Constraint Relaxation for Bayesian Modeling with Parameter Constraints

Leo Duan, Alexander L Young, Akihiko Nishimura, David Dunson

Abstract: Prior information often takes the form of parameter constraints. Bayesian methods include such information through prior distributions having constrained support. By using posterior sampling algorithms, one can quantify uncertainty without relying on asymptotic approximations. However, outside of narrow settings, parameter constraints make it difficult to develop new prior and/or efficient posterior sampling algorithms. In this work, we first describe a general approach to utilize the large pool of unconstrained distributions in constrained space, then we propose to relax the parameter support into the neighborhood surrounding constrained space for convenient posterior estimation. The constraint relaxation can be done using data augmentation technique or with an approximation function. General off the shelf posterior sampling algorithms, such as Hamiltonian Monte Carlo (HMC), can then be used directly. We illustrate this approach through multiple examples involving equality and inequality constraints. While existing methods tend to rely on conjugate families or sophisticated reparameterization, our proposed approach frees us up to define new classes of models for constrained problems. We illustrate this through application to a variety of simulated and real datasets.

KEY WORDS: Simplex, Stiefel Manifold, Parameter Expansion

1 Introduction

It is extremely common to have prior information available on parameter contraints in statistical models. For example, one may have prior knowledge that a vector of parameters lies on the probability simplex or satisfies a particular set of inequality constraints. Other common examples include shape constraints on functions, positive semidefiniteness of matrices and orthogonality. There is a very rich literature on optimization subject to parameter contraints. One common approach is to rely on Lagrange and Karush-Kuhn-Tucker multipliers (**?**). However, simply producing a point estimate is often insufficient, as uncertainty quantification (UQ) is a key component of most statistical analyses. Usual large sample asymptotic theory, for example showing asymptotic normality of statistical estimators, tends to break down in constrained inference problems. Instead, limiting distributions may have a complex form that needs to be rederived for each new type of constraint, and may be intractable. An appealing alternative is to rely on Bayesian methods

for UQ, including the constraint through a prior distribution having restricted support, and then applying Markov chain Monte Carlo (MCMC) to avoid the need for large sample approximations. Although MCMC is conceptually simple, except for a few limited cases, it is generally difficult to generate random variable strictly inside constrained space.

To overcome this difficulty, one common strategy is to reparameterize with un-/less constrained parameters at equal or less dimension. The new parameters form functions that can always satisfy the constraint. The transformation, if bijective, is known as 'coordinate system' in manifold embedding literature (**??**). Examples include the polar coordinates for data on a hyper-sphere, or stick-breaking construction for Dirichlet distribution on probability simplex (**?**). One can then directly assign prior on the less constrained parameters. Although this strategy has been successful, convenient coordinate system does not always exist; and heavy reparameterzation tends to makes it more difficult to induce prior property on the original space. For example, uniformity of unconstrained parameter in a compact space may not be equivalent to uniformity on the constrained space via transformation. **?** provide a useful tutorial and cautious guide on this subject.

Alternatively, it is typical to rely on customized solution for specific constraints. One popular strategy is to restrict focus to a prior and likelihood such that posterior sampling is tractable. For example, for modeling of data on Stiefel manifolds, von Mises-Fisher and matrix Bingham-von Mises-Fisher distribution (**??**) are routinely used. Besides limiting consideration to specialized models, another drawback is that the tractable computation, especially posterior conjugacy, tends to break down under common modeling/data complication, such as matrix symmetry, hierarchical structures, etc.

For these reasons, it is appealing to consider approaches that do not rely on conjugate constrained distributions. Early work (**?**) suggested using general unconstrained distribution inside a simple truncated space, and running Gibbs sampling ignoring the constraint but only accepting the draws that fall into truncated space. Unfortunately, this method can be highly inefficient if constrained space has a small or zero measure, which will create a low or zero acceptance probability. A recent idea is to run MCMC ignoring the constraint, and then project draws from the unconstrained posterior to the appropriately constrained space. Such an approach was proposed for generalized linear models with order constraints by **?**, extended to functional data with monotone or unimodal constraints (**?**), and recently modified to nonparametric regression with monotonicity (**?**) or manifold (**?**) constraints. A third independent direction utilizes Hamiltonian Monte Carlo (HMC) that incorporates geometric structure with a Riemannian metric (**?**), making proposals strictly inside the constrained space by solving a large linear system. Although simpler algorithms using geodesic flow were proposed for a few selected constrained space (**?**), compared to the first two strategies that operates in unconstrained space, strictly accommodating the constrained geometry tend to require more customization, such as computing the metric tensor for different manifolds.

The goal of this article is to dramatically expand the families of constrained priors one could use and develop simple computational strategy for more general constraints. We first introduce a general strategy to adapt common existing distributions into constrained space. To enable simple posterior computation, we *relax* the parameter into the neighborhood surrounding the constrained space. This approach enjoys the advantages of unconstrained space sampling while approximately takes into account of the geometry of the constrained space. This relaxation either produces an approximation for posterior under general constraints formed by equality and/or inequality, or an exact solution for several common constrained space such as simplex and Stiefel manifolds. Theoretic studies are conducted and comparison with existing approaches are shown in simulations and data applications.

Alex:

2 Constrained Relaxation Methodology

Suppose θ is an $\mathcal{R}$-valued random variable with $\dim(\mathcal{R}) = r < \infty$ and that θ is subject to some constraints which restrict it to a subset $\mathcal{D} \subset \mathcal{R}$. In the Bayesian setting, of principle interest here, θ is a parameter which is known to satisfy some constraints such that it resides in $\mathcal{D}$. In this case, a common approach is to choose a prior distribution with support $\mathcal{D}$. However, aside from some special cases, a suitable choice of prior may be limited. Leo: In my opinion, constructing/choosing prior for constrained space is a separate issue, which does not involve 'relaxation', so perhaps we can have a short section before and motivated for more computational stuffs. Moreover, sampling θ from the constrained space, when possible, may be difficult or computationally intractable.

One potential strategy to alleviate this issue is to construct an approximate distribution which places a high probability on $\mathcal{D}$ but has support in $\mathcal{R}$ by 'relaxing' the constraints. As a motivating example, consider the case where θ has density $\pi_{\mathcal{R}}(\theta)$ with support $\mathcal{R}$ and $\mathcal{D}$ is a measureable subset with positive measure. The posterior density of θ given data Y and $\theta \in \mathcal{D}$ is,

$$\pi(\theta|\theta \in \mathcal{D}, Y) \propto \mathcal{L}(\theta; Y)\pi_{\mathcal{R}}(\theta)\mathbb{1}_{\mathcal{D}}(\theta)$$

for some likelihood function $\mathcal{L}(\theta; Y)$ and data Y. As an approximation, suppose we used the density

$$\tilde{\pi}(\theta) \propto \mathcal{L}(\theta; Y)\pi_{\mathcal{R}}(\theta) \exp\left(-\frac{1}{\lambda} v_{\mathcal{D}}(\theta) \right) \tag{1}$$

where $v_{\mathcal{D}}(\theta) = \inf_{x \in \mathcal{D}} ||\theta - x||$ is a measure of the distance from θ to the constrained space for some metric $||\cdot||$.

