Constraint Relaxation for Bayesian Modeling with Parameter Constraints

Leo L Duan, Alexander L Young, Akihiko Nishimura, David B 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, the choices of priors and efficient sampling algorithms are severely limited. Motivated to addresses these problems, we propose to relax the parameter support into the neighborhood surrounding constrained space. This allows us to utilize the much larger family of prior distributions and off–the–shelf sampling algorithms available in general unconstrained space. The relaxed model can be directly used for statistical inference, or viewed as an approximate solution to the original constrained problem. We study the constrained and relaxed distributions under multiple settings, and theoretically quantify their differences. Popular state–of–art sampling algorithms such as leap-frog Hamiltonian Monte Carlo can be directly utilized with almost no customization. We illustrate this approach through multiple novel modeling examples involving equality and inequality constraints.

KEY WORDS: Constrained Bayes, Constraint Functions, Shrinkage on Manifold, Support Expansion, Ordered Simplex

1 Introduction

It is extremely common to have prior information available on parameter constraints 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 semi-definiteness of matrices and orthogonality. There is a very rich literature on optimization subject to parameter constraints. One common approach is to rely on Lagrange and Karush-Kuhn-Tucker multipliers (Boyd and Vandenberghe, 2004). However, simply producing a point estimate is often insufficient, as uncertainty quantification (UQ) is a key component of most statistical analysis. 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 re-derived 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. This strategy appears conceptually simple, however, except for a few limited cases, it is generally very difficult to find appropriate prior distribution in constrained space and to develop tractable posterior sampling algorithm.

To overcome these difficulties, 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 is known as 'coordinate system' in manifold embedding literature (Nash, 1954; Do Carmo, 2016). Examples include the polar coordinates for a unit sphere, or stick-breaking construction for Dirichlet distribution on probability simplex (Ishwaran and James, 2001). One can then directly assign prior on the less constrained parameters. Despite a few successes, in general, it is difficult to assign prior on the new parameterization while preserving property on the original constrained space. For example, uniformity on a compact constrained space often requires a non-uniform distribution based on the new parameterization, which can be intractable (See Diaconis et al. (2013) for details). Moreover, convenient reparameterzation does not always exist, especially when multiple constraints are involved.

As a result, the solutions are often limited to a few distributions for specific constraints. For example, for modeling of data on Stiefel manifolds, one routinely relies on von Mises-Fisher and matrix Bingham-von Mises-Fisher distributions (Khatri and Mardia, 1977; Hoff, 2009), which are both derived from constrained Gaussian. Often, the flexibility of these distributions is questionable in accommodating the modern complexity of data, such as possible sparsity, presence of outlier, multi-modality, etc.. Motivated to address this issue, non-parametric Bayes approach has been exploited to extend the flexibility, such as the work by Lin et al. (2016) on Stiefel manifold; although such cases are rare for other constraints.

Besides the modeling difficulty, posterior sampling has been another challenge under parameter constraint. Early work (Gelfand et al., 1992) 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 Dunson and Neelon (2003), extended to functional data with monotone unimodal constraints (Gunn and Dunson, 2005), and recently modified to nonparametric regression with monotonicity (Lin and Dunson, 2014) or manifold (Lin et al., 2016) constraints. Another direction utilizes a customized Hamiltonian Monte Carlo (HMC) with geometric flow for a few selected constraints (Byrne and Girolami, 2013). On the other hand, those approaches are arguably highly customized, which cannot directly exploit the rapidly developing automatic tools such as STAN (Carpenter et al., 2016).

The goal of this article is to dramatically simplify the constrained modeling problem, by *relaxing* the parameter into the neighborhood surrounding the constrained space. One could exploit the larger family of prior distributions and off–the–shelf sampling algorithms available in general unconstrained space. Due to the proximity to the constrained space, the relaxed model enjoy property similar to the constrained model, such as an improved convergence in latent factor model as shown in data application. Alternatively, one could view the collected posterior samples as approximation to ones under constrained model. We carefully quantify the difference in the theoretic study.

2 Constraint Relaxation Methodology

In conventional models under constraints, assume that θ is an $\mathcal{R}$-valued random variable with dimensionality $\dim(\mathcal{R}) = r < \infty$ and that θ is subject to some constraints restricting it to a subset $\mathcal{D} \subset \mathcal{R}$. In Bayesian setting, θ is a parameter where constraints arise from some given information.

As a motivating example, consider the case where θ has prior density $\pi_{\mathcal{D}}(\theta)$ with support on $\mathcal{D}$. The posterior density of θ given data Y and $\theta \in \mathcal{D}$ is,

$$\pi_{\mathcal{D}}(\theta \mid Y) \propto \mathcal{L}(\theta; Y)\pi_{\mathcal{D}}(\theta).$$

There are two primary problems motivating this work: the choice of $\pi_{\mathcal{D}}$ is very limited, which severely restricts the scope of modeling; posterior sampling often is difficult. To address these challenges, we consider an alternative model that *relaxes* the constraints by placing a high probability in the neighborhood of $\mathcal{D}$ but has support in $\mathcal{R}$. Suppose we used the density

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

where $\pi_{\mathcal{R}}(\theta)$ is proportional to $\pi_{\mathcal{D}}(\theta)$ without the constraint, i.e. $\pi_{\mathcal{D}}(\theta) \propto \pi_{\mathcal{R}}(\theta)\mathbb{1}_{\mathcal{D}}(\theta)$. We assume $\pi_{\mathcal{R}}(\theta)$ is proper, so that the posterior is proper as well. We use $\|v_{\mathcal{D}}(\theta)\|$ as a distance from θ to the constrained space (e.g. $\inf_{x\in\mathcal{D}} \|\theta - x\|_2$ with $\|.\|$ an appropriate metric).

Note that $\mathbb{1}_{\mathcal{D}}(\theta)$ is the pointwise limit of $\exp\big(-\lambda^{-1}\|v_{\mathcal{D}}(\theta)\|\big)$ (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. As shown in later sections, the immediate benefits are that one could use general unconstrained prior for $\pi_{\mathcal{R}}(\theta)$, and posterior sampling becomes much simpler. We refer to this strategy as Bayes constraint relaxation (**CORE**).

We investigate a number of questions about Bayes CORE: (i) For what types of distributions and constraints is CORE suitable? (ii) Is there a general approach for constructing the 'relaxed' distribution? (iii) How well do samples from the relaxed model approximate the original constrained distribution? (iv) How does the amount of relaxation 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 $[0,\infty)^d$, $\mathbb{R}^{n\times k}$) and $\pi_{\mathcal{R}}(\theta)$ is absolutely continuous with respect to Lebesgue measure on $\mathcal{R}$ hereby denoted as $\mu_{\mathcal{R}}$. We investigate two general types of constraints.

2.1 Constrained Space with Positive Measure

We first consider the simpler case where $\mathcal{D}$ has positive measure, i.e. $\int_{\mathcal{D}} \mathcal{L}(\theta; y)\pi(\theta)d\mu_{\mathcal{R}}(\theta) > 0$. Generally, inequality constraints (e.g. $a^T\theta < 0$, $\|\theta\|_2^2 < 1$) fall into this category.

The constrained posterior density is a result of simple renormalization:

$$\pi_{\mathcal{D}}(\theta \mid Y) = \frac{\mathcal{L}(\theta; Y)\pi_{\mathcal{R}}(\theta)\mathbb{1}_{\mathcal{D}}(\theta)}{\int_{\mathcal{D}} \mathcal{L}(\theta; Y)\pi_{\mathcal{R}}(\theta)d\mu_{\mathcal{R}}(\theta)} \propto \mathcal{L}(\theta; Y)\pi_{\mathcal{R}}(\theta)\mathbb{1}_{\mathcal{D}}(\theta),$$

which is defined with respect to $\mu_{\mathcal{R}}$.

We now consider a relaxed density

$$\tilde{\pi}_\lambda(\theta) = \frac{\mathcal{L}(\theta; Y)\pi_{\mathcal{R}}(\theta) \exp\big(-\lambda^{-1}\|v_{\mathcal{D}}(\theta)\|\big)}{\int_{\mathcal{R}} \mathcal{L}(\theta; Y)\pi_{\mathcal{R}}(\theta) \exp\big(-\lambda^{-1}\|v_{\mathcal{D}}(\theta)\|\big)d\mu_{\mathcal{R}}(\theta)} \propto \mathcal{L}(\theta; Y)\pi_{\mathcal{R}}(\theta) \exp\big(-\lambda^{-1}\|v_{\mathcal{D}}(\theta)\|\big) \tag{2}$$

which is also absolutely continuous with respect to $\mu_{\mathcal{R}}$. Here $\lambda > 0$ and $v_{\mathcal{D}}(\theta)$ is a scalar-valued function which measures the distance from θ to the constrained space $\mathcal{D}$, i.e. $v_{\mathcal{D}}(\theta) = 0 \quad \forall\theta \in \mathcal{D}$ and is positive

otherwise. Formally, as $\lambda \to 0^+$, $\exp(-v_{\mathcal{D}}(\theta)/\lambda) \to \mathbb{1}_{\mathcal{D}}(\theta)$ pointwise. If $\mathcal{D}$ is an open subset of $\mathcal{R}$, this limit may not hold on the boundary of $\mathcal{D}$, denoted $\partial\mathcal{D}$; however in general, $\mu_{\mathcal{R}}(\partial\mathcal{D}) = 0$.

There are many possible choices for $\|v_{\mathcal{D}}\|$ which may be selected for different reasons. Perhaps the most intuitive choice is to take

$$\|v_{\mathcal{D}}(\theta)\| = \inf_{x\in\mathcal{D}} \|\theta - x\|_k \qquad \text{or } \|v_{\mathcal{D}}(\theta)\| = \inf_{x\in\mathcal{D}} \sqrt{(x-\theta)^T A(x-\theta)}$$

where $\|\cdot\|_k$ denotes the distance using the k-norm, used for isotropic relaxation; A is positive definite for directional relaxation. Alternatively, one may choose distance for convenient computation, such as closed-form. For example, for inequality $\theta_1 + \theta_2 < 0$, one could use $\|v_{\mathcal{D}}(\theta)\| = \max(\theta_1 + \theta_2, 0)$. As long as $\|v_{\mathcal{D}}(\theta)\|$ is zero for $\theta \in \mathcal{D}$ and positive for $\theta \notin \mathcal{D}$, it follows that $\pi_{\mathcal{D}}$ is the pointwise limit of $\tilde{\pi}_\lambda$ for $\mu_{\mathcal{R}}$ a.e. θ in $\mathcal{R}$.

One utility of CORE is to obtain approximate estimate $\mathbb{E}[g(\theta) \mid \theta \in \mathcal{D}]$ via the relaxed posterior density, $\tilde{\pi}_\lambda$. For small λ, it is natural to anticipate

$$\int_{\mathcal{R}} g(\theta)\tilde{\pi}_\lambda(\theta)d\mu_{\mathcal{R}}(\theta) \approx \int_{\mathcal{D}} g(\theta)\pi_{\mathcal{D}}(\theta)d\mu_{\mathcal{R}}(\theta).$$

Rigorous details for its validity, a suitable class of functions for which it applies and rates of convergence in λ are contained in Section 3.1. For now, we illustrate a common case of modeling under linear inequality.

Example: Constrained Gaussian under Linear Inequality

Suppose that $\theta = (\theta_1, \theta_2) \sim \mathrm{No}_{\mathcal{D}}(\mu, \Sigma)$ is a truncated bivariate Gaussian, parameterized by $\mu \in \mathbb{R}^2$ and covariance matrix, $\Sigma = \sigma^2 I_2$, $\sigma > 0$, with $\mathcal{D} = \{(\theta_1, \theta_2) \mid \theta_1 + \theta_2 \leq 1, \theta_1 \geq 0, \theta_2 \geq 0\}$.

The constrained density of θ is

$$\pi_{\mathcal{D}}(\theta_1, \theta_2) = \frac{\exp\left(-\frac{\|\theta-\mu\|^2}{2\sigma^2}\right)\mathbb{1}_{\mathcal{D}}(\theta)}{\int_0^1 \int_0^{1-\theta_1} \exp\left(-\frac{\|\theta-\mu\|^2}{2\sigma^2}\right) d\theta_2 d\theta_1}$$

$$\propto \exp\left(-\frac{\|\theta-\mu\|^2}{2\sigma^2}\right)\mathbb{1}_{\{\theta_1+\theta_2\leq 1\}\cap\{\theta_1\in[0,\infty)\}\cap\{\theta_2\in[0,\infty)\}}.$$

When considering $\mathcal{R} = \mathbb{R}^2$, observe that $\pi_{\mathcal{R}}(\theta)$ is the bivariate Gaussian density with mean μ and Σ. We take a simple distance function $\|v_{\mathcal{D}}(\theta_1, \theta_2)\| = \max(\theta_1 + \theta_2 - 1, 0) + \max(-\theta_1, 0) + \max(-\theta_2, 0)$. Then,

$\|v_{\mathcal{D}}(\theta_1, \theta_2)\| = 0 \ \forall \theta \in \mathcal{D}$. Otherwise, $\|v_{\mathcal{D}}\|$ is positive. The relaxed density, $\tilde{\pi}_\lambda(\theta)$, is

$$\begin{aligned}\tilde{\pi}_\lambda(\theta) &= \frac{\exp\left(- \frac{\|\theta-\mu\|^2}{2\sigma^2} - \frac{1}{\lambda}\|v_{\mathcal{D}}(\theta)\| \right)}{\int_{\mathbb{R}^2} \exp\left(- \frac{\|\theta-\mu\|^2}{2\sigma^2} - \frac{1}{\lambda}\|v_{\mathcal{D}}(\theta)\| \right) d\theta_2 d\theta_1} \\ &\propto \exp\left(- \frac{\|\theta-\mu\|^2}{2\sigma^2} \right) \exp\left(- \frac{1}{\lambda}[\max(\theta_1+\theta_2-1, 0) + \max(-\theta_1, 0) + \max(-\theta_2, 0)] \right).\end{aligned} \tag{3}$$

Figure 1 depicts a few plots of the relaxed density as λ decreases. For $\lambda = 10^{-2}$, the relaxed density still places non-negligible probability outside of the constrained region, as fuzzy edges due to slowly decreasing density can be observed near the boundary of the triangle. For $\lambda = 10^{-4}$, the edges become sharp as the density rapidly drops to 0 outside of $\mathcal{D}$.

