Computational and Graphical Statistics 18*(2), 438–456.

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.

Kelly, C. and J. Rice (1990). Monotone smoothing with application to dose-response curves and the assessment of synergism. *Biometrics*, 1071–1085.

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*.

Lin, L., V. Rao, and D. B. Dunson (2016). Bayesian nonparametric inference on the stiefel manifold. *Statistica Sinica*.

Lin, L., B. St Thomas, H. Zhu, and D. B. Dunson (2016). Extrinsic local regression on manifold-valued data. *Journal of the American Statistical Association* (just-accepted).

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.