Note that $\mathbb{1}_{\mathcal{D}}(\theta)$ is the pointwise limit of $\exp\big(-\nu_{\mathcal{D}}(\theta)/\lambda\big)$ (except perhaps on the boundary of $\mathcal{D}$) as $\lambda \to 0^+$. However, (1) has support $\mathcal{R}$ for all $\lambda > 0$, hence 'relaxing' the constraint. Since (1) is supported on

$\mathcal{R}$ it is more suitable for off-the-shelf MCMC sampling strategies. Ideally, one would hope that samples from (1) could be easily generated and that they would behave as if drawn from the fully conditioned distribution when λ is sufficiently small. We consider this approach when adapted to a number of settings, but generally we refer to it as constraint relaxation (**CORE**).

These observations motivate a number of questions about **CORE** which we investigate in the article. (i) For what types of distributions and constraints is CORE suitable? (ii) Is there a general approach for constructing the 'relaxed' constraint? (iii) How well do samples from the relaxed constraint represent those from the fully conditioned distribution? (iv) How does the approximation depend on the tuning parameter λ?

The answers to (ii) - (iv) depend largely upon (i). Therefore, beginning with (i), we assume θ is a continuous random variable (e.g. $\mathcal{R}$ is $\mathbb{R}^d$, $[0,\infty)^d$, $\mathbb{R}^{n\times k}$) and θ has an unconstrained prior density $\pi_{\mathcal{R}}(\theta)$ which is absolutely continuous with respect to Lebesque measure on $\mathcal{R}$ hereby denoted as $\mu_{\mathcal{R}}$. We investigate two general types of constraints.

First, we consider the simpler case where $\mathcal{D}$ has positive measure, i.e. $\int_{\mathcal{D}} \pi(\theta)\mu_{\mathcal{R}}(d\theta) > 0$. Leo: Generally, Inequality constraints (e.g. $a_i < \theta_i < b_i$, $||\theta||_2^2 < 1$) fall into this category. The analysis in this case will be more straightforward as we can follow traditional approaches to conditional probability. Here, the construction of the relaxed constraint will follow the form of Eq. (1).

Secondly, we consider the case where $\mathcal{D}$ is a measure zero subset of $\mathcal{R}$. In particular, we restrict ourselves to the setting where $\mathcal{D}$ can be represented implicitly as the solution set of a consistent system of equations $\{\nu_i(\theta) = 0\}_{i=1}^s$ so that $\mathcal{D} = \{\theta | v_j(\theta) = 0,\, j = 1,\ldots,s\}$ is a co-dimension s submanifold of $\mathcal{R}$. For a given constrained space, $\mathcal{D}$, there may be multiple choices of the constraints v_i. However, there are technical requirements, discussed in Section 2.1, which limit the potential choice of the constraint questions Leo: perhaps avoid cross-referencing later section, by simplifying this sentence to 'There are some limitations on the types of constraint one could use, however we note that ...'. While these criteria may seem restrictive, we note that many common constraints (e.g. $||\theta||^2 = 1$, $\sum_i \theta_i = 1$, $\theta \in V_k(\mathbb{R}^n)$ where $V_k(\mathbb{R}^n)$ is the Stiefel manifold) fall into this category. In this case, the conditional distribution of θ given $\theta \in \mathcal{D}$, must be handled with care since $\int_D \pi_R(\theta) d\mu_{\mathcal{R}}(\theta) = 0$. However, the requirement that $\mathcal{D}$ has codimension s will serve two purposes. First, it will make the construction of conditional distributions on $\mathcal{D}$ using the tools of geometric measure theory more intuitive. Secondly, it will motivate a general strategy for choosing appropriate constraint equations and in constructing a relaxed density similar to (1).

3 Analytical Results

3.1 Positive measure constrained spaces

Theorem 1. *Suppose $g \in \mathbb{L}^1(\mathcal{R}, \pi_{\mathcal{R}} d\mu_{\mathcal{R}})$ and that $\pi_{\mathcal{D}}$ and $\tilde{\pi}_\lambda$ are taken as above. Then,*

$$\left| E[g(\theta)|\theta \in \mathcal{D}] - \tilde{E}[g(\theta)] \right| \leq \frac{\int_{\mathcal{R}\setminus\mathcal{D}} (E|g(\theta)| + |g(\theta)|)\mathcal{L}(\theta; Y)\pi_{\mathcal{R}}(\theta) \exp(-v(\theta)/\lambda) d\mu_{\mathcal{R}}(\theta)}{\left[\int_{\mathcal{D}} \mathcal{L}(\theta; Y)\pi_{\mathcal{R}}(\theta) d\mu_{\mathcal{R}}(\theta)\right]^2}$$

where $\tilde{E}[g(\theta)] = \int_{\mathcal{R}} g(\theta)\tilde{\pi}_\lambda(\theta) d\mu_{\mathcal{R}}(\theta)$ is the expected value of $g(\theta)$ with respect to the relaxed density $\tilde{\pi}_\lambda$.

Corollary 1. *Suppose $g \in \mathbb{L}^2(\mathcal{R}, \pi_{\mathcal{R}} d\mu_{\mathcal{R}})$, $\pi_{\mathcal{D}}$ and $\tilde{\pi}_\lambda$ are as above, and $v_d(\theta)$ has the form of Eq. (??) with $k = 2$. Then for $0 < \lambda \ll 1$,*

$$\left| E[g(\theta)|\theta \in \mathcal{D}] - \tilde{E}[g(\theta)] \right| = O(\lambda |\log(\lambda)|).$$

3.2 Measure zero constrained spaces

Theorem 2. *Alex: We'll need to decide on a choice of notation for this definition:*

Assume that $J(v(\theta)) > 0$ and that for each $z \in \mathbb{R}^s$ there is a finite non-negative integer p_z such that,

$$m^{p_z}(z) = \int_{\mathbb{R}^s} \frac{\mathcal{L}(\theta; Y)\pi_{\mathcal{R}}(\theta)\mathbb{1}_{v(\theta)=z}}{J(v(\theta))} d\bar{\mathcal{H}}^p(z) \in (0, \infty).$$

Then, for any Borel subset, E, of $\mathcal{R}$, it follows that

$$P(E|v(\theta) = z) = \begin{cases} \frac{1}{m^{p_z}(z)} \int_E \frac{\mathcal{L}(\theta;Y)\pi_{\mathcal{R}}(\theta)\mathbb{1}_{v(\theta)=z}}{J(v(\theta))} d\bar{\mathcal{H}}^p(z) & m^s(z) \in (0, \infty) \\ \delta(E) & m^p(z) \in \{0, \infty\} \end{cases}$$

is a valid regular conditional probability for $\theta \in \mathcal{D}$. Here, $\delta(E) = 1$ if $0 \in E$ and 0 otherwise. Leo: I suggest we use the r and $(r - s)$ for dimensionality as before, this changes to:

Assume that $J(v(\theta)) > 0$ and that there exists $z \in \mathbb{R}^s$ such that,

$$m^s(z) = \int_{\mathbb{R}^{r-s}} \frac{\mathcal{L}(\theta; Y)\pi_{\mathcal{R}}(\theta)\mathbb{1}_{v(\theta)=z}}{J(v(\theta))} d\bar{\mathcal{H}}^{(r-s)}(\theta) \in (0, \infty).$$

Then, for any Borel subset, E, of $\mathcal{R}$, it follows that

$$P(E|v(\theta) = z) = \begin{cases} \frac{1}{m^s(z)} \int_E \frac{\mathcal{L}(\theta;Y)\pi_{\mathcal{R}}(\theta)\mathbb{1}_{v(\theta)=z}}{J(v(\theta))} d\bar{\mathcal{H}}^{(r-s)}(\theta) & m^s(z) \in (0, \infty) \\ \delta(E) & m^s(z) \in \{0, \infty\} \end{cases}$$

is a valid regular conditional probability for $\theta \in \mathcal{D}$. Here, $\delta(E) = 1$ if $0 \in E$ and 0 otherwise.