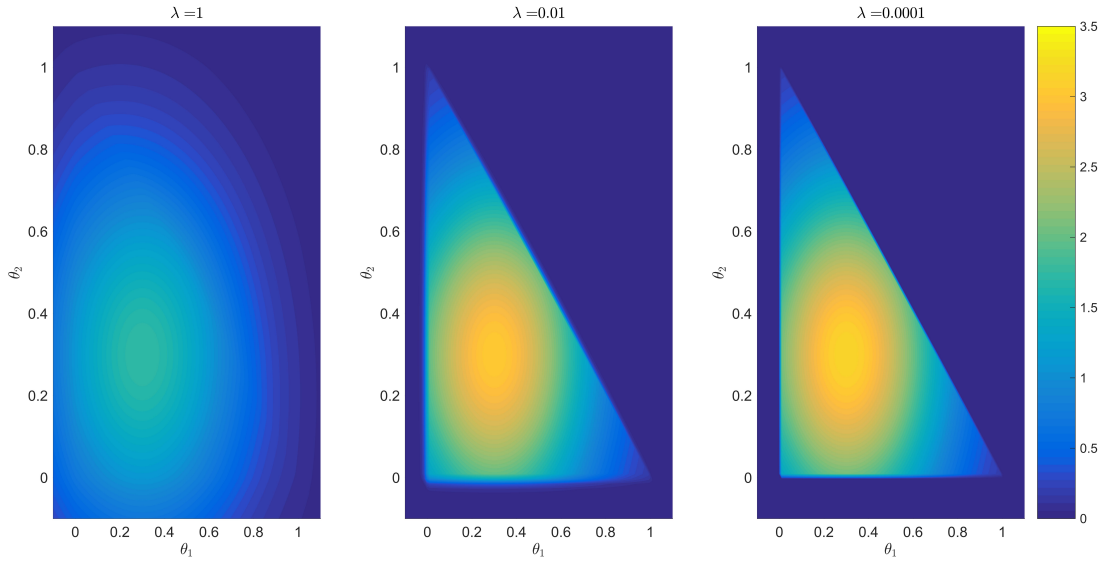


Figure 1: The relaxed distribution from a Gaussian $\text{No}([0.3, 0.3]', 1/10 I_2)$ subject to linear inequality, with density represented by color. As λ decreases toward 0, the amount of relaxation is reduced and the triangular constrained region becomes more apparent.

2.2 Constrained Space with Zero Measure

In the second case, we consider when $\mathcal{D}$ is a measure zero subset of $\mathcal{R}$, i.e. $\int_{\mathcal{D}} \mathcal{L}(\theta; Y)\pi_{\mathcal{R}}(\theta) d\mu_{\mathcal{R}}(\theta) = 0$. Recall that $\mathcal{R}$ have dimensionality r, we now restrict ourselves to the setting where $\mathcal{D}$ can be represented implicitly as the solution set of a consistent system of equations $\{v_j(\theta) = 0\}_{j=1}^s$, so that $\mathcal{D} = \{\theta \mid v_j(\theta) = 0, \ j = 1, \ldots, s\}$ is a $(r-s)$-dimensional submanifold of $\mathcal{R}$. While we impose some restrictions, the result applies on many common constraints (e.g. $\sum_i \theta_i = 1$, $\theta^T\theta = I$).

Due to the zero measure, one cannot obtain conditional probability as before by simply re-normalizing $\left[\int_{\mathcal{D}} \mathcal{L}(\theta; Y)\pi_{\mathcal{R}} d\mu_{\mathcal{R}}(\theta)\right]^{-1}$. Instead, we resort to the generalized definition of conditional probability, named *regular conditional probability* (r.c.p.) (Kolmogorov, 1950) to derive a constrained density coherent with $\pi_{\mathcal{R}}$.

More technical definition of the r.c.p. is provided in the appendix. For now, the following intuition is sufficient. While $\mathcal{D}$ has zero r-dimensional volume (i.e. zero Lebesgue measure), it has a positive $(r-s)$-dimensional 'surface area', formally known as normalized Hausdorff measure, denoted by $\bar{\mathcal{H}}^{(r-s)}$. We can use it as the normalizing constant to obtain a r.c.p density:

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

where $J(\nu_{\mathcal{D}}(\theta)) = \sqrt{(D\nu_{\mathcal{D}})'(D\nu_{\mathcal{D}})}$ is the Jacobian of $\nu_{\mathcal{D}}$, which we assume is positive. This term is introduced as we fix $\{v_j(\theta) = 0\}_{j=1}^{s}$. The density is defined with respect to $\bar{\mathcal{H}}^{(r-s)}$.

Similar to the constrained space with positive measure, we face the same modeling and computing challenges here as well. Additionally, although Hausdorff measure is a standard tool in geometric measure theory (Federer, 2014), the available distributions are scarce in Bayesian literature.

We now take a similar strategy by considering $\|\nu_{\mathcal{D}}(\theta)\|$ as the distance from θ to $\mathcal{D}$. Thus, $\|v_{\mathcal{D}}(\theta)\| = 0$ implies that $\theta \in \mathcal{D}$, otherwise $\|v_{\mathcal{D}}(\theta)\|_1 > 0$ implies $\theta \notin \mathcal{D}$. To construct a relaxed density, we expand support in the neighborhood of $\{\theta : \|v_{\mathcal{D}}(\theta)\| = 0\}$ by replacing the indicator with $\exp(-\lambda^{-1}\|v(\theta)\|)\mathbb{1}_{\mathcal{X}}(v(\theta))$. The truncation of the image of $v(.)$ to $\mathcal{X} \subset v(\mathcal{R})$ serves two purpose: (i) to make sure $\{\theta : v_{\mathcal{D}}(\theta) = x\}$ still has dimension $(r-s)$ for any $x \in \mathcal{X}$; (ii) to make sure that $v(.)$ is applicable in the following transformation (details to be discussed in next section).

To derive the density, we first consider a Lebesgue measure over a Borel set $\mathcal{F} \subset \mathcal{R}$:

$$\int_{\mathcal{F}} \tilde{\pi}_\lambda(\theta)d\mu(\theta) = \frac{\int_{\mathcal{X}} \left[\int_{\{\theta: v(\theta)=x\}\cap\mathcal{F}} \mathcal{L}(y;\theta)\pi_{\mathcal{R}}(\theta)J^{-1}(\nu_{\mathcal{D}}(\theta))d\bar{\mathcal{H}}^{(r-s)}(\theta)\right] \exp\left(-\lambda^{-1}\|x\|\right)dx}{\int_{\mathcal{X}} \left[\int_{\{\theta: v(\theta)=x\}} \mathcal{L}(y;\theta)\pi_{\mathcal{R}}(\theta)J^{-1}(\nu_{\mathcal{D}}(\theta))d\bar{\mathcal{H}}^{(r-s)}(\theta)\right] \exp\left(-\lambda^{-1}\|x\|\right)dx}$$

Using co-area formula (Federer, 2014), we can now transform double integrals to single integral. Omitting the integral over $\mathcal{F}$, this simplifies to a density:

$$\tilde{\pi}_\lambda(\theta) = \frac{\mathcal{L}(y;\theta)\pi_{\mathcal{R}}(\theta)\exp\left(-\frac{1}{\lambda}\|\nu_{\mathcal{D}}(\theta)\|\right)\mathbb{1}_{\mathcal{X}}(v(\theta))}{\int_{\mathcal{R}} \mathcal{L}(y;\theta)\pi_{\mathcal{R}}(\theta)\exp\left(-\frac{1}{\lambda}\|\nu_{\mathcal{D}}(\theta)\|\right)\mathbb{1}_{\mathcal{X}}(v(\theta))d\mu_{\mathcal{R}}(\theta)} \propto \mathcal{L}(y;\theta)\pi_{\mathcal{R}}(\theta)\exp\left(-\frac{1}{\lambda}\|\nu_{\mathcal{D}}(\theta)\|\right)\mathbb{1}_{\mathcal{X}}(v(\theta)), \tag{4}$$

which is defined with respect to $\mu_{\mathcal{R}}$. Note the Jacobian term vanishes and the density is with respect to common Lebesgue measure. The relaxed density is very similar to (2) in the last section.

Much like the positive measure case, $\exp\left(-\lambda^{-1}\|\nu_{\mathcal{D}}(\theta)\|\right)$ converges pointwise to $\mathbb{1}_{\mathcal{D}}(\theta)$. As $\lambda \to 0^+$, this multiplicative factor is concentrating the probability to a small layer around the constrained space. As a result, for small λ, one could expect that

$$\int_{\mathcal{R}} g(\theta)\tilde{\pi}_\lambda(\theta)d\mu_{\mathcal{R}}(\theta) \approx \int_{\mathcal{D}} g(\theta)\pi_{\mathcal{D}}(\theta)d\bar{\mathcal{H}}^{(r-s)}(\theta)$$

We provide details of this result and suitable class of functions in the next section. We first provide another example to illustrate.

Example: Constrained Gaussian on Unit Circle

Let $\theta = (\theta_1, \theta_2)$ be a bivariate Gaussian parameterized by mean $\mu \in \mathbb{R}^2$ and covariance matrix, $\Sigma = \sigma^2 I_2$, $\sigma > 0$, except it is constrained to the unit circle $\mathcal{D} = \{(\theta_1, \theta_2) \mid \theta_1^2 + \theta_2^2 = 1\}$. Since the unit circle is one-dimensional and $\theta = (\theta_1, \theta_2)$ is two-dimensional, we use a (2-1)=1-dimensional constraint function

$$v_{\mathcal{D}}(\theta_1, \theta_2) = \theta_1^2 + \theta_2^2 - 1.$$

Then $v_{\mathcal{D}}(\theta_1, \theta_2) = 0\ \forall \theta \in \mathcal{D}$. Otherwise, $v_{\mathcal{D}}$ is non-zero. Furthermore, $J(\nu_{\mathcal{D}}(\theta)) = 2\|\theta\|_2 = 2$ for $\theta \in \mathcal{D}$. The constrained density of θ given that $\theta \in \mathcal{D}$ is then,

$$\begin{aligned}
\pi_{\mathcal{D}}(\theta_1, \theta_2) = &= \frac{\exp\left(-\frac{\|\theta-\mu\|^2}{2\sigma^2}\right)\mathbb{1}_{\mathcal{D}}(\theta)}{\int_{\mathcal{D}} \exp\left(-\frac{\|\theta-\mu\|^2}{2\sigma^2}\right) d\bar{\mathcal{H}}^1(\theta)} \\
&\propto \exp\left(-\frac{\|\theta-\mu\|^2}{2\sigma^2}\right)\mathbb{1}_{\theta_1^2+\theta_2^2=1} \\
&\propto \exp\left(\frac{\theta'\mu}{\sigma^2}\right)\mathbb{1}_{\theta_1^2+\theta_2^2=1}.
\end{aligned}$$

This density can be interpreted with respect to the normalized Hausdorff-1 measure on the unit circle which coincides with arclength in this case. Observe this is the von Mises–Fisher distribution on the unit circle with location $\mu/\|\mu\|_2$ and concentration $\|\mu\|_2/\sigma^2$.

We make the relaxed space compact by $\mathcal{R} = (-a, a)^2$, with $a > 1$; and $\mathcal{X} = [-1, 0) \cup (0, 2a^2 - 1]$. Clearly, $\{\theta_1^2 + \theta_2^2 = x\}$ still has dimension 1 for all $x \in \mathcal{X}$.

The relaxed density $\tilde{\pi}_\lambda(\theta)$ is

$$\tilde{\pi}_\lambda(\theta) = \frac{\exp\left(-\frac{\|\theta-\mu\|^2}{2\sigma^2} - \frac{1}{\lambda}\|v_{\mathcal{D}}(\theta)\|\right)\mathbb{1}_{\mathcal{X}}(v_{\mathcal{D}}(\theta))}{\int_{\mathbb{R}^2}\exp\left(-\frac{\|\theta-\mu\|^2}{2\sigma^2} - \frac{1}{\lambda}\|v_{\mathcal{D}}(\theta)\|\right)\mathbb{1}_{\mathcal{X}}(v_{\mathcal{D}}(\theta))d\theta_2 d\theta_1} \tag{5}$$

$$\propto \exp\left(-\frac{\|\theta-\mu\|^2}{2\sigma^2}\right)\exp\left(-\frac{1}{\lambda}|\theta_1^2+\theta_2^2-1|\right)\mathbb{1}_{\mathcal{X}}(\theta_1^2+\theta_2^2-1).$$

Figure 2 depicts a few plots of the relaxed density as λ decreases. For $\lambda = 10^{-2}$ the constraint along the circle is clear. While the relaxed density still places some small probability outside of the constrained region, the rightmost plot becomes similar to the von Mises–Fisher distribution on the circle plotted in two dimensions.