Theorem 3. *Let $m : \mathbb{R}^s \to \mathbb{R}$ and $m_g : \mathbb{R}^s \to \mathbb{R}$ be defined as follows*

$$m(x) = \int_{\nu^{-1}(x)} \frac{\pi_{\mathcal{H}}(\theta)\mathcal{L}(y;\theta)}{J(\nu(\theta))} d\mathcal{R}^{r-s}(\theta)$$

$$m_g(x) = \int_{\nu^{-1}(x)} g(\theta) \frac{\pi_{\mathcal{H}}(\theta)\mathcal{L}(y;\theta)}{J(\nu(\theta))} d\mathcal{R}^{r-s}(\theta).$$

Suppose that both m and m_g are continuous on an open interval containing the origin and that $g \in \mathbb{L}^1(\mathcal{R}, \pi_R\mathcal{L}(y;\theta)d\mu_{\mathcal{R}}) \cap \mathbb{L}^1(\mathcal{D}, \frac{\pi_R\mathcal{L}(y;\theta)}{J(\nu(\theta))}d\mu_{\mathcal{R}})$. Then,

$$E_{\tilde{\Pi}}[g] \to E[g|\theta \in \mathcal{D}].$$

Corollary 2. *In addition to the assumptions of Theorem 3, suppose that both m and m_g are differentiable at 0. Then*

$$|E_{\tilde{\Pi}}[g] - E[g|\theta \in \mathcal{D}]| = O()$$

as $\lambda \to 0^+$.

4 Posterior Computation

Adapting unconstrained density into space $\mathcal{D}$ often disrupts its posterior conjugacy. Since one can now sample the posterior in $\mathcal{R}$ using CORE, one can exploit conventional sampling tools such as slice sampling, adaptive Metropolis-Hastings and Hamiltonian Monte Carlo (HMC). In this section, we focus on HMC for its easiness to use and good performance in block updating of parameters.

4.1 Hamiltonian Monte Carlo under Constraint Relaxation

We provide a brief overview of HMC for continuous θ^* under constraint relaxation. Discrete extension is possible via recent work of **?**. For easy notation, we use q to represent θ^* under approximation-CORE (**??**), and $\{\theta^*, w\}$ under DA-CORE (**??**).

In order to sample q, HMC introduces an auxillary momentum variable $p \sim \text{No}(0, M)$. The covariance matrix M is referred to as a *mass matrix* and is typically chosen to be the identity or adapted to approximate the inverse covariance of q. HMC then sample from the joint target density $\pi(q,p) = \pi(q)\pi(p) \propto \exp(-H(q,p))$ where, in the case of the posterior under relaxation,

$$\begin{aligned} &H(q,p) = U(q) + K(p), \\ &\text{where } U(q) = -\log \pi(q), \\ &K(p) = \frac{p' M^{-1} p}{2}. \end{aligned} \tag{2}$$

with $\pi(q)$ is the unnormalized density in (**??**) or (**??**).

From the current state $(q^{(0)}, p^{(0)})$, HMC generates a proposal for Metropolis-Hastings algorithm by simulating Hamiltonian dynamics, which is defined by a differential equation:

$$\begin{aligned} \frac{\partial q^{(t)}}{\partial t} &= \frac{\partial H(q,p)}{\partial p} = M^{-1}p, \\ \frac{\partial p^{(t)}}{\partial t} &= -\frac{\partial H(q,p)}{\partial q} = -\frac{\partial U(q)}{\partial q}. \end{aligned} \tag{3}$$

The exact solution to (3) is typically intractable but a valid Metropolis proposal can be generated by numerically approximating (3) with a reversible and volume-preserving integrator (**?**). The standard choice is the *leapfrog* integrator which approximates the evolution $(q^{(t)}, p^{(t)}) \to (q^{(t+\epsilon)}, p^{(t+\epsilon)})$ through the following update equations:

$$p \leftarrow p - \frac{\epsilon}{2}\frac{\partial U}{\partial q}, \quad q \leftarrow q + \epsilon M^{-1}p, \quad p \leftarrow p - \frac{\epsilon}{2}\frac{\partial U}{\partial q} \tag{4}$$

Taking L leapfrog steps from the current state $(q^{(0)}, p^{(0)})$ generates a proposal $(q^*, p^*) \approx (q^{(L\epsilon)}, p^{(L\epsilon)})$, which is accepted with the probability

$$1 \wedge \exp\Big(-H(q^*, p^*) + H(q^{(0)}, p^{(0)}))\Big)$$

4.2 Computing Efficiency and Support Expansion

Since CORE expands the support from $\mathcal{D}$ to $\mathcal{R}$, it is useful to study the effect of space expansion on the computing efficiency. In this section, we provide some quantification of the effects and provide a practical guidance on choosing $\pi(w)$ or λ in the two strategies.

In understanding the computational efficiency of HMC, it is useful to consider the number of leapfrog steps to be a function of ϵ and set $L = \lfloor \tau/\epsilon \rfloor$ for a fixed integration time $\tau > 0$. In this case, the mixing rate of HMC is completely determined by τ in the limit $\epsilon \to 0$ (**?**). In practice, while a smaller stepsize ϵ leads to a more accurate numerical approximation of Hamiltonian dynamics and hence a higher acceptance rate, it takes a larger number of leapfrog steps and gradient evaluations to achieve good mixing. For an optimal computational efficiency of HMC, therefore, the stepsize ϵ should be chosen only as small as needed to achieve a reasonable acceptance rate (**??**). A critical factor in determining a reasonable stepsize is the *stability limit* of the leapfrog integrator (**?**). When ϵ exceeds this limit, the approximation becomes unstable and the acceptance rate drops dramatically. Below the stability limit, the acceptance rate $a(\epsilon)$ of HMC increases to 1 quite rapidly as $\epsilon \to 0$ and in fact satisfies $a(\epsilon) = 1 - \mathcal{O}(\epsilon^4)$ (**?**).

For simplicity, the following discussions assume the mass matrix M is taken to be the identity. Let $\mathbf{H}_U(q)$ denote the hessian matrix of $U(q) = -\log \pi(q)$ and let $\xi_1(q)$ denotes the first largest eigenvalue of $\mathbf{H}_U(q)$. While analyzing stability and accuracy of an integrator is highly problem specific, the linear stability analysis and empirical evidences suggest that, for stable approximation of Hamiltonian dynamics by the leapfrog integrator in $\mathbb{R}^p$, the condition $\epsilon < 2\xi_1(\theta)^{-1/2}$ must hold on most regions of the parameter space (**?**). Besides the eigenvalue, if the support of q is a constrained space $\mathcal{Q}$, another limiting factor is roughly the shortest distance to the boundary $\eta(\theta; \mathcal{Q}) = \inf_{q' \notin \mathcal{Q}} \|q' - q\|$. If either $\eta(\theta; \mathcal{Q})$ or $\xi_1(\theta)^{-1/2}$ is close to 0, the upper bound would be too low to obtain efficient computation. In constrained model, the parameter space $\mathcal{D}$ can have very small $\eta(\theta; \mathcal{D})$. Constraint relaxation can reduce this problem via support expansion.

For approximation-CORE (**??**), to control approximation error, one can choose to relax a subset of constringent constraints. Observing $\mathcal{D} = \cap_k D_k$, each approximation $\exp(-\frac{|v_k(\theta^*)|^\alpha}{\lambda_k})$ corresponds to a constrained space $\mathcal{D}_k$. One practical strategy is that, for $\mathcal{D}_k$'s with $\eta(\theta; \mathcal{D}_k) \approx 0$, one uses moderate λ_k to induce some support expansion (denoted by $\lambda_k \geq \zeta$ with ζ moderately small but not too close to 0); for $\mathcal{D}_k$'s without this issue, one uses very small $\lambda_k \approx 0$ to almost always uphold the constraint. The latter was also suggested by **?** as creating a high 'energy wall'. Noting this could create inaccuracy of HMC near the boundary with $\lambda_k \approx 0$ under fixed step size, we use random step size ϵ at each iteration to reduce the error.

The Hessian $\mathbf{H}_U(q)$ under approximation-CORE is given by

$$\mathbf{H}_U(q) = -\mathbf{H}_{\log L(y;\theta^*)\pi_{\mathcal{R}}(\theta^*)/J(v(\theta^*))}(\theta^*) + \sum_k \lambda_k^{-1} \mathbf{H}_{|v_k|^\alpha}(\theta^*) \mathbb{1}_{\theta \notin \mathcal{D}_k}, \tag{5}$$

where the second term $\lambda_k^{-1} \mathbf{H}_{v_k}(\theta) \mathbb{1}_{\theta \notin \mathcal{D}_k}$ is 0 unless $\theta \notin \mathcal{D}_k$. Since the λ_k^{-1}'s in the second term often dominate the eigenvalue, hence $\xi_1^{-1/2}(\theta^*) \approx \min_{\lambda_k \geq \zeta} \lambda_k^{1/2}$. A trade-off between approximation accuracy and computational efficiency is involved. Fortunately, as quantified above, the approximation error is often $\mathcal{O}(\max_{\lambda_k \geq \zeta} \lambda_k)$ and decreases faster than the efficiceny cap $\mathcal{O}(\min_{\lambda_k \geq \zeta} \lambda_k^{1/2})$, as λ_k decreases. In our experiments, we did find changing from $\lambda_k = 10^{-4}$ to 10^{-5} requires approximately 3 times of computing budget, due the effect on stability bound.

On the other hand, since DA-CORE (**??**) does not involve such error trade-off, it is preferred when it is applicable. Letting $\mathcal{Q}_\theta \subset \mathcal{D}$ denote the support for the constrained $\theta \in \mathcal{D}$, the reparameterization changes the support to $Q_{\theta^*} = \{g(\theta; w) : \theta \in \mathcal{Q}\}$. Therefore, one could choose $\pi(w)$ to substantially increase $\eta(\theta^*; Q_{\theta^*})$. Since the augmented variable w is redundant, DA-CORE can be considered as one type of parameter expansion discovered by **?**, who originally focused on accelerating Gibbs sampling of probit regression. Although for greater space expansion, it is possible to use diffuse or even improper prior for $\pi(w)$ on Q_{θ^*}, we recommend

assigning $\pi(w)$ loosely centered at $w_0 : g(\theta; w_0) = \theta$ (corresponding to when $\theta^* = \theta$), which makes θ^* a mild relaxation of θ. This ensures no substantial change in $\xi_1^{-1/2}(\theta^*)$ in HMC.

5 Simulations

In order to compare against existing approaches on computing efficiency and provide empirical evidence supporting our previous result, we run simulations on several toy examples in this section.

5.1 Gaussian under Linear Inequality

We first consider linear models under linear inequality constraints. Although recent work proposed a new customized prior with posterior conjugacy (**?**), via our framework, one can simply exploit general Gaussian prior. We sample a bivariate Gaussian $\theta \sim \text{No}\left(\mu, I\sigma^2\right)$ subject to linear inequality $\theta \in (0,1)^2, \theta_1 + \theta_2 < 1$, which forms a triangle. In two separate settings, choosing (μ, σ^2) as $([0.3, 0.3], 1/10)$ induces a wide-spread distribution centered in the interior of $\mathcal{D}$, while $([0.7, 0.3]', 1/10^4)$ induces a distribution concentrated on the boundary of $\mathcal{D}$. As DA-CORE does not appear straightforward in this case, we use the approximation-CORE (**??**) with $\exp(-\frac{|v(\theta)|}{\lambda})$, $v(\theta) = |\theta_1 + \theta_2 - 1|_+ + |-\theta_1|_+ + |-\theta_2|_+ + |\theta_1 - 1|_+ + |\theta_2 - 1|_+$. Since the triangle has wide support with $\eta(\theta; \mathcal{D})$ away from 0, small $\lambda = 10^{-8}$ guarantees almost no approximation error. Figure 1 plots the posterior sample and its contour. Clearly, all posterior fall inside $\mathcal{D}$. To compare, we ran simple rejection sampling with untruncated normal proposal $\text{No}(\mu, I\sigma^2)$. As expected, it suffers from a rapidly growing rejection rate from 12% to 51%, as μ moves further away from the center of $\mathcal{D}$.

(a) Constrained $\text{No}([0.3, 0.3], 1/10)$ (b) Constrained $\text{No}([0.7, 0.3], 1/10^4)$

Figure 1: Posterior sample of bivariate normal distribution subject to linear inequality constraints $\theta \in (0,1)^2, \theta_1 + \theta_2 < 1$, using HMC with constraint relaxation. Posterior is spread out around the center (panel (a)) or concentrated on the boundary (panel (b)) of the region.

5.2 von Mises–Fisher on Unit Circle

To illustrate equality constraint relaxation, we generate a simple von Mises–Fisher distribution $\pi_{\mathcal{D}}(\theta) \propto \exp(F'\theta)$ on a unit circle $\{(\theta_1, \theta_2) : \theta_1^2 + \theta_2^2 = 1\}$. We use $F = (5, 5)$ to induce a relatively spread-out θ on the manifold. For sampling, we compare three strategies: approximate-CORE using $\exp(-\frac{|\theta'\theta - 1|}{\lambda})$ for approximating the indicator, DA-CORE using $\theta_1 = \frac{\theta_1^*}{w}, w = \sqrt{(\theta_1^*)^2 + (\theta_2^*)^2}$ and $\pi(w) \sim \text{No}(1,1)\mathbb{1}_{w>0}$ and exact von Mises–Fisher obtained using 'movMF' package.