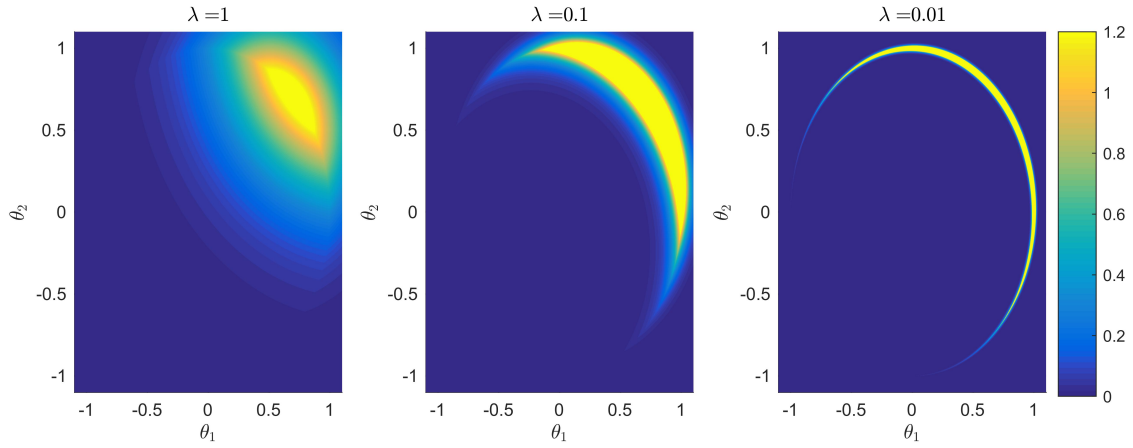


Figure 2: The relaxed distribution from a von Mises–Fisher distribution vMF($[1/\sqrt{2}, 1/\sqrt{2}]', 1/25$). With decreasing λ, the relaxation is reduced and the circular constraint becomes clear.

3 Theory

In this section, we provide theoretic details, mainly on two aspects: (i) the suitable constraints for relaxation; (ii) the error when using relaxed model as an approximation to constrained model.

3.1 Constrained Space with Positive Measure

For constrained space with positive measure, generally, as long as a tractable distance function exists such that $\mathcal{D} = \{\theta : \|v(\theta)\| = 0\}$, CORE is applicable.

We now focus on quantifying the difference between constrained and relaxed densities. Both of these densities are absolutely continuous with respect to Lebesgue measure on $\mathcal{R}$. Thus, the expectation of g with respect to constrained density is

$$\mathbb{E}[g(\theta)|\theta \in \mathcal{D}] = \int_{\mathcal{D}} g(\theta)\pi_{\mathcal{D}}(\theta)d\mu_{\mathcal{R}} = \frac{\int_{\mathcal{D}} g(\theta)\mathcal{L}(\theta;Y)\pi_{\mathcal{R}}(\theta)d\mu_{\mathcal{R}}(\theta)}{\int_{\mathcal{D}} \mathcal{L}(\theta;Y)\pi_{\mathcal{R}}(\theta)d\mu_{\mathcal{R}}(\theta)}. \tag{6}$$

Similarly, the expected value of g with respect to the relaxed density,

$$\mathbb{E}_{\tilde{\pi}_\lambda}[g(\theta)] = \int_{\mathcal{R}} g(\theta)\tilde{\pi}_\lambda(\theta)d\mu_{\mathcal{R}} = \frac{\int_{\mathcal{R}} g(\theta)\mathcal{L}(\theta;Y)\pi_{\mathcal{R}}(\theta)\exp(-\|v_{\mathcal{D}}(\theta)\|/\lambda)d\mu_{\mathcal{R}}(\theta)}{\int_{\mathcal{R}} \mathcal{L}(\theta;Y)\exp(-\|v_{\mathcal{D}}(\theta)\|/\lambda)\pi_{\mathcal{R}}(\theta)d\mu_{\mathcal{R}}(\theta)}. \tag{7}$$

We can now consider the behavior of $E_{\tilde{\pi}_\lambda}[g]$ as $\lambda \to 0^+$.

Lemma 1. *Suppose $g \in \mathbb{L}^1(\mathcal{R}, \mathcal{L}(\theta;Y)\pi_{\mathcal{R}}d\mu_{\mathcal{R}})$. Then,*

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

where $E|g(\theta)| \propto \int_{\mathcal{R}} |g(\theta)|\mathcal{L}(\theta;Y)\pi_{\mathcal{R}}d\mu_{\mathcal{R}}(\theta)$ is the expected value of $|g(\theta)|$ with respect to the unconstrained posterior density and $C_{\mathcal{R}} = \int_{\mathcal{R}} \mathcal{L}(\theta;Y)\pi_{\mathcal{R}}(\theta)d\mu_{\mathcal{R}}(\theta)$ is the normalizing constant of this unconstrained posterior density. Furthermore, if $\|v_{\mathcal{D}}(\theta)\|$ is zero for all $\theta \in \mathcal{D}$ and positive for $\theta \in (\mathcal{R}\setminus\mathcal{D})^o$, it follows from the dominated convergence theorem that

$$\left|\mathbb{E}[g(\theta)|\theta \in \mathcal{D}] - \mathbb{E}_{\tilde{\pi}_\lambda}[g(\theta)]\right| \to 0 \text{ as } \lambda \to 0^+.$$

Thus, one can obtain sufficiently accurate estimates of $\mathbb{E}[g|\theta \in \mathcal{D}]$ by sampling from $\tilde{\pi}_\lambda$ when λ is sufficiently small. From a practical standpoint, it is desirable to understand the rate at which $\mathbb{E}_{\tilde{\pi}_\lambda}[g(\theta)]$ converges to $\mathbb{E}[g(\theta) \in \mathcal{D}]$. This question is addressed in the following theorem.

Theorem 1. *Suppose $g \in \mathbb{L}^2(\mathcal{R}, \mathcal{L}(\theta;Y)\pi_{\mathcal{R}}d\mu_{\mathcal{R}})$, $v_{\mathcal{D}}(\theta) = \inf_{x\in\mathcal{D}} \|\theta - x\|_2$, $\mathcal{D}$ has a piecewise smooth boundary, and that $\mathcal{L}(\theta;Y)\pi_{\mathcal{R}}(\theta)$ is continuous on a open neighborhood containing $\mathcal{D}$. Then for $0 < \lambda \ll 1$,*

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

This theorem follows by applying the Cauchy-Schwartz inequality to the term in the numerator of the bound given in Lemma 1. One can attain a bound depending on the surface area of $\mathcal{D}$ when it is bounded. The proofs of Lemma 1 and Theorem 1 are contained in Appendix A.

These results have some important implications both analytically and numerically. First, in addition to point estimates, $\mathbb{E}[\theta \mid \theta \in \mathcal{D}]$, it is possible to approximate probabilities $P(\theta \in \mathcal{F}|\theta \in \mathcal{D})$ and higher moments, e.g. $\mathbb{E}[\Pi_j \theta_j^{k_j}|\theta \in \mathcal{D}]$, so long as these moments exist for the unconstrained posterior density. Second, these

bounds demonstrate that the error in using the relaxed density to approximate $\mathbb{E}[g(\theta)|\theta \in \mathcal{D}]$ is proportional to $\sqrt{\lambda}[\int_{\mathcal{D}} \mathcal{L}(\theta; Y)\pi_{\mathcal{R}}(\theta)d\mu_{\mathcal{R}}(\theta)]^{-2}$ although this rate may not be optimal. In practice, λ may need to be very small, particularly in the case where $0 < P(\theta \in \mathcal{D}) \ll 1$. Of course, specific details of the scaling of $\left|\mathbb{E}[g(\theta)|\theta \in \mathcal{D}] - \mathbb{E}_{\tilde{\pi}_\lambda}[g(\theta)]\right|$ will depend upon $\mathcal{D}$ and $\|v_{\mathcal{D}}(\theta)\|$.

3.2 Constrained Space with Zero Measure

For the constrained space with zero measure, we review a few important concepts of geometric measure theory which are used throughout this section. First, recall the definition of Hausdorff measure.

Definition - Hausdorff Measure. *Let $A \subset \mathbb{R}^r$. Fix $s \leq r$. Then*

$$\mathcal{H}^s(A) = \lim_{\delta \to 0} \inf \left\{ \sum [diam(S_i)]^s : A \subseteq \cup S_i, diam(S_i) \leq \delta, diam(S_i) = \sup_{x,y \in S} \|x - y\| \right\}$$

.

We denote the normalized Hausdorff measure as $\bar{\mathcal{H}}^s(A) = \frac{\Gamma(\frac{1}{2})^s}{2^s \Gamma(\frac{s}{2}+1)} \mathcal{H}^s(A)$. When $s = r$, Lebesgue and normalized Hausdorff measures coincide $\mu_{\mathbb{R}^m}(A) = \bar{\mathcal{H}}^s(A)$ (Evans and Gariepy, 2015). Additionally, for a subset $\mathcal{D}$, there exists a unique, critical value d such that

$$\bar{\mathcal{H}}^s(\mathcal{D}) = \begin{cases} 0, & s > d \\ \infty, & s < d. \end{cases}$$

The critical value, d, is referred to as the Hausdorff dimension of $\mathcal{D}$. We note that, when $\mathcal{D}$ is a compact, d-dimensional submanifold of $\mathbb{R}^m$, it will have Hausdorff dimension d and $\bar{\mathcal{H}}^d(\mathcal{D})$ is the d-dimensional surface area of A.

We now state the co-area formula which is used to define a regular conditional probability on the measure zero constrained space $\mathcal{D}$ and is pivotal in all of the proofs of the theorems.

Theorem 2. *Co-area formula (Diaconis et al., 2013; Federer, 2014) Suppose $\nu : \mathbb{R}^r \to \mathbb{R}^s$ with $s < r$ is Lipschitz and that $g \in \mathbb{L}^1(\mathbb{R}^r, \mu_{\mathbb{R}^r})$. Assume $J[\nu(\theta)] > 0$, then*

$$\int_{\mathbb{R}^r} g(\theta) J[\nu(\theta)] d\mu_{\mathbb{R}^r}(\theta) = \int_{\mathbb{R}^s} \left(\int_{\nu^{-1}(y)} g(\theta) d\bar{\mathcal{H}}^{r-s}(\theta) \right) d\mu_{\mathbb{R}^s}(y), \tag{8}$$

The behavior of the pre-images $\nu^{-1}(y)$ in the co-area formula are important for the convergence results presented later in this section. As such, we assume that $\mathcal{D}$ can be defined implicitly as the solution set to a system of s equations, $\{\nu_j(\theta) = 0\}_{j=1}^s$, where

(a) $\nu_j : \mathcal{R} \to \mathbb{R}$ is Lipschitz continuous,

(b) $v_j(\theta) = 0$ only for $\theta \in \mathcal{D}$,

(c) for $j = 1, \ldots, s$, the pre-image $v_j^{(-1)}(x)$ is a co-dimension 1 sub-manifold of $\mathcal{R}$ for $\mu_{\mathbb{R}}$-a.e. x in the range of v_j,

(d) $v_j^{(-1)}(0)$ and $v_k^{(-1)}(0)$ intersect transversally for $1 \leq j < k \leq s$.

We refer to the functions $v_1, \ldots, v_s$ as constraint functions. In this case, if we let $\nu : \mathcal{R} \to \mathbb{R}^s$ be the vector-valued function $\nu(\theta) = [\nu_1(\theta), \ldots, \nu_s(\theta)]^T$, then $\mathcal{D} = \ker(v)$ is a co-dimension s submanifold of $\mathcal{R}$ for $\mu_{\mathbb{R}^s}$-a.e. x the range of v. Recall, the ambient space, $\mathcal{R}$, is $r-$dimensional. Therefore, it follows that $\mathcal{D}$ is a $(r-s)-$dimensional submanifold of $\mathcal{R}$, and it is natural to discuss the $(r-s)$-dimensional surface area of $\mathcal{D}$.

Property (a), guarantees that ν is itself Lipschitz. The remaining properties (b)-(d) are constructed so that $\nu^{(-1)}(x)$ for $x \in \mathbb{R}^s$ is also a submanifold which is close to $\mathcal{D} = \nu^{(-1)}(0)$ when x is near zero. In particular, the assumption of transversality ensures that $\nu^{(-1)}(x)$ with also be $r - s$ dimensional for x sufficiently close to 0, defined by the set $\mathcal{X}$.

The existence and uniqueness of the constraints must be addressed. In the case where $\mathcal{D}$ is specified by a collection of equality constraints – such as the probability simplex or the Stiefel manifold for example– it is not difficult to find a suitable set of constraint functions. Table 1 contains a number of examples of common constrained spaces and appropriate choices of constraint functions.