Unlike the previous linear inequality constraint, the unit circle has narrow $\eta(\theta; \mathcal{D}) = 0$ for all $\theta \in \mathcal{D}$, therefore, some support expansion is needed for HMC. We test $\lambda = 10^{-3}$, 10^{-4} and 10^{-5} for approximation-

CORE. To compare the efficiency of HMC, we fix the number of leap-frog steps to 20 within one iteration HMC, and let software STAN automatically tune for stable step size. Table 1 shows the effective sample size per 1000 iterations, the effective 'violation' $|v(\theta)| = |\theta_1^2 + \theta_2^2 - 1|$ and the 1-Wasserstein distance W_1 as the approximation error. As W_1 is numerically computed, to provide a baseline error, we also calculate the average W_1 comparing two independent samples from the same exact distribution. The approximation error W_1 based on $\lambda = 10^{-5}$ approximation is indistinguishable from this low numerical error, while the other approximations have slightly larger error but more effective samples. As expected, the DA-CORE is exact and has high effective sample size.

	HMC based on CORE				Exact
	Approximation			DA-CORE	
	$\lambda = 10^{-3}$	$\lambda = 10^{-4}$	$\lambda = 10^{-5}$		
W_1	0.050 (0.019, 0.095)	0.034 (0.027, 0.037)	0.014 (0.013,0.025)	0.017 (0.0012,0.026)	0.015 (0.0014,0.025)
$\lvert v(\theta)\rvert \mid y$	9×10^{-4} $(2.6 \cdot 10^{-5}, 3.3 \cdot 10^{-3})$	9×10^{-5} $(2.0 \cdot 10^{-6}, 3.4 \cdot 10^{-4})$	9×10^{-6} $(2.7 \cdot 10^{-7}, 3.5 \cdot 10^{-5})$	0	0
ESS /1000 Iterations	751.48	260.54	57.10	788.30	

Table 1: Benchmark of constraint relaxation methods on sampling von–Mises Fisher distribution on a unit circle. For each approximation CORE, average approximation error (with 95% credible interval, out of 10 repeated experiments) is computed, and numeric error of W_1 is shown under column 'exact' as comparing two independent copies from the exact distribution. Effective sample size shows DA-CORE and approximation-CORE with relatively large λ have high computing efficiency.

5.3 Dirichlet on a Simplex

Lastly, we experiment with a particularly challenging distribution on a $(p-1)$-simplex, defined by $\{\theta : \theta \in (0,\infty)^p, \sum_{i=1}^p \theta_i = 1\}$. We consider Dirichlet distribution $\mathrm{Dir}(\alpha)$, with $\pi_{\mathcal{D}}(\theta) \propto \prod_{i=1}^p \theta_i^{\alpha-1}$. When the concentration parameter $\alpha < 1$, $\mathrm{Dir}(\alpha)$ exhibits sparse property that some θ_i's become very close to 0, which is exploited in topic modeling (**?**) and shrinkage (**?**) literature. Despite the simple form, the computation can be quite difficult if there is large uncerntainty associated with θ on top of sparsity. The distribution will be multi-modal with distribution scattered along the boundary of the simplex (Figure 2(a)).

To illustrate, we consider $p = 3$ and various values of $\alpha \in \{1, 0.5, 0.1, 0.01\}$. We test the performance of approximation-CORE and DA-CORE. To compare, we also test the standard HMC using coordinate system $\theta_1 = \cos^2(\theta_1^*), \theta_2 = \sin^2(\theta_1^*)\cos^2(\theta_2^*), \theta_3 = \sin^2(\theta_1^*)\sin^2(\theta_2^*)$ for $\theta^* \in (0, 2\pi)^2$, which is equivalent to stick-breaking representation (**?**); and the geodesic HMC utilizing the geometric flow directly on the simplex (**?**). For all HMCs, we fix the number of steps in each iteration to be 30 and tune the step size to have effective sample size as large as possible.

Table 2 lists the effective sample sizes under different α's. As α becomes smaller than 1, approximation-CORE and geodesic HMC become worse in performance, while DA-CORE and coordinate system are much less impacted. Figure 2(b) shows at $\alpha = 0.01$, the approximation-CORE and geodesic HMC are stuck for a long time, while DA-CORE works substantially better. As a well-tested reparameterization, HMC based on

coordinate system still works acceptably well in this case.

The difficulty that approximation-CORE encountered was anticipated. **?** have previously reported similar slow-down of geodesic HMC computing on hyper-Dirichlet distribution (**?**) with $\alpha < 1$. Comparing these two approaches, geodesic HMC relies on restricting the kinetic flow on $\mathcal{D}$ via its product with the metric tensor, and approximation-CORE relies on creating high energy wall in the potential energy. The latter can be viewed as an approximation to the former, which explains the similarity in performance.

(a) $2,000$ samples from Dir(0.01) on 2-simplex. (b) Traceplot of θ_1 using 4 types of HMCs.

Figure 2: Sampling of Dirichlet on an simplex with distribution concentrated on the boundaries. Panel(a) illustrates the distribution under Dir(0.01); Panel(b) compares the traceplots of 4 different types of HMCs, which are based on: approximation-CORE with $\lambda = 10^{-3}$, DA-CORE, geodesic flow on simplex (**?**) and coordinate system.

	HMC based on CORE			Geodesic HMC	Coord System HMC
	Approx $\lambda = 10^{-3}$	Approx $\lambda = 10^{-4}$	DA		
ESS /1000 Iter. ($\alpha = 1$)	511.43	146.07	947.53	174.14	**961.08**
ESS /1000 Iter. ($\alpha = 0.5$)	145.15	33.16	**912.94**	31.47	846.92
ESS /1000 Iter. ($\alpha = 0.1$)	88.32	26.88	**992.75**	28.70	875.83
ESS /1000 Iter. ($\alpha = 0.01$)	20.54	3.91	**722.44**	17.26	128.55

Table 2: Average effective sample size per 1000 iterations in Dir(α), under different α.

6 Application: Finding Sparse Basis in a Population of Networks

We now consider a real data application in brain network analysis. The brain connectivity structures are obtained in the data set KKI-42 (Landman et al. 2011), which consists of $n = 21$ healthy subjects without any history of neurological disease. We take the first scan out of the scan-rescan data as the input. Each observation is a $V \times V$ symmetric network, recorded as an adjacency matrix A_i for $i = 1, \ldots, n$. The regions are constructed via the Desikan et al. (2006) atlas, for a total of V = 68 nodes. For the ith matrix A_i, $A_{i,k,l} \in \{0, 1\}$ is the element on the kth row and lth column of A_i, with $A_{(i,k,l)} = 1$ indicating there is an connection between kth and lth region, $A_{(i,k,l)} = 0$ if there is no connection. The matrix is symmetric due to the undirectedness of the network, but the diagonal records $A_{(i,k,k)}$ for all i and k are missing due to the lack of meaning for self-connectivity.

One scientific interest in neuroscience is to quantify the variation of brain networks and identify the regions (nodes) that contribute to it. Extending factor analysis to multiple matrices, one appealing approach is to have the networks share a common factor matrix but let the loadings vary across subjects. This can be considered as a simplified equivalent of three-way tensor factorization (**?**). Then to selectively identify the important nodes, one natural way is to apply shrinkage on the elements of factor matrix.

Geometrically, the factor matrix, denoted by $\{U_1, \ldots, U_d\}$, reside on a Stiefel manifold $\mathcal{V}(n, d) = \{U :$

$U'U = I_d\}$, where $U = [U_1, \ldots, U_d]$ is the $n \times d$ matrix. Using r to index $1, \ldots, d$, each frame U_r represents a $(n-1)$-hypersphere. Applying shrinkage forces some of its sub-coordinates to be close to 0, which is reducing each U_r onto a lower-dimensional hypersphere. Although previous work was done using sparse PCA (**?**) for continuous outcome, little work has been done in a probabilistic model for binary matrices.