$\mathcal{R}$	$\mathcal{D}$	$\dim(\mathcal{R})$	$\dim(\mathcal{D})$	Constraint functions
$[0,1]^r$	Probability simplex, Δ^{r-1}	r	$r-1$	$v(\theta) = \sum(\theta) - 1$
$\mathbb{R}^r$	Line, span$\{\vec{u}\}$ $\vec{u} \neq \vec{0}$	r	1	$\nu_j(\vec{\theta}) = \vec{\theta}^T \vec{b}_j$ $\{\vec{b}_1, \ldots, \vec{b}_{r-1}\}$ a basis for span$\{\vec{u}\}^\perp$
$[-1,1]^r$	Unit sphere, $\mathbb{S}^{r-1}$	r	$r-1$	$v(\theta) = (\|\theta\|^2 - 1)$
$[-1,1]^{n\times k}$	Stiefel manifold, $\mathcal{V}(n,k)$	nk	$nk - \frac{1}{2}k(k+1)$	$v_{i,j}(\theta) = (\vec{\theta}_i' \vec{\theta}_j - \delta_{i,j})$ $1 \leq i \leq j \leq k$ and $\delta_{i,j} = \mathbb{1}_{i=j}$

Table 1: Table of constraints for some commonly used constrained spaces.

With regards to uniqueness, we note that the constraints cannot be unique in any case. For example, rescaling in each $v_j(\theta)$ will also satisfy (a)-(d). Naturally, an optimal choice will depend largely on the properties of the constrained distribution that one wishes to estimate making the choice of $\{\nu_j\}_{j=1}^s$ context dependent.

Under the given construction of the constrained space, we can now specify the regular conditional probability of θ, given $\theta \in \mathcal{D}$.

Theorem 3. *(Diaconis et al., 2013) Assume that $J(v(\theta)) > 0$ and that for each $z \in \mathbb{R}^s$ there is a finite non-negative p_z such that,*

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

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

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

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

By construction, $\{\theta : v(\theta) = z\}$ is a $(r-s)$ dimensional submanifold of $\mathcal{R}$ for $\mu_{\mathbb{R}^s}$-a.e. z in $\mathcal{X}$, the suitable range of v. As such, it follows that one should take $p_z = r - s$. It is possible that $m^p(z) \in \{0, \infty\}$ for some $z \notin \mathcal{X}$; however, they are excluded during our construction. Most importantly, $0 \in \mathcal{X}$, therefore, Theorem 3 allows us to define

$$\pi_{\mathcal{D}}(\theta | \theta \in \mathcal{D}, Y) = \frac{1}{m^{r-s}(0)} \frac{\mathcal{L}(\theta; Y)\pi_{\mathcal{R}}(\theta)\mathbb{1}_{v(\theta)=0}}{J(v(\theta))} \tag{9}$$

as the constrained posterior density.

As a result, we can define the conditional expectation of $g(\theta)$ given $\theta \in \mathcal{D}$ as

$$\mathbb{E}[g(\theta)|\theta \in \mathcal{D}] = \mathbb{E}[g(\theta)|\nu(\theta) = 0\,] = \int_{\mathcal{R}} g(\theta)\pi_{\mathcal{D}}(\theta) d\bar{\mathcal{H}}^{r-s}(\theta).$$

The expected value of $g(\theta)$ with respect to the relaxed density, denote $\mathbb{E}_{\tilde{\Pi}}[g(\theta)]$, is

$$\mathbb{E}_{\tilde{\Pi}}[g(\theta)] = \frac{1}{m_\lambda} \int_{\mathcal{R}} g(\theta)\pi_{\mathcal{R}}(\theta)\mathcal{L}(Y; \theta) \exp\left(-\frac{1}{\lambda}\|\nu(\theta)\|_1 \right) d\mu_{\mathcal{R}}(\theta)$$

with $m_\lambda = \int_{\mathcal{R}} \pi_{\mathcal{R}}(\theta)\mathcal{L}(Y; \theta) \exp(-\lambda^{-1}\|\nu(\theta)\|_1) d\mu_{\mathcal{R}}(\theta)$. The primary results of the section are the following statements regarding the use of $\mathbb{E}_{\tilde{\Pi}}[g]$ to estimate $\mathbb{E}[g|\theta \in \mathcal{D}]$.

Theorem 4. *Let $m : \mathbb{R}^s \to \mathbb{R}$ and $G : \mathbb{R}^s \to \mathbb{R}$ be defined as follows*

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

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

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

$$\left| \mathbb{E}_{\tilde{\Pi}}[g] - \mathbb{E}[g|\theta \in \mathcal{D}] \right| \to 0 \text{ as } \lambda \to 0^+.$$

Corollary 1. *In addition to the assumptions of Theorem 4, suppose that both m and G are differentiable at 0. Then*

$$\left| \mathbb{E}_{\tilde{\Pi}}[g] - \mathbb{E}[g|\theta \in \mathcal{D}] \right| = O\left(\frac{\lambda}{|\log \lambda|^s} \right)$$

as $\lambda \to 0^+$.

Like the results from Section 3.1, the convergence rates are sub-linear. Unlike the positive measure case, the convergence rates are dimension dependent. We assess the approximation error with different λ in the von Mises–Fisher distribution as described above. The result is provided in the appendix.

4 Posterior Computation

Compared to constrained density in space $\mathcal{D}$, relaxed density is supported in $\mathcal{R}$ and can be directly sampled via off–the–shelf 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 Nishimura et al. (2017).

In order to sample θ, HMC introduces an auxiliary 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 θ. HMC then sample from the joint target density $\pi(\theta, p) = \pi(\theta)\pi(p) \propto \exp(-H(\theta, p))$ where, in the case of the posterior under relaxation,

$$\begin{aligned} &H(\theta, p) = U(\theta) + K(p), \\ &\text{where } U(\theta) = -\log \pi(\theta), \\ &\qquad K(p) = \frac{p' M^{-1} p}{2}. \end{aligned} \tag{10}$$

with $\pi(\theta)$ is the unnormalized density in (2) or (4).

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

$$\begin{aligned} \frac{\partial \theta^{(t)}}{\partial t} &= \frac{\partial H(\theta, p)}{\partial p} = M^{-1} p, \\ \frac{\partial p^{(t)}}{\partial t} &= -\frac{\partial H(\theta, p)}{\partial \theta} = -\frac{\partial U(\theta)}{\partial \theta}. \end{aligned} \tag{11}$$

The exact solution to (11) is typically intractable but a valid Metropolis proposal can be generated by numerically approximating (11) with a reversible and volume-preserving integrator (Neal, 2011). The standard choice is the *leapfrog* integrator which approximates the evolution $(\theta^{(t)}, p^{(t)}) \to (\theta^{(t+\epsilon)}, p^{(t+\epsilon)})$ through the following update equations:

$$p \leftarrow p - \frac{\epsilon}{2} \frac{\partial U}{\partial \theta}, \quad \theta \leftarrow \theta + \epsilon M^{-1} p, \quad p \leftarrow p - \frac{\epsilon}{2} \frac{\partial U}{\partial \theta} \tag{12}$$

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

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

We refer to this algorithm as CORE-HMC.

4.2 Computing Efficiency in CORE-HMC

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 of HMC. In this subsection, we provide some quantification.

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 $\epsilon \to 0$ (Betancourt, 2017). 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 (Beskos et al., 2013; Betancourt et al., 2014). A critical factor in determining a reasonable stepsize is the *stability limit* of the leapfrog integrator (Neal, 2011). 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 $a(\epsilon) = 1 - \mathcal{O}(\epsilon^4)$ (Beskos et al., 2013).

For simplicity, the following discussions assume the mass matrix M is taken to be the identity, and $\mathcal{D} = \cap_{j=1}^{s}\{\theta : v_j(\theta) = 0\}$. We denote $\mathcal{D}_j = \{\theta : v_j(\theta) = 0\}$ and consider a directional relaxation, which is equivalent to replacing a single λ with several λ_j's in the relaxation part, i.e. $\exp(-\sum_j \|v_j(\theta^*)\|\lambda_j^{-1})$. There are generally two factors limiting the efficiency of HMC: (i) the width of support in constrained space; (ii) the largest eigenvalue of the Hessian matrix. For the former, using $\mathcal{Q}$ to denote a support, the width of support is related to the shortest distance to the boundary $\eta(\theta; \mathcal{Q}) = \inf_{\theta' \notin \mathcal{Q}} \|\theta' - \theta\|$. If $\eta(\theta; \mathcal{Q}) \approx 0$ for all $\theta \in \mathcal{Q}$, the proposal would likely be rejected using a large leap-frog step size. In such case, it is useful to utilize CORE to expand support and increase $\eta(\theta; \mathcal{Q})$ for better computing efficiency. For the eigenvalue, let $\mathbf{H}_U(\theta)$ denote the hessian matrix of $U(\theta) = -\log \pi(\theta)$. 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 (Hairer et al., 2006), with $\xi_1(\theta)$ the largest eigenvalue of $\mathbf{H}_U(\theta)$. The hessian is

$$\mathbf{H}_U(\theta) = -\mathbf{H}_{\log(\mathcal{L}(\theta;y)\pi_{\mathcal{R}}(\theta))}(\theta) + \sum_j \lambda_j^{-1}\mathbf{H}\|v_j(\theta)\|\mathbb{1}_{\theta \notin \mathcal{D}_j}. \tag{13}$$

Note the second term is zero unless θ is outside of $\mathcal{D}_k$. As λ_j^{-1} in the second term often dominates the eigenvalue in the first term, hence the effective eigenvalue often is proportional to $\min_{j:\theta \notin \mathcal{D}_j} \lambda_j^{1/2}$.

Lastly, one may want to use CORE for approximate estimation under constrained model, while maintaining some computing efficiency. For this purpose, we now provide a practical guide on choosing λ_j. For $\mathcal{D}_j$'s with very small distance to support boundary $\eta(\theta; \mathcal{D}_j) \approx 0$, one should use moderate λ_j to increase support width; for $\mathcal{D}_j$'s without this issue, one should use very small $\lambda_j \approx 0$ to keep $\theta \in \mathcal{D}_j$, so that it has almost no influence on the hessian eigenvalue. The high rejection of HMC near the boundary of $\mathcal{D}_j$ can be avoided with random step size ϵ at each iteration, which is recommended for HMC in general (Livingstone et al., 2016). We evaluate the computing efficiency with different λ in sampling the von Mises–Fisher distribution. The result is provided in the appendix.

5 Simulated Examples

The simple computation of CORE frees up the modeling flexibility. We now illustrate more utility of the method via simulated examples.

Example: Sphere t Distribution

We now derive a new distribution on a $(p-1)$-sphere $\mathcal{D} = \{\theta \in \mathbb{R}^p : \|\theta\|_2 = 1\}$. Recall that von Mises–Fisher distribution (Khatri and Mardia, 1977) is the result of constraining a multivariate Gaussian

$\theta \sim \text{No}(F, I\sigma^2)$ with $F \in \mathcal{D}$

$$\pi_{\mathcal{D}}(\theta) \propto \exp(-\frac{\|F-\theta\|^2}{2\sigma^2})\mathbb{1}_{\theta'\theta=1} \propto \exp(\frac{F'}{\sigma^2}\theta)\mathbb{1}_{\theta'\theta=1}.$$

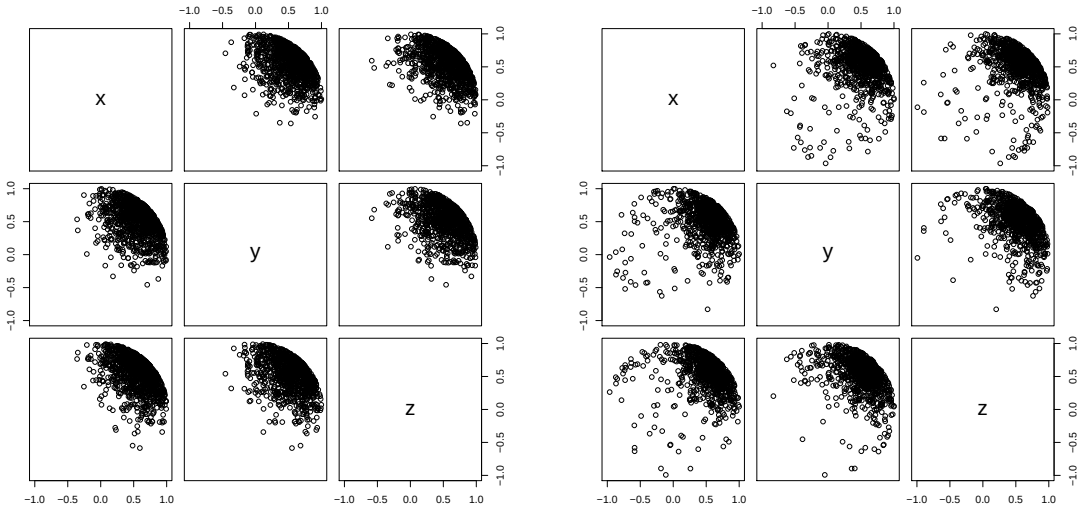
Although the final form appears more like an exponential, the behavior of von Mises–Fisher on sphere can be largely explained by its unconstrained parent Gaussian. In the Gaussian $\pi_{\mathcal{R}}(\theta)$, θ is symmetrically distributed around F, with density decaying exponentially as $\|\theta - F\|^2$ increases with rate $(2\sigma^2)^{-1}$; as the constrained density $\pi_{\mathcal{D}}(\theta)$ is proportional $\pi_{\mathcal{R}}(\theta)$, it concentrates similarly.

This naturally suggests we could use another distribution to induce different behavior on the sphere; then one could use CORE to generate approximate sample. We start from a multivariate t-distribution $\pi_{\mathcal{R}}(\theta)$, $t_m(F, I\sigma^2)$ with m degrees of freedom, mean $F \in \mathcal{D}$ and variance $I\sigma^2$, using (9) to generate a density

$$\begin{aligned}\pi_{\mathcal{D}}(\theta) \propto & (1 + \frac{\|F-\theta\|^2}{m\sigma^2})^{-\frac{(m+p)}{2}}\mathbb{1}_{\theta'\theta=1} \\ \propto & (1 - \frac{F'\theta}{1+m\sigma^2/2})^{-\frac{(m+p)}{2}}\mathbb{1}_{\theta'\theta=1}\end{aligned} \tag{14}$$

As in the t-distribution, the density decays polynomially as $\|F-\theta\|^2$ increases, as opposed to the exponential decay in Gaussian. We refer to this new distribution as sphere t-distribution.

CORE allows us to easily obtain approximate sample via relaxation function $\exp(-\lambda^{-1}\|\theta'\theta - 1\|)$. Figure 3 shows that the sphere t-distribution with $m = 3$ exhibits much less concentration than von Mises–Fisher on the sphere,. This can be useful for robust modeling when there could be 'outlier' on the sphere.



(a) von Mises–Fisher distribution.

(b) Sphere t-distribution with $m = 3$.

Figure 3: Sectional view of random samples from constrained distributions on a unit sphere inside $\mathbb{R}^3$. The distributions are derived through conditioning on $\theta'\theta = 1$ based on unconstrained densities of (a) $\text{No}(F, \text{diag}\{0.1\})$, (b) $t_3(F, \text{diag}\{0.1\})$, where $F = [1/\sqrt{3}, 1/\sqrt{3}, 1/\sqrt{3}]'$. The samples are generated via CORE-HMC with $\lambda = 10^{-3}$.

Example: Ordered Dirichlet Distribution

We derive an ordered Dirichlet distribution. We build it upon the canonical Dirichlet distribution $\text{Dir}(\alpha)$ with $\pi_{\mathcal{D}}(\theta) \propto \prod_{j=1}^{J} \theta_j^{\alpha-1} \mathbb{1}_{\sum_{j=1}^{J} \theta_j = 1}$ and further impose order constraint, $1 > \theta_1 \geq \ldots \geq \theta_J > 0$, yielding

$$\pi_{\mathcal{D}}(\theta) \propto \prod_{j=1}^{J} \theta_j^{\alpha-1} \cdot \mathbb{1}_{\sum_{j=1}^{J} \theta_j = 1} \cdot \prod_{j=1}^{J-1} \mathbb{1}_{\theta_j \geq \theta_{j+1}} \,. \tag{15}$$

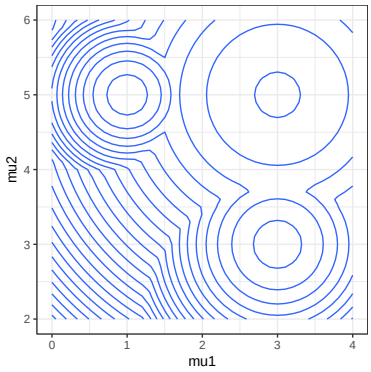
As commonly used in mixture model, canonical Dirichlet prior has its index j exchangeable. Since its permutation does not change the likelihood, label-switching problem often occurs (reviewed in Jasra et al. (2005)). Naturally, order constraint in θ can alleviate this problem, especially in preventing the switch between large θ_j and small $\theta_{j'}$.

To illustrate, we consider a hierarchical normal distribution with a common variance but the mean from a mixture, for data $y_i \in \mathbb{R}^2$ indexed by $i = 1, \ldots, n$:

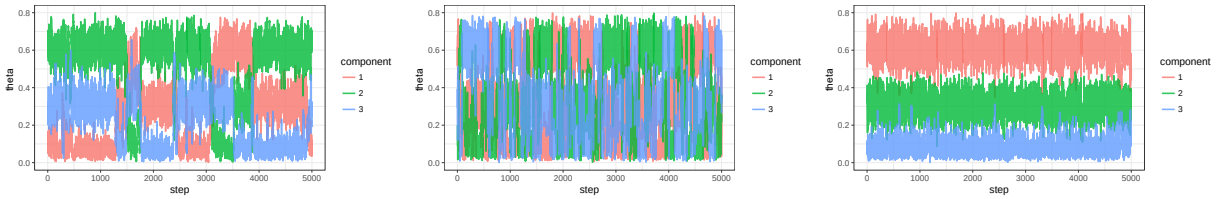
$$y_i \overset{indep}{\sim} \text{No}(\mu_i, \Sigma), \qquad \mu_i \overset{iid}{\sim} G, \qquad G(.) = \sum_{j=1}^{J} \theta_j \delta_{\mu_j}(.),$$

We generate $n = 100$ samples from 3 components with $\{\theta_1, \theta_2, \theta_3\} = \{0.6, 0.3, 0.1\}$, $\{\mu_1, \mu_2, \mu_3\} = \{[1, 5], [3, 3], [3, 5]\}$ and $\Sigma = I_2$. We assign weakly informative priors $\text{No}(0, 10I_2)$ for each μ_j and inverse–

Gamma prior for the diagonal element in $\Sigma = \text{diag}(\sigma_1^2, \sigma_2^2)$ with $\sigma_1^2, \sigma_2^2 \sim \text{IG}(2, 1)$. Figure 4(a) shows the contour of posterior density of μ. The small component sample size leads to large overlap among the posterior.



(a) Posterior density of the component means $\{\mu_j\}_{j=1}^3$.



(b) Gibbs sampling of unordered Dirichlet

(c) HMC sampling of unordered Dirichlet

(d) HMC sampling of ordered Dirichlet

Figure 4: Contour of the posterior density of component means and traceplot of the posterior sample for the component weights w, in a 3-component normal mixture model. Panel (a) shows that there is significant overlap among component means $\{\mu_j\}_{j=1}^3$. Without ordering in θ, its traceplot shows label-switching issue in both Gibbs (b) and HMC (c) sampling of Dirichlet distribution. The ordered Dirichlet distribution has significantly less label-switching issue (d), where we utilize CORE to obtain approximate posterior sample.

The ordering disrupts the traditional Gibbs sampling (Ishwaran and James, 2001), however, one could still obtain approximate posterior using CORE. We use $\exp(-\lambda_1^{-1} \| \sum_{j=1}^J \theta_j - 1) \|) \prod_{j=1}^{J-1} \exp[-\lambda_2^{-1} \max(\theta_{j+1} - \theta_j, 0)]$ to relax the constraints. We use $\lambda_1 = 10^{-3}$ on simplex constraint to allow efficient sampling and $\lambda_2 = 10^{-6}$ to induce almost no relaxation on the ordering. The posterior estimates of θ in CORE are close to the true values and indistinguishable from the other methods, except for a very small relaxation $\sum_{j=1}^J \theta_j - 1$ at posterior mean 0.001 ($(-0.001, 0.003)$ for 95% credible interval).

We compare the traceplots of ordered Dirichlet and unordered Dirichlet. Without the order constraint, significant label-switching occur in both Gibbs and HMC (traceplots in Figure 4(b,c)), whereas ordered Dirichlet has almost no label-switching(Figure 4(d)).

6 Application: Sparse Latent Factor Model in a Population of Brain Networks

We apply CORE in a real data application of analyzing a population of brain networks. 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. For each subject, we take the first scan out of the scan-rescan data as the input data, and reserve the second scan for model validation later. 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 (brain regions). 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 with the diagonal records empty $A_{(i,k,k)}$ for all i and k.

One interest in neuroscience is to quantify the variation of brain networks and identify the brain regions contributing to difference. Extending latent factor model to multiple matrices, one appealing approach is to have the networks share a common factor matrix but let the loadings vary across subjects.

$$A_{(i,k,l)} \sim \text{Bern}(\frac{1}{1+\exp(-\psi_{(i,k,l)} - z_{(k,l)})})$$

$$\psi_{(i,k,l)} = \sum_{r=1}^{d} v_{(i,r)} u_{(k,r)} u_{(l,r)}$$

$$z_{(k,l)} \sim \text{No}(0, \sigma_z^2), \quad \sigma_z^2 \sim \text{IG}(2, 1)$$

$$v_{(i,r)} \sim \text{No}(0, \sigma_{v,(r)}^2), \quad \sigma_{v,(r)}^2 \sim \text{IG}(2, 1)$$

for $k > l$, $k = 2, \ldots, V$, $i = 1, \ldots, n$; we choose weakly informative prior inverse Gamma IG$(2, 1)$, as appropriate for the scale parameters $\sigma_\cdot^2$ under the logistic link; $Z = \{z_{(k,l)}\}_{k=1,\ldots,V;l=1,\ldots,V}$ 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 $v_{(i,r)} > 0$; $U = \{u_{(k,r)}\}_{k=1,\ldots,V;r=1,\ldots,d}$ is the $V \times d$ shared factor matrix.

To help convergence, it is common to let the factor matrix U on a Stiefel manifold $\mathcal{V}(n, d) = \{U : U'U = I_d\}$, so that free rotation or rescaling of U is less likely to occur (Hoff et al., 2016). Obviously, slightly relaxing this constraint via CORE can still retain this convergence property, at this time, much more flexible priors can be assigned for U.

We now consider apply a shrinkage prior near the Stiefel manifold, in order to identify the important nodes. We use the Dirichlet-Laplace prior (Bhattacharya et al., 2015):

$$u_{(k,r)} = \eta_{(k,r)}\kappa_{(k,r)}\sigma_u$$

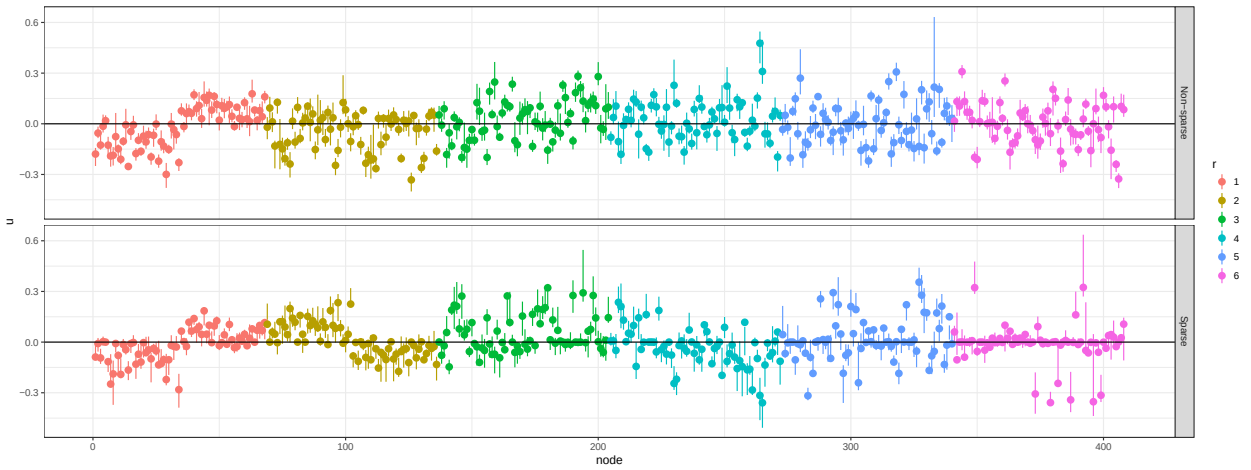
$$\eta_{(k,r)} \sim \text{Lap}(0,1), \quad \{\kappa_{(1,r)} \dots \kappa_{(V,r)}\} \sim \text{Dir}(\alpha), \quad \sigma_u^2 \sim \text{IG}(2,1)$$

for $k = 1, \dots, V$, $\text{Lap}(0,1)$ denotes the Laplace distribution centered at 0 with scale 1. To induce sparsity in each Dirichlet, we use $\alpha = 0.1$ as suggested by Bhattacharya et al. (2015). We replace the constraint by relaxation function $\exp(-\lambda^{-1}\|U'U - I\|)$, with $\lambda = 10^{-3}$.

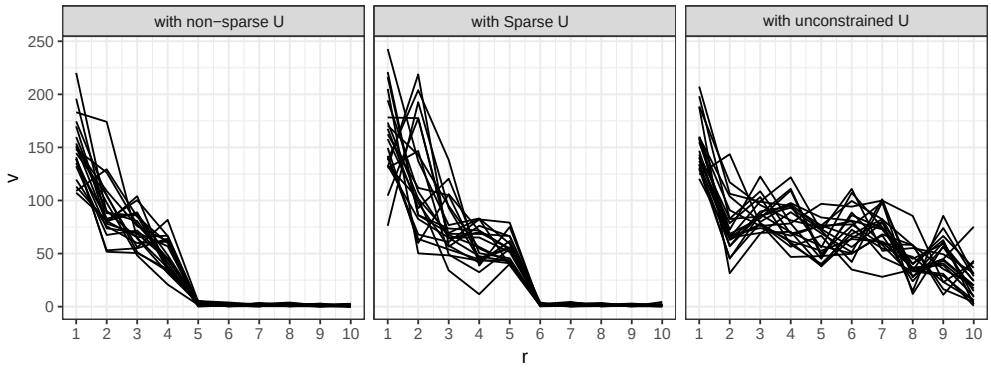
For comparison, we test with the specified model against two baseline models: one without shrinkage prior; one without the shrinkage prior and close-to orthonormality constraint. We use simple $u_{(k,r)} \sim \text{No}(0,1)$ in those two models. We run all models for $10,000$ iterations and discard the first $5,000$ iteration as burn-in. For each iteration, we run 300 leap-frog steps. For efficient computing, we truncated $d = 20$.