To apply shrinkage in the constrained space, we adopt the induced prior as common in Bayesian literature (reviewed by **?**), which usually takes the form hierarchical structure $\theta_i \mid \kappa_i, \sigma \sim \text{No}(0, \kappa_i\sigma), \quad \kappa_i \sim G_1, \quad \sigma \sim G_2$ with κ_i, σ as the local and global scale parameters. However, when constraining θ_i, one caveat would be only adapting the conditional density $\text{No}(\theta_i; \kappa_i\sigma)$, which yields intractable normalizing constant involving $\kappa_i\sigma$ in the conditional. This difficulty can be avoided by reparameterizing $\theta_i = \eta_i\kappa_i\sigma$ with $\eta_i \sim \text{No}(0, 1)$, and adapting the *joint* density of $\{\eta_i, \kappa_i, \sigma\}$ on constrained space instead. The joint density will not have intractable constant as long as the hyper-parameters in G_1 and G_2 are fixed.

We now take the Dirichlet-Laplace prior (**?**) as unconstrained distribution $\pi_{\mathcal{R}}$ and adapt it onto Stiefel manifold via (**??**).

$$
\begin{aligned}
&A_{(i,k,l)} \sim \text{Bern}(\frac{1}{1+\exp(-\psi_{(i,k,l)} - z_{(k,l)})}) \\
&\psi_{(i,k,l)} = \sum_{r=1}^{d} v_{(i,r)} u_{(k,r)} u_{(l,r)} \\
&U'U = I_d \text{ with } U = \{u_{(k,r)}\}_{k=1,\ldots,n;r=1,\ldots,d} \\
&u_{(k,r)} = \eta_{(k,r)}\kappa_{(k,r)}\sigma_u \\
&\eta_{(k,r)} \sim \text{Lap}(0,1), \quad \{\kappa_{(1,r)} \ldots \kappa_{(V,r)}\} \sim \text{Dir}(\alpha), \quad \sigma_u^2 \sim \text{IG}(2,1) \\
&z_{(k,l)} \sim \text{No}(0, \sigma_z^2), \quad \sigma_z^2 \sim \text{IG}(2,1) \\
&v_{(i,r)} \sim \text{No}(0, \sigma_{v,(r)}^2), \quad \sigma_{v,(r)}^2 \sim \text{IG}(2,1)
\end{aligned}
$$

for $k > l$, $k = 2, \ldots, V$, $i = 1, \ldots, n$; $\text{Lap}(0, 1)$ denotes the Laplace distribution centered at 0 with scale 1; $Z = \{z_{(k,l)}\}_{k=1,\ldots,V;l=1,\ldots,V}$ is a symmetric unstructured matrix that serves as the latent mean; $\{v_{(i,r)}\}_{r=1,\ldots,d}$ is the loading for the ith network, with each $v_{(i,r)} > 0$; for all other scale parameters $\sigma_\cdot^2$, we choose weakly informative prior inverse Gamma $\text{IG}(2, 1)$, as appropriate for the scale under the logistic link. To induce sparsity in each Dirichlet, we use $\alpha = 0.1$ as suggested by **?**.

There are two types of constraints in the model, $U'U = I_d$ and $\sum_{k=1}^{V} \kappa_{(k,r)} = 1$ for $r = 1, \ldots, d$. Taking $v_1(U) = U'U - I_d$ and $v_2(\kappa_{(k,r)}) = \sum_{k=1}^{V} \kappa_{(k,r)} - 1$ for each r, the Jacobian is constant in (**??**). For posterior computation, we use DA-CORE as described above. Using latent variable w_U d-by-d upper triangular and

positive diagonal matrix, and $w_{\kappa,(r)} > 0$ for $r = 1, \ldots, d$, we relax the parameters to

$$U^* = U w_U, \quad \kappa^*_{(k,r)} = \kappa_{(k,r)} w_{\kappa,(r)},$$

which yields re-parameterization via projection

$$\begin{aligned} U &= U^* w_U^{-1}, \quad w_U = \text{QR.R}(U^*), \\ \kappa_{(k,r)} &= \frac{\kappa^*_{(k,r)}}{w_{\kappa,(r)}}, \quad w_{\kappa,(r)} = \sum_{k=1}^{V} \kappa^*_{(k,r)} \\ \eta_{k,r} &= \frac{u_{(k,r)}}{\kappa_{(k,r)} \sigma_u}, \end{aligned} \tag{6}$$

where QR.R denotes the function that outputs R matrix in QR decomposition. To control the amount of relaxation, we assign w_U near I_d via $\pi(w_U) \propto \text{etr}\left[-\frac{(w_U - I_d)'(w_U - I_d)}{\lambda}\right]$ and $w_{\kappa,(r)}$ near 1 via $\pi(w_{\kappa,(r)}) \propto \exp\left[-\frac{(w_{\kappa,(r)}-1)^2}{\lambda}\right]$ and set $\lambda = 10^{-3}$.

For comparison, we test with the specified model (i) against (ii) the same model except with simple $u_{(k,r)} \sim \text{No}(0, \sigma_u^2)$ instead of the shrinkage prior and (iii) the same model except without the orthonormality constraint $U'U = I$ and the shrinkage prior. We run all models for $10,000$ iterations and discard the first $5,000$ iteration as burn-in. For each iteration, we run 300 leap-frog steps. For efficient computing, we truncated $d = 20$.

Table 3 lists the benchmark results. Compared to (i) and (ii), the unconstrained model (iii) suffers from very low effective sample size, due to the serious convergence issue in the factor matrix U. As explained by previous findings in matrix/tensor factorization (**?**), the factor matrix could scale and rotate without changing the likelihood, and substantial improvement could be obtained by applying orthonormality constraint.

Figure 3(a) plots the posterior mean loadings $v_{(i,r)}$, with each line representing one subject. For all $i = 1, \ldots, 21$, the lines drop quickly to near 0 after $r \geq 5$ in model (i) and (ii), but only do so until $r \geq 10$ in model (iii). This indicates that independent factors are more effective representation of the span, compared to non-orthogonal ones. Clearly, (i) shows more variability than (ii) in the loading $v_{(i,r)}$. We validate these models by calculating area under the receiver operating characteristic curve (AUC) based on the mean predicted probability and the binary outcome $A_{(i,k,l)}$, using the fitted data and the other unused rescan data from the 21 subjects. The models (i) and (ii) with orthonomality constraint perform similarly well, and clearly better than the unconstrained model (iii) in prediction AUC.

Model	(i).with shrinkage & orthonormality	(ii).with orthonormality only	(iii).unconstrained
Fitted AUC	97.9%	97.1%	96.9%
Prediction AUC	96.2%	96.2%	93.6%
ESS /1000 Iterations	193.72	188.10	8.15

Table 3: Comparing 3 models for 21 brain networks

Figure 3(b) compares the models (i) and (ii) over the top 6 frames of U_r, with r re-ordered such that $\sigma^2_{v,(1)} \geq \sigma^2_{v,(2)} \geq \ldots \geq \sigma^2_{v,(d)}$. The posterior of U_1, U_2, U_3 look very similar between the two, whereas U_4, U_5, U_6 have a considerable subset of points close to 0 in the model with shrinkage prior.