Figure 5(a) plots the top 6 factors U_r estimated near the Stiefel manifold, without and with shrinkage prior. Under the shrinkage prior, sparsity starts to show as early as the third factor U_3; the second factor U_2 shows a clear partition of first 34 and latter 34 nodes, which correspond to the two hemispheres of the brain. Accordingly, in the estimated loadings $v_{(i,r)}$ (Figure 5(b)), model under shrinkage prior detects more variability in the subject-specific loadings (represented by each line), especially over the second factor.

For the model with a completely unconstrained U, the factors and loadings fail to converge. And the loadings have a much slower drop to 0, compared to the two models with U near the Stiefel manifold. This indicates that near-orthogonal factors are more efficient representation of the span.



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



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

Figure 5: Factors and loadings estimates of the network models. Panel (a) shows that shrinkage model shows difference starting from the second factor (model with unconstrained U is omitted due to non-convergence in the factor); Panel (b) compares the varying loadings of the subjects in three models.

We further validate the models by assessing the area under the receiver operating characteristic curve (AUC). We compute the posterior mean of estimated connectivity probability for each individual, then evaluate AUC against the observed binary data (fitted AUC) and the unobserved binary data from the second scan of the same subjects (prediction AUC). Table 2 lists the benchmark results. The two models with near-orthonormality show much better performance, especially in prediction. Although we do not see a clear improved prediction by further using shrinkage prior, the sparse loadings it discover could be more useful for scientific interpretation. In terms of computing efficiency, CORE-HMC generates reasonable effective samples (ESS) per 1000 iterations; while the model with no constraint suffers from extremely small ESS due to non-convergence.

Model	(i).with shrinkage & near-orthonormality	(ii).with near-orthonormality only	(iii).completely 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 2: Benchmark of 3 models for 21 brain networks. Models with near-orthonormality show much better performance in both AUC and computing efficiency.

7 Discussion

Using constraint relaxation, we circumvent the common difficulties of constrained modeling, such as prior specification and posterior estimation. One interesting further direction perhaps is to tackle the 'doubly intractable' problem. This issue emerges when data (instead of parameters) are on the constrained space, forcing some associated parameters into an intractable constant. It is worth studying how to exploit CORE to approximate this normalization constant. Another task under the CORE framework may involve development of a formal test on whether the parameters reside on the constrained space.

References

Beskos, A., N. Pillai, G. Roberts, J. M. Sanz-Serna, and A. Stuart (2013, 11). Optimal Tuning of the Hybrid Monte Carlo algorithm. *Bernoulli 19*(5A), 1501–1534.

Betancourt, M. (2017). A Conceptual Introduction to Hamiltonian Monte Carlo. *arXiv:1701.02434*.

Betancourt, M., S. Byrne, and M. Girolami (2014). Optimizing the Integrator Step Size for Hamiltonian Monte Carlo. *arXiv:1411.6669*.

Bhattacharya, A., D. Pati, N. S. Pillai, and D. B. Dunson (2015). Dirichlet–Laplace Priors for Optimal Shrinkage. *Journal of the American Statistical Association 110*(512), 1479–1490.

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.

Carpenter, B., A. Gelman, M. Hoffman, D. Lee, B. Goodrich, M. Betancourt, M. A. Brubaker, J. Guo, P. Li, and A. Riddell (2016). STAN: a Probabilistic Programming Language. *Journal of Statistical Software 20*, 1–37.

Desikan, R. S., F. Ségonne, B. Fischl, B. T. Quinn, B. C. Dickerson, D. Blacker, R. L. Buckner, A. M. Dale, R. P. Maguire, B. T. Hyman, et al. (2006). An Automated Labeling System for Subdividing The Human Cerebral Cortex On MRI Scans Into Gyral Based Regions Of Interest. *Neuroimage 31*(3), 968–980.

Diaconis, P., S. Holmes, M. Shahshahani, et al. (2013). Sampling from a Manifold. In *Advances in Modern Statistical Theory and Applications: A Festschrift in honor of Morris L. Eaton*, pp. 102–125. Institute of Mathematical Statistics.

Do Carmo, M. P. (2016). *Differential Geometry of Curves and Surfaces: Revised and Updated Second Edition*. Courier Dover Publications.

Dunson, D. B. and B. Neelon (2003). Bayesian Inference on Order-Constrained Parameters in Generalized Linear Models. *Biometrics 59*(2), 286–295.

Evans, L. C. and R. F. Gariepy (2015). *Measure Theory and Fine Properties of Functions*. CRC press.

Federer, H. (2014). *Geometric Teasure Theory*. Springer.

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.

Gunn, L. H. and D. B. Dunson (2005). A Transformation Approach for Incorporating Monotone or Unimodal Constraints. *Biostatistics 6*(3), 434–449.

Hairer, E., C. Lubich, and G. Wanner (2006). *Geometric Numerical Integration. Structure-Preserving Algorithms for Ordinary Differential Equations*. Springer-Verlag.

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.

Ishwaran, H. and L. F. James (2001). Gibbs Sampling Methods for Stick-breaking Priors. *Journal of the American Statistical Association 96*(453), 161–173.

Jasra, A., C. C. Holmes, and D. A. Stephens (2005). Markov Chain Monte Carlo Methods and The Label Switching Problem in Bayesian Mixture Modeling. *Statistical Science*, 50–67.

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.

Kolmogorov, A. N. (1950). Foundations of the Theory of Probability.

Landman, B. A., A. J. Huang, A. Gifford, D. S. Vikram, I. A. L. Lim, J. A. Farrell, J. A. Bogovic, J. Hua, M. Chen, S. Jarso, et al. (2011). Multi-parametric Neuroimaging Reproducibility: A 3-T Resource Study. *Neuroimage 54*(4), 2854–2866.

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*.

Nash, J. (1954). C1 Isometric Imbeddings. *Annals of mathematics*, 383–396.

Neal, R. M. (2011). MCMC using Hamiltonian Dynamics. *Handbook of Markov Chain Monte Carlo 2*, 113–162.

Nishimura, A., D. Dunson, and J. Lu (2017). Discontinuous Hamiltonian Monte Carlo for Sampling Discrete Parameters. *arXiv preprint arXiv:1705.08510*.

A Proofs for Section 3.1

Proof. Proof of Lemma 1

Recall, that the distance function $\|v_{\mathcal{D}}(\theta)\|$ is chosen so that $\|v_{\mathcal{D}}(\theta)\|$ is zero for all $\theta \in \mathcal{D}$. It follows that for any function g

$$\begin{aligned}&\int_{\mathcal{R}} g(\theta)\mathcal{L}(\theta;Y)\pi_{\mathcal{R}}(\theta)\exp(-\|v_{\mathcal{D}}(\theta)\|/\lambda)d\mu_{\mathcal{R}}(\theta)\\ &= \int_{\mathcal{R}\setminus\mathcal{D}} g(\theta)\mathcal{L}(\theta;Y)\pi_{\mathcal{R}}(\theta)\exp(-\|v_{\mathcal{D}}(\theta)\|/\lambda)d\mu_{\mathcal{R}}(\theta) + \int_{\mathcal{D}} g(\theta)\mathcal{L}(\theta;Y)\pi_{\mathcal{R}}(\theta)\exp(-\|v_{\mathcal{D}}(\theta)\|/\lambda)d\mu_{\mathcal{R}}(\theta).\end{aligned} \tag{16}$$

Then,

$$\left|E[g(\theta)|\theta \in \mathcal{D}] - E_{\tilde{\pi}_\lambda}[g(\theta)]\right|$$

$$= \left| \frac{\int_{\mathcal{D}} g(\theta)\mathcal{L}(\theta;Y)\pi_{\mathcal{R}}(\theta)d\mu_{\mathcal{R}}(\theta)}{\int_{\mathcal{D}} \mathcal{L}(\theta;Y)\pi_{\mathcal{R}}(\theta)d\mu_{\mathcal{R}}(\theta)} - \frac{\int_{\mathcal{R}} g(\theta)\mathcal{L}(\theta;Y)\pi_{\mathcal{R}}(\theta)\exp\big(-\|v_{\mathcal{D}}(\theta)\|/\lambda)d\mu_{\mathcal{R}}(\theta)}{\int_{\mathcal{R}} \mathcal{L}(\theta;Y)\pi_{\mathcal{R}}(\theta)\exp\big(-\|v_{\mathcal{D}}(\theta)\|/\lambda)d\mu_{\mathcal{R}}(\theta)} \right|$$

$$= \left| \frac{\int_{\mathcal{R}\setminus\mathcal{D}} \mathcal{L}(\theta;Y)\pi_{\mathcal{R}}(\theta)\exp(-\|v_{\mathcal{D}}(\theta)\|/\lambda)d\mu_{\mathcal{R}}(\theta) \cdot \int_{\mathcal{D}} g(\theta)\mathcal{L}(\theta;Y)\pi_{\mathcal{R}}(\theta)d\mu_{\mathcal{R}}(\theta) - \int_{\mathcal{D}} \mathcal{L}(\theta;Y)\pi_{\mathcal{R}}(\theta)d\mu_{\mathcal{R}}(\theta) \cdot \int_{\mathcal{R}\setminus\mathcal{D}} g(\theta)\mathcal{L}(\theta;$$

where the second equality follows from combining the fractions and making use of (3). We can bound the denominator from below by $\big[\int_{\mathcal{D}} \mathcal{L}(\theta;Y)\pi_{\mathcal{R}}(\theta)d\mu_{\mathcal{R}}(\theta)\big]^2 > 0$ so that

$$\big|\mathbb{E}[g(\theta)|\theta \in \mathcal{D}] - \mathbb{E}_{\tilde{\pi}_\lambda}[g(\theta)]\big|$$

$$\leq \frac{\big|\int_{\mathcal{R}\setminus\mathcal{D}} \mathcal{L}(\theta;Y)\pi_{\mathcal{R}}(\theta)\exp(-\|v_{\mathcal{D}}(\theta)\|/\lambda)d\mu_{\mathcal{R}}(\theta) \cdot \int_{\mathcal{D}} g(\theta)\mathcal{L}(\theta;Y)\pi_{\mathcal{R}}(\theta)d\mu_{\mathcal{R}}(\theta) - \int_{\mathcal{D}} \mathcal{L}(\theta;Y)\pi_{\mathcal{R}}(\theta)d\mu_{\mathcal{R}}(\theta) \cdot \int_{\mathcal{R}\setminus\mathcal{D}} g(\theta)\mathcal{L}(\theta;Y)\pi_{\mathcal{R}}$$

where $C_{\mathcal{D}} = \int_{\mathcal{D}} \mathcal{L}(\theta;Y)\pi_{\mathcal{R}}(\theta)d\mu_{\mathcal{R}}(\theta)$. If we add and subtract

$$\int_{\mathcal{R}\setminus\mathcal{D}} \mathcal{L}(\theta;Y)\pi_{\mathcal{R}}(\theta)\exp(-\|v_{\mathcal{D}}(\theta)\|/\lambda)d\mu_{\mathcal{R}}(\theta) \cdot \int_{\mathcal{R}\setminus\mathcal{D}} g(\theta)\mathcal{L}(\theta;Y)\pi_{\mathcal{R}}(\theta)\exp(-\|v_{\mathcal{D}}(\theta)\|/\lambda)d\mu_{\mathcal{R}}(\theta)$$

within the numerator, we can apply the triangle inequality. Thus,

$$\big|\mathbb{E}[g(\theta)|\theta \in \mathcal{D}] - \mathbb{E}_{\tilde{\pi}_\lambda}[g(\theta)]\big|$$

$$\leq \frac{\left|\int_{\mathcal{R}\setminus\mathcal{D}} \mathcal{L}(\theta;Y)\pi_{\mathcal{R}}(\theta)\exp(-\|v_{\mathcal{D}}(\theta)\|/\lambda)d\mu_{\mathcal{R}}(\theta)\right| \cdot \left|\int_{\mathcal{D}} g(\theta)\mathcal{L}(\theta;Y)\pi_{\mathcal{R}}(\theta)d\mu_{\mathcal{R}}(\theta) - \int_{\mathcal{R}\setminus\mathcal{D}} g(\theta)\mathcal{L}(\theta;Y)\pi_{\mathcal{R}}(\theta)\exp(-\|v_{\mathcal{D}}(\theta)\right.}{C_{\mathcal{D}}^2}$$

$$+ \frac{\left|\int_{\mathcal{R}\setminus\mathcal{D}} g(\theta)\mathcal{L}(\theta;Y)\pi_{\mathcal{R}}(\theta)\exp(-\|v_{\mathcal{D}}(\theta)\|/\lambda)d\mu_{\mathcal{R}}(\theta)\right| \cdot \left|\int_{\mathcal{D}} \mathcal{L}(\theta;Y)\pi_{\mathcal{R}}(\theta)d\mu_{\mathcal{R}}(\theta) - \int_{\mathcal{R}\setminus\mathcal{D}} \mathcal{L}(\theta;Y)\pi_{\mathcal{R}}(\theta)\exp(-\|v_{\mathcal{D}}(\theta)\|\right.}{C_{\mathcal{D}}^2}$$