(a) Posterior mean of the loadings $v_{i,r}$ for 21 subjects using three models. Each line represents the loadings for one subject over $r = 1, \ldots 10$.

(b) Posterior mean and pointwise 95% credible interval of the factors $U_1, \ldots, U_6$ in the two constrained models.

Figure 3: Loadings and factors estimates of the network models. Panel (a) compares the varying loadings of the subjects in three models; Panel (b) compares the estimated shared factors with and without the shrinkage prior (model (iii) is omitted due to non-convergence in the factors).

7 Discussion

what Parameter constraint often limits the flexibility to develop new model and creates huge burden in developing efficient posterior sampling algorithms. In this article, we develop a formal strategy to utilize the large pool of distributions in the constrained space, and propose a constraint relaxation approach to allow simple implementation for posterior estimation. For common constrained space that can be projected to via a function, we propose an exact algorithm based on data augmentation; for more general problem, we propose an approximation approach. This strategy works well for general equality and inequality constraints.

The future work of this research may include tackling the 'doubly intractable' problem. This issue is common when the data is on the constrained space, or the constrained prior has hyper-parameters to estimate. In the data application, we show that a reparameterization strategy works for some shrinkage priors, but clearly, more general treatment is needed. We expect our work to be compatible to the existing solutions (**???**).

Alex:

A Proofs from Section 3.1

B Proofs from Section 3.2

We have two densities. The first is the fully constrained density for $\theta \in \mathcal{D}$.

$$\pi_{\mathcal{D}}(\theta) = \frac{1}{m_0} \frac{\mathcal{L}(y;\theta)\pi_{\mathcal{R}}(\theta)}{J(\nu(\theta))} \mathbb{1}_{\mathcal{D}}(\theta)$$

where the normalizing constant m_0 is calculated w.r.t. Hausdorff measure

$$m_0 = \int_{\mathcal{R}} \frac{\mathcal{L}(y;\theta)\pi_{\mathcal{R}}(\theta)}{J(\nu(\theta))} \mathbb{1}_{\mathcal{D}}(\theta) d\bar{\mathcal{H}}^{r-s}(\theta).$$

This density defines a probability measure, Π, w.r.t. Hausdorff measure. For a measurable set E,

$$\Pi(E) = \int_E \frac{g(\theta)}{m_0} \frac{\mathcal{L}(y;\theta)\pi_{\mathcal{R}}(\theta)}{J(\nu(\theta))} \mathbb{1}_{\mathcal{D}}(\theta) d\bar{\mathcal{H}}^{r-s}(\theta).$$

Secondly, we have the relaxed distribution

$$\tilde{\pi}_{\mathcal{D}}(\theta) = \frac{1}{m_\lambda} \mathcal{L}(y;\theta)\pi_{\mathcal{R}}(\theta) \exp\left(- \frac{||v(\theta)||_1}{\lambda} \right)$$

where the normalizing constant is calculated w.r.t. Lebesgue measure on $\mathcal{R}$

$$m_\lambda = \int_{\mathcal{R}} \mathcal{L}(y;\theta)\pi_{\mathcal{R}}(\theta) \exp\left(- \frac{||v(\theta)||_1}{\lambda} \right) d\mu_{\mathcal{R}}(\theta).$$

We use $\mu_{\mathcal{R}}$ to denote Lebesgue measure on $\mathcal{R}$. This density defines a probability measure, $\tilde{\Pi}$, w.r.t. $\mu_{\mathcal{R}}$. For a measurable set E,

$$\tilde{\Pi}(E) = \int_E \frac{g(\theta)}{m_\lambda} \mathcal{L}(y;\theta)\pi_{\mathcal{R}}(\theta) \exp\left(- \frac{||v(\theta)||_1}{\lambda} \right) d\mu_{\mathcal{R}}(\theta).$$

Definition of expectation of $g(\theta)$ w.r.t. constrained and relaxed measures

For a given function, g, we can define the exact and approximate expectations of g, respectively E_Π and $E_{\tilde{\Pi}}$, as

$$\begin{aligned}
E_\Pi[g(\theta)|\theta \in \mathcal{D}] &= \int_{\mathcal{R}} \frac{g(\theta)}{m_0} \frac{\mathcal{L}(y;\theta)\pi_{\mathcal{R}}(\theta)}{J(\nu(\theta))} \mathbb{1}_{\mathcal{D}}(\theta) d\bar{\mathcal{H}}^{r-s}(\theta) \\
&= \int_{\mathcal{D}} \frac{g(\theta)}{m_0} \frac{\mathcal{L}(y;\theta)\pi_{\mathcal{R}}(\theta)}{J(\nu(\theta))} d\bar{\mathcal{H}}^{r-s}(\theta) \\
E_{\tilde{\Pi}}[g(\theta)] &= \int_{\mathcal{R}} \frac{g(\theta)}{m_\lambda} \mathcal{L}(y;\theta)\pi_{\mathcal{R}}(\theta) \exp\left(- \frac{||v(\theta)||_1}{\lambda} \right) d\mu_{\mathcal{R}}(\theta) \\
&= \int_{\mathbb{R}^s} \frac{1}{m_\lambda} \int_{\nu^{-1}(x)} g(\theta) \frac{\pi_{\mathcal{R}}(\theta)\mathcal{L}(y;\theta)}{J(\nu(\theta))} \exp\left(- \frac{||\nu(\theta)||_1}{\lambda} \right) d\bar{\mathcal{H}}^{r-s}(\theta) d\mu_{\mathbb{R}^s}(x) \\
&= \int_{\mathbb{R}^s} \frac{\exp\left(- \frac{||x||_1}{\lambda} \right)}{m_\lambda} \int_{\nu^{-1}(x)} g(\theta) \frac{\pi_{\mathcal{R}}(\theta)\mathcal{L}(y;\theta)}{J(\nu(\theta))} d\bar{\mathcal{H}}^{r-s}(\theta) d\mu_{\mathbb{R}^s}(x)
\end{aligned}$$

Let,

$$m(x) = m^{r-s}(x) = \int_{\nu^{-1}(x)} \frac{\pi_{\mathcal{R}}(\theta)\mathcal{L}(y;\theta)}{J(\nu(\theta))} d\bar{\mathcal{H}}^{r-s}(\theta).$$

By construction, $m(x) > 0$ for $\mu_{\mathbb{R}^s}$-a.e. $x \in \mathbb{R}^s$ and $m(0) > 0$. Then by Theorem 1,

$$E[g(\theta)|\nu(\theta) = x] = \frac{1}{m(x)} \int_{\nu^{-1}(x)} g(\theta) \frac{\pi_{\mathcal{R}}(\theta)\mathcal{L}(y;\theta)}{J(\nu(\theta))} d\bar{\mathcal{H}}^{r-s}(\theta). \tag{7}$$

As such, we may may express $E_{\tilde{\Pi}}[g(\theta)]$ as

$$E_{\tilde{\Pi}} = \int_{\mathbb{R}^s} \frac{m(x)}{m_\lambda} \exp\left(-\frac{||x||_1}{\lambda}\right) E\big[g(\theta)|\nu(\theta) = x\big] d\mu_{\mathbb{R}^s}(x). \tag{8}$$

At this point, I'm going to try to demonstrate the asymptotic behavior of (2) for small λ. To write this up properly, we'll need to begin with absolute values of differences and bound above. This will require some statements about the integrability of g. I'm going to wait on writing the argument up in that manner until we determine what can be said about the continuity/uniform continuity/differentiability of m and $E[g|\nu = x]$ near $x = 0$. For now, I'll make the following assumptions.