Since $g \in \mathbb{L}^1(\mathcal{R}, \mathcal{L}(\theta;Y)\pi_{\mathcal{R}}d\mu_{\mathcal{R}})$, we can then bound the numerators as follows. First,

$$\left|\int_{\mathcal{R}\setminus\mathcal{D}} \mathcal{L}(\theta;Y)\pi_{\mathcal{R}}(\theta)\exp(-\|v_{\mathcal{D}}(\theta)\|/\lambda)d\mu_{\mathcal{R}}(\theta)\right| \cdot \left|\int_{\mathcal{D}} g(y_i)\mathcal{L}(\theta;Y)\pi_{\mathcal{R}}(\theta)d\mu_{\mathcal{R}}(\theta) - \int_{\mathcal{R}\setminus\mathcal{D}} g(\theta)\mathcal{L}(\theta;Y)\pi_{\mathcal{R}}(\theta)\exp(-\|v_{\mathcal{D}}(\theta)\|/\right.$$

$$\leq \left|\int_{\mathcal{R}\setminus\mathcal{D}} \mathcal{L}(\theta;Y)\pi_{\mathcal{R}}(\theta)\exp(-\|v_{\mathcal{D}}(\theta)\|/\lambda)d\mu_{\mathcal{R}}(\theta)\right| \cdot \left(\left|\int_{\mathcal{D}} g(\theta)\mathcal{L}(\theta;Y)\pi_{\mathcal{R}}(\theta)d\mu_{\mathcal{R}}(\theta)\right| + \left|\int_{\mathcal{R}\setminus\mathcal{D}} g(\theta)\mathcal{L}(\theta;Y)\pi_{\mathcal{R}}(\theta)\exp(-\|v_{\mathcal{I}}\right.\right.$$

$$\leq \int_{\mathcal{R}\setminus\mathcal{D}} \mathcal{L}(\theta;Y)\pi_{\mathcal{R}}(\theta)\exp(-\|v_{\mathcal{D}}(\theta)\|/\lambda)d\mu_{\mathcal{R}}(\theta) \cdot \left(\int_{\mathcal{D}} |g(y_i)|\mathcal{L}(\theta;Y)\pi_{\mathcal{R}}(\theta)d\mu_{\mathcal{R}}(\theta) + \int_{\mathcal{R}\setminus\mathcal{D}} |g(\theta)|\mathcal{L}(\theta;Y)\pi_{\mathcal{R}}(\theta)\exp(-\|v_{\mathcal{D}}(\right.\right.$$

$$\leq \int_{\mathcal{R}\setminus\mathcal{D}} \mathcal{L}(\theta;Y)\pi_{\mathcal{R}}(\theta)\exp(-\|v_{\mathcal{D}}(\theta)\|/\lambda)d\mu_{\mathcal{R}}(\theta) \cdot \int_{\mathcal{R}} |g(x_i)|\mathcal{L}(\theta;Y)\pi_{\mathcal{R}}(\theta)d\mu_{\mathcal{R}}(\theta) = C_{\mathcal{R}}\mathbb{E}|g(x_i)| \int_{\mathcal{R}\setminus\mathcal{D}} \mathcal{L}(\theta;Y)\pi_{\mathcal{R}}(\theta)\exp(-$$

Here, $C_{\mathcal{R}} = \int_{\mathcal{R}} \mathcal{L}(\theta; Y)\pi_{\mathcal{R}}(\theta)d\mu_{\mathcal{R}}(\theta)$ is the normalizing constant of $\mathcal{L}(\theta; Y)\pi_{\mathcal{R}}$. Secondly,

$$\begin{aligned}
&\left| \int_{\mathcal{R}\setminus\mathcal{D}} g(\theta)\mathcal{L}(\theta; Y)\pi_{\mathcal{R}}(\theta)\exp(-\|v_{\mathcal{D}}(\theta)\|/\lambda)d\mu_{\mathcal{R}}(\theta) \right| \cdot \left| \int_{\mathcal{D}} \mathcal{L}(\theta; Y)\pi_{\mathcal{R}}(\theta)d\mu_{\mathcal{R}}(\theta) - \int_{\mathcal{R}\setminus\mathcal{D}} \mathcal{L}(\theta; Y)\pi_{\mathcal{R}}(\theta)\exp(-\|v_{\mathcal{D}}(\theta)\|/\lambda)d\mu \right| \\
&\leq \int_{\mathcal{R}\setminus\mathcal{D}} |g(\theta)|\mathcal{L}(\theta; Y)\pi_{\mathcal{R}}(\theta)\exp(-\|v_{\mathcal{D}}(\theta)\|/\lambda)dx \cdot \left| \int_{\mathcal{D}} \mathcal{L}(\theta; Y)\pi_{\mathcal{R}}(\theta)d\mu_{\mathcal{R}}(\theta) \right| + \left| \int_{\mathcal{R}\setminus\mathcal{D}} \mathcal{L}(\theta; Y)\pi_{\mathcal{R}}(\theta)\exp(-\|v_{\mathcal{D}}(\theta)\|/\lambda)d\mu \right| \\
&\leq \int_{\mathcal{R}\setminus\mathcal{D}} |g(\theta)|\mathcal{L}(\theta; Y)\pi_{\mathcal{R}}(\theta)\exp(-\|v_{\mathcal{D}}(\theta)\|/\lambda)d\mu_{\mathcal{R}}(\theta) \left(\int_{\mathcal{D}} \mathcal{L}(\theta; Y)\pi_{\mathcal{R}}(\theta)d\mu_{\mathcal{R}}(\theta) + \int_{\mathcal{R}\setminus\mathcal{D}} \mathcal{L}(\theta; Y)\pi_{\mathcal{R}}(\theta)d\mu_{\mathcal{R}}(\theta) \right) \\
&= \int_{\mathcal{R}\setminus\mathcal{D}} |g(\theta)|\mathcal{L}(\theta; Y)\pi_{\mathcal{R}}(\theta)\exp(-\|v_{\mathcal{D}}(\theta)\|/\lambda)d\mu_{\mathcal{R}}(\theta).
\end{aligned}$$

Thus, we have the bounds specified by the theorem,

$$\begin{aligned}
&\left|\mathbb{E}[g(\theta)|\theta \in \mathcal{D}] - \mathbb{E}_{\tilde{\pi}_\lambda}[g(\theta)]\right| \\
&\leq \frac{C_{\mathcal{R}}\mathbb{E}|g(\theta)| \int_{\mathcal{R}\setminus\mathcal{D}} \mathcal{L}(\theta; Y)\pi_{\mathcal{R}}(\theta)\exp(-\|v_{\mathcal{D}}(\theta)\|/\lambda)d\mu_{\mathcal{R}}(\theta)}{C_{\mathcal{D}}^2} + \frac{\int_{\mathcal{R}\setminus\mathcal{D}} |g(\theta)|\mathcal{L}(\theta; Y)\pi_{\mathcal{R}}(\theta)\exp(-\|v_{\mathcal{D}}(\theta)\|/\lambda)d\mu_{\mathcal{R}}(\theta)}{C_{\mathcal{D}}^2} \\
&= \frac{\int_{\mathcal{R}\setminus\mathcal{D}} (C_{\mathcal{R}}\mathbb{E}|g(\theta)| + |g(\theta)|)\mathcal{L}(\theta; Y)\pi_{\mathcal{R}}(\theta)\exp(-\|v_{\mathcal{D}}(\theta)\|/\lambda)d\mu_{\mathcal{R}}(\theta)}{C_{\mathcal{D}}^2}.
\end{aligned}$$

It remains to be shown that

$$\left|\mathbb{E}[g(\theta)|\theta \in \mathcal{D}] - \mathbb{E}_{\tilde{\pi}_\lambda}[g(\theta)]\right| \to 0 \text{ as } \lambda \to 0^+.$$

Again, by the assumptions that $g \in \mathbb{L}^1(\mathcal{R}, \mathcal{L}(\theta; Y)\pi_{\mathcal{R}}d\mu_{\mathcal{R}})$ and $\|v_{\mathcal{D}}(\theta)\| > 0$ for $\mu_{\mathcal{R}}$ a.e. $\theta \in \mathcal{R} \setminus \mathcal{D}$, it follows that $(C_{\mathcal{R}}\mathbb{E}|g(x_i)|+|g(x_i)|)\mathcal{L}(\theta; Y)\pi_{\mathcal{R}}(\theta)$ is a dominating function of $(C_{\mathcal{R}}\mathbb{E}|g(x_i)|+|g(x_i)|)\mathcal{L}(\theta; Y)\pi_{\mathcal{R}}(\theta)\exp(-\|v_{\mathcal{D}}(\theta)\|/\lambda)$ which converges to zero for $\mu_{\mathcal{R}}$-a.e. $\theta \in \mathcal{R} \setminus \mathcal{D}$ as $\lambda \to 0^+$. Thus, by the dominated convergence theorem, $\left|\mathbb{E}[g(\theta)|\theta \in \mathcal{D}] - \mathbb{E}_{\tilde{\pi}_\lambda}[g(\theta)]\right| \to 0$ as $\lambda \to 0^+$.

□

Proof. Proof of Theorem 1

We begin with the bound from Lemma 1.

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

For the moment, let us focus on the numerator of the previous expression. By the Cauchy-Schwartz inequality,

$$\int_{\mathcal{R}\setminus\mathcal{D}}(C_{\mathcal{R}}\mathbb{E}|g(\theta)|+|g(\theta)|)\mathcal{L}(\theta;Y)\pi_{\mathcal{R}}(\theta)\exp(-\|v_{\mathcal{D}}(\theta)\|/\lambda)d\mu_{\mathcal{R}}(\theta)$$

$$\leq\Big(\int_{\mathcal{R}\setminus\mathcal{D}}(C_{\mathcal{R}}\mathbb{E}|g(\theta)|+|g(\theta)|)^2\mathcal{L}(\theta;Y)\pi_{\mathcal{R}}(\theta)d\mu_{\mathcal{R}}(\theta)\Big)^{1/2}\Big(\int_{\mathcal{R}\setminus\mathcal{D}}\exp(-2\|v_{\mathcal{D}}(\theta)\|/\lambda)\mathcal{L}(\theta;Y)\pi_{\mathcal{R}}(\theta)d\mu_{\mathcal{R}}(\theta)\Big)^{1/2}$$

$$\leq\Big(\int_{\mathcal{R}}(C_{\mathcal{R}}\mathbb{E}|g(\theta)|+|g(\theta)|)^2\mathcal{L}(\theta;Y)\pi_{\mathcal{R}}(\theta)d\mu_{\mathcal{R}}(\theta)\Big)^{1/2}\Big(\int_{\mathcal{R}\setminus\mathcal{D}}\exp(-2\|v_{\mathcal{D}}(\theta)\|/\lambda)\mathcal{L}(\theta;Y)\pi_{\mathcal{R}}(\theta)d\mu_{\mathcal{R}}(\theta)\Big)^{1/2}$$

By assumption, $g\in\mathbb{L}^2(\mathcal{R},\mathcal{L}(\theta;Y)\pi_{\mathcal{R}}\mu_{\mathcal{R}})$. Thus,

$$\int_{\mathcal{R}\setminus\mathcal{D}}(C_{\mathcal{R}}\mathbb{E}|g(\theta)|+|g(\theta)|)\mathcal{L}(\theta;Y)\pi_{\mathcal{R}}(\theta)\exp(-\|v_{\mathcal{D}}(\theta)\|/\lambda)d\mu_{\mathcal{R}}(\theta)$$

$$=\underbrace{\Big([C_{\mathcal{R}}^3+2C_{\mathcal{R}}^2](\mathbb{E}|g|)^2+C_{\mathcal{R}}\mathbb{E}[|g|^2]\Big)^{1/2}}_{C_{\mathcal{R},g}<\infty}\Big(\int_{\mathcal{R}\setminus\mathcal{D}}\exp(-2\|v_{\mathcal{D}}(\theta)\|/\lambda)\mathcal{L}(\theta;Y)\pi_{\mathcal{R}}(\theta)d\mu_{\mathcal{R}}(\theta)\Big)^{1/2}$$

$$=C_{\mathcal{R},g}\Big(\int_{\mathcal{R}\setminus\mathcal{D}}\exp(-2\|v_{\mathcal{D}}(\theta)\|/\lambda)\mathcal{L}(\theta;Y)\pi_{\mathcal{R}}(\theta)d\mu_{\mathcal{R}}(\theta)\Big)^{1/2}$$