Assumption: Both $m(x)$ and $E\big[g(\theta)|\nu(\theta) = x\big]$ are continuous at $x = 0$.

Asymptotic behavior of m_λ

Fix $\epsilon > 0$. Since $-\lambda \log(\lambda^{s+1}) \to 0$ as $\lambda \to 0^+$, we may choose λ small enough so that $||m(x) - m(0)||_1 < \epsilon$ for all $x \in B_1(0; \lambda|\log(\lambda^{s+1}|)$ and $\lambda < \max(\epsilon^{1/s}, \epsilon^{1/(s+1)})$, the 1-ball of radius $\lambda|\log(\lambda^{s+1})|$ centered at 0.

Let us first consider the small λ behavior of m_λ. We begin by re-expressing m_λ in terms of $m(x)$.

$$\begin{aligned} m_\lambda &= \int_{\mathcal{R}} \pi_{\mathcal{R}}(\theta)\mathcal{L}(y;\theta) \exp\left(-\frac{||v(\theta)||_1}{\lambda}\right) d\mu_{\mathcal{R}} \\ &= \int_{\mathbb{R}^s} \exp\left(-\frac{||x||_1}{\lambda}\right) \int_{\nu^{-1}(x)} \frac{\pi_{\mathcal{R}}(\theta)\mathcal{L}(y;\theta)}{J(\nu(\theta))} d\bar{\mathcal{H}}^{r-s}(\theta) d\mu_{\mathbb{R}^s}(x) \\ &= \int_{\mathbb{R}^s} m(x) \exp\left(-\frac{||x||_1}{\lambda}\right) \end{aligned}$$

Now, split the above integral into two regions: the interior and exterior of $B_1(0; \lambda|\log(\lambda^{s+1})|)$. Note that

outside of B_1, $\exp(-||x||_1/\lambda) \leq \lambda^{s+1}$.

$$
\begin{aligned}
m_\lambda &= \int_{\mathbb{R}^s \setminus B_1(0;\lambda|\log(\lambda^{s+1})|)} m(x) \exp\left(-\frac{||x||_1}{\lambda}\right) d\mu_{\mathbb{R}^s}(x) + \int_{B_1(0;\lambda|\log(\lambda^{s+1})|)} m(x) \exp\left(-\frac{||x||_1}{\lambda}\right) d\mu_{\mathbb{R}^s}(x) \\
&= O\left(\lambda^{s+1} \int_{\mathbb{R}^s \setminus B_1(0;\lambda|\log(\lambda^{s+1})|)} m(x) d\mu_{\mathbb{R}^s}(x)\right) + \int_{B_1(0;\lambda|\log(\lambda^{s+1})|)} (m(0) + O(\epsilon)) \left[1 + O\left(\frac{1}{\lambda} \exp\left(-\frac{||x||_1}{\lambda}\right)\right)\right] d\mu_{\mathbb{R}^s}(x \\
&= O\left(\lambda^{s+1} \underbrace{\int_{\mathbb{R}^s} m(x) d\mu_{\mathbb{R}^s}(x)}_{=C}\right) + \int_{B_1(0;\lambda|\log(\lambda^{s+1})|)} [m(0) + O(\epsilon)][1 + O(\lambda^s)] d\mu_{\mathbb{R}^s}(x) \\
&= O(C\lambda^{s+1}) + (m(0) + O(\epsilon))(1 + +O(\lambda^s)) \underbrace{\frac{|2(s+1)\lambda \log \lambda|^s}{\Gamma(s+1)}}_{Vol(B_1(0;\lambda|\log(\lambda^{s+1})|)} \\
&\sim m(0) \frac{|2(s+1)\lambda \log \lambda|^s}{\Gamma(s+1)}
\end{aligned}
$$

at leading order for small λ.

Asymptotic behavior of $E_{\tilde{\Pi}}[g(\theta)]$

We know turn to the small λ behavior of $\tilde{E}[g(\theta)]$. Fix $\epsilon > 0$, choose $\lambda < \epsilon^{1/s}$ such that $x \in B_1(0, \lambda|\log(\lambda)^{s+1}|)$ implies

$$
\begin{aligned}
|m(x) - m(0)| &< \epsilon \\
|E[g(\theta)|\nu(\theta) = x] - E[g(\theta)|\nu(\theta) = 0| &< \epsilon
\end{aligned}
$$

Similar to the study of $m(x)$, separate the $\tilde{E}[g(\theta)]$ into integrals over the interior and exterior of $B_1(0, \lambda|\log(\lambda)^{s+1}|)$

$$
\begin{aligned}
E_{\tilde{\Pi}}[g(\theta)] &= \int_{\mathbb{R}^s \setminus B_1(0;\lambda|\log(\lambda^{s+1})|)} \frac{m(x)}{m_\lambda} \exp\left(-\frac{||x||_1}{\lambda}\right) E\big[g(\theta)|\nu(\theta) = x\big] d\mu_{\mathbb{R}^s}(x) \\
&\quad + \int_{B_1(0;\lambda|\log(\lambda^{s+1})|)} \frac{m(x)}{m_\lambda} \exp\left(-\frac{||x||_1}{\lambda}\right) E\big[g(\theta)|\nu(\theta) = x\big] d\mu_{\mathbb{R}^s}(x) \\
&= O\left(\frac{\lambda^{s+1}}{m_\lambda} \int_{\mathbb{R}^s} m(x) E[g(\theta)|\nu(\theta) = x] d\mu_{\mathbb{R}^s}(x)\right) + \int_{B_1} \frac{m(0) + O(\epsilon)}{m_\lambda} (1 + O(\lambda^s)] \Big(E[g(\theta)|\nu(\theta) = 0] + O(\epsilon)\Big) d\mu_{\mathbb{R}^s}(\\
&= O\left(CE|g| \frac{\lambda}{|\log \lambda|^s}\right) + \frac{1 + O(\epsilon)}{\frac{|2(s+1)\lambda \log \lambda|^s}{\Gamma(s+1)}} E[g(\theta)|\nu(\theta) = 0] \int_{B_1} (1 + O(\epsilon)) d\mu_{\mathbb{R}^s}(x) \\
&= O\left(CE|g| \frac{\lambda}{|\log \lambda|^s}\right) + E[g(\theta)|\nu(\theta) = 0] \frac{1 + O(\epsilon)}{1 - C_1 \epsilon}
\end{aligned}
$$

Therefore, λ may be chosen sufficiently small so that $CE|g|\lambda/|\log\lambda^s| < \epsilon$ which implies that

$$\left|E_{\tilde{\Pi}}[g] - E[g(\theta)|\nu(\theta) = 0]\right| < C\epsilon.$$

And we may conclude that

$$E_{\tilde{\Pi}}[g] \to E[g|\nu = 0] \text{ as } \lambda \to 0^+.$$