We separate the integral

$$\int_{\mathcal{R}\setminus\mathcal{D}}\exp(-2\|v_{\mathcal{D}}(\theta)\|/\lambda)\mathcal{L}(\theta;Y)\pi_{\mathcal{R}}(\theta)d\mu_{\mathcal{R}}(\theta)$$

over the sets $\{\theta:\|v_{\mathcal{D}}(\theta)\|>-\lambda\log\lambda\}$ and $\{\theta:\,0<\|v_{\mathcal{D}}(\theta)\|<-\lambda\log\lambda\}$.

$$\int_{\mathcal{R}\setminus\mathcal{D}}\exp(-2\|v_{\mathcal{D}}(\theta)\|/\lambda)\mathcal{L}(\theta;Y)\pi_{\mathcal{R}}(\theta)d\mu_{\mathcal{R}}(\theta)$$

$$=\int_{\{\theta:\|v_{\mathcal{D}}(\theta)\|>-\lambda\log\lambda\}}\exp(-2\|v_{\mathcal{D}}(\theta)\|/\lambda)\mathcal{L}(\theta;Y)\pi_{\mathcal{R}}(\theta)d\mu_{\mathcal{R}}(\theta)+\int_{\{\theta:\,0<\|v_{\mathcal{D}}(\theta)\|<-\lambda\log\lambda\}}\exp(-2\|v_{\mathcal{D}}(\theta)\|/\lambda)\mathcal{L}(\theta;Y)\pi_{\mathcal{R}}(\theta)$$

$$\leq\lambda^2\int_{\{\theta:\|v_{\mathcal{D}}(\theta)\|>-\lambda\log\lambda\}}\mathcal{L}(\theta;Y)\pi_{\mathcal{R}}(\theta)d\mu_{\mathcal{R}}(\theta)+\int_{\{\theta:\,0<\|v_{\mathcal{D}}(\theta)\|<-\lambda\log\lambda\}}\exp(-2\|v_{\mathcal{D}}(\theta)\|/\lambda)\mathcal{L}(\theta;Y)\pi_{\mathcal{R}}(\theta)d\mu_{\mathcal{R}}(\theta)$$

$$\leq C_{\mathcal{R}}\lambda^2+\int_{\{\theta:\,0<\|v_{\mathcal{D}}(\theta)\|<-\lambda\log\lambda\}}\exp(-2\|v_{\mathcal{D}}(\theta)\|/\lambda)\mathcal{L}(\theta;Y)\pi_{\mathcal{R}}(\theta)d\mu_{\mathcal{R}}(\theta)$$

To review, to this point we have shown that

$$\big|\mathbb{E}[g(\theta)|\theta\in\mathcal{D}]-\mathbb{E}_{\tilde{\pi}_\lambda}[g(\theta)]\big|\leq\frac{C_{\mathcal{R},g}}{D_{\mathcal{D}}^2}\Big(C_{\mathcal{R}}\lambda^2+\int_{\{\theta:\,0<\|v_{\mathcal{D}}(\theta)\|<-\lambda\log\lambda\}}\exp(-2\|v_{\mathcal{D}}(\theta)\|/\lambda)\mathcal{L}(\theta;Y)\pi_{\mathcal{R}}(\theta)d\mu_{\mathcal{R}}(\theta)\Big)^{1/2}\tag{17}$$

From the requirements of Theorem 1, we now let $\|v_{\mathcal{D}}(\theta)\|=\inf_{x\in\mathcal{D}}||\theta-x||_2$ and assume that $\mathcal{D}$ has a piecewise smooth boundary. In this case, the set $\{\theta:\,0<\|v_{\mathcal{D}}(\theta)\|<-\lambda\log\lambda\}$ forms a 'shell' of thickness $-\lambda\log\lambda$ which encases $\mathcal{D}$.

For the moment, suppose that $\mathcal{D}$ is a bounded subset of $\mathcal{R}$. Furthermore, suppose we take λ sufficiently small so that $\mathcal{L}(\theta;Y)\pi_{\mathcal{R}}(\theta)$ is continuous on $V_\lambda = \{\theta : 0 < \|v_{\mathcal{D}}(\theta)\| < -\lambda \log \lambda\}$. Observe that

$$\int\limits_{\{\theta: 0<\|v_{\mathcal{D}}(\theta)\|<-\lambda\log\lambda\}} \exp(-2\|v_{\mathcal{D}}(\theta)\|/\lambda)\mathcal{L}(\theta;Y)\pi_{\mathcal{R}}(\theta)d\mu_{\mathcal{R}}(\theta) \leq \sup_{V_\lambda} |\mathcal{L}(\theta;Y)\pi_{\mathcal{R}}(\theta)| \int_{V\lambda} \exp(-2\|v_{\mathcal{D}}(\theta)\|/\lambda)d\mu_R(\theta)$$

$$\leq \sup_{V_\lambda} |\mathcal{L}(\theta;Y)\pi_{\mathcal{R}}(\theta)| \int_{V\lambda} d\mu_R(\theta) = \sup_{V_\lambda} |\mathcal{L}(\theta;Y)\pi_{\mathcal{R}}(\theta)| \cdot Vol(V_\lambda)$$

$$= \sup_{V_\lambda} |\mathcal{L}(\theta;Y)\pi_{\mathcal{R}}(\theta)| S_{\mathcal{D}} \cdot \lambda|\log\lambda|$$

Here,$S_{\mathcal{D}}$ is the surface area of boundary of $\mathcal{D}$, which is finite by the assumptions that $\mathcal{D}$ is bounded and has a piecewise smooth boundary. Additionally, since V_λ is relatively compact, it follows that $\sup_{V_\lambda} |\mathcal{L}(\theta;Y)\pi_{\mathcal{R}}(\theta)| < \infty$.

Consider the more general case where $\mathcal{D}$ is not a bounded subset of $\mathcal{R}$. Since $\int_{\mathcal{R}} \mathcal{L}(\theta;Y)\pi_{\mathcal{R}}(\theta)d\mu_{\mathcal{R}}(\theta)$, there exists a radius ρ such that $\int_{||\theta||_2>\rho} \mathcal{L}(\theta;Y)\pi_{\mathcal{R}}(\theta)d\mu_{\mathcal{R}}(\theta) < \lambda^2$. Note that, for $\theta \in V_\lambda$, $J(v_{\mathcal{D}}(\theta)) = \sqrt{(Dv_{\mathcal{D}})'(Dv_{\mathcal{D}})} = 2$. By the co-area formula Diaconis et al. (2013); Federer (2014)

$$\int\limits_{\{\theta: 0<\|v_{\mathcal{D}}(\theta)\|<-\lambda\log\lambda\}} \exp(-2\|v_{\mathcal{D}}(\theta)\|/\lambda)\mathcal{L}(\theta;Y)\pi_{\mathcal{R}}(\theta)d\mu_{\mathcal{R}}(\theta) = \int_0^{-\lambda\log\lambda} e^{-\frac{x}{\lambda}} \left(\int_{v_{\mathcal{D}}^{-1}(x)} \frac{1}{2}\mathcal{L}(\theta;Y)\pi_{\mathcal{R}}(\theta)d\bar{\mathcal{H}}^{r-1}(\theta) \right) dx$$

Again, we may take λ sufficiently small so that $\mathcal{L}(\theta;Y)\pi_{\mathcal{R}}(\theta)$ is continuous on V_λ. As such, the function $\int_{v_{\mathcal{D}}^{-1}(x)} \frac{1}{2}\mathcal{L}(\theta;Y)\pi_{\mathcal{R}}(\theta)d\bar{\mathcal{H}}^{r-1}(\theta)$ is a continuous map from the closed interval, $[0, -\lambda\log\lambda]$, to $\mathbb{R}$. Hence it is bounded. As a result,

$$\int\limits_{\{\theta: 0<\|v_{\mathcal{D}}(\theta)\|<-\lambda\log\lambda\}} \exp(-2\|v_{\mathcal{D}}(\theta)\|/\lambda)\mathcal{L}(\theta;Y)\pi_{\mathcal{R}}(\theta)d\mu_{\mathcal{R}}(\theta)$$

$$\leq \sup_{x\in[0,-\lambda\log\lambda]} \left(\int_{v_{\mathcal{D}}^{-1}(x)} \frac{1}{2}\mathcal{L}(\theta;Y)\pi_{\mathcal{R}}(\theta)d\bar{\mathcal{H}}^{r-1}(\theta) \right) \cdot \int_0^{-\lambda\log\lambda} e^{-\frac{x}{\lambda}} dx$$

$$= \sup_{x\in[0,-\lambda\log\lambda]} \left(\int_{v_{\mathcal{D}}^{-1}(x)} \frac{1}{2}\mathcal{L}(\theta;Y)\pi_{\mathcal{R}}(\theta)d\bar{\mathcal{H}}^{r-1}(\theta) \right) \cdot (\lambda - \lambda^2) = O(\lambda)$$

This result also applies to the case where $\mathcal{D}$ is bounded. Thus, we may conclude that

$$\left|\mathbb{E}[g(\theta)|\theta\in\mathcal{D}] - \mathbb{E}_{\tilde{\pi}_\lambda}[g(\theta)]\right|$$

$$\leq \frac{C_{\mathcal{R},g}}{D_{\mathcal{D}}^2} \left(C_{\mathcal{R}}\lambda^2 + \sup_{x\in[0,-\lambda\log\lambda]} \left(\int_{v_{\mathcal{D}}^{-1}(x)} \frac{1}{2}\mathcal{L}(\theta;Y)\pi_{\mathcal{R}}(\theta)d\bar{\mathcal{H}}^{r-1}(\theta) \right) \cdot (\lambda - \lambda^2) \right)^{1/2}$$

$$= \frac{C_{\mathcal{R},g}}{D_{\mathcal{D}}^2} \cdot \sup_{x\in[0,-\lambda\log\lambda]} \left(\int_{v_{\mathcal{D}}^{-1}(x)} \frac{1}{2}\mathcal{L}(\theta;Y)\pi_{\mathcal{R}}(\theta)d\bar{\mathcal{H}}^{r-1}(\theta) \right) \cdot \sqrt{\lambda} + o(\sqrt{\lambda})$$

Since $\sup_{x\in[0,-\lambda\log\lambda]}\left(\int_{v_{\mathcal{D}}^{-1}(x)}\frac{1}{2}\mathcal{L}(\theta;Y)\pi_{\mathcal{R}}(\theta)d\bar{\mathcal{H}}^{r-1}(\theta)\right)$ is a decreasing function in λ, we may conclude that $\left|\mathbb{E}[g(\theta)|\theta\in\mathcal{D}]-\mathbb{E}_{\tilde{\pi}_\lambda}[g(\theta)]\right|=O(\sqrt{\lambda})$. □

B Proofs from Section 3.2

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

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

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

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

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 $\mathcal{R}$, denote by $\mu_{\mathcal{R}}$,

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

For a given function, $g:\mathcal{R}\to\mathbb{R}$, we can define the exact and approximate expectations of g, respectively $\mathbb{E}_\Pi$ and $\mathbb{E}_{\tilde{\Pi}}$, as

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

Let,

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

By construction, $m(x) > 0$ for $\mu_{\mathbb{R}^s}$-a.e. $x \in Range(v)$. In particular, $m_0 = m(0) > 0$. By Theorem 1,

$$\mathbb{E}[g(\theta)|v(\theta) = x] = \frac{1}{m(x)} \int_{v^{-1}(x)} g(\theta) \frac{\pi_{\mathcal{R}}(\theta)\mathcal{L}(y;\theta)}{J(v(\theta))} d\bar{\mathcal{H}}^{r-s}(\theta). \tag{18}$$

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

$$\mathbb{E}_{\tilde{\Pi}}[g(\theta)] = \int_{\mathbb{R}^s} \frac{m(x)}{m_\lambda} \exp\left(- \frac{||x||_1}{\lambda} \right) \mathbb{E}\big[g(\theta)|v(\theta) = x\big] d\mu_{\mathbb{R}^s}(x). \tag{19}$$

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

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

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

$$\begin{aligned}
m_\lambda &= \int_{\mathbb{R}^s \setminus B_1(0;\lambda|\log(\lambda^{s+1})|)} m(x) \exp\left(- \frac{||x||_1}{\lambda} \right) d\mu_{\mathbb{R}^s}(x) + \int_{B_1(0;\lambda|\log(\lambda^{s+1})|)} m(x) \exp\left(- \frac{||x||_1}{\lambda} \right) d\mu_{\mathbb{R}^s}(x) \\
&= O\bigg(\lambda^{s+1} \int_{\mathbb{R}^s \setminus B_1(0;\lambda|\log(\lambda^{s+1})|)} m(x) d\mu_{\mathbb{R}^s}(x) \bigg) \\
&\qquad + \int_{B_1(0;\lambda|\log(\lambda^{s+1})|)} m(x) \bigg[1 + O\bigg(\frac{1}{\lambda} \exp\left(- \frac{||x||_1}{\lambda} \right) \bigg) \bigg] d\mu_{\mathbb{R}^s}(x) \\
&= O\bigg(\lambda^{s+1} \int_{\mathbb{R}^s \setminus B_1(0;\lambda|\log(\lambda^{s+1})|)} m(x) d\mu_{\mathbb{R}^s}(x) \bigg) \\
&\qquad + \int_{B_1(0;\lambda|\log(\lambda^{s+1})|)} m(x) \Big[1 + O(\lambda^s) \Big] d\mu_{\mathbb{R}^s}(x)
\end{aligned}$$

Since $m(x)$ is continuous on an open neighborhood containing the origin, we may choose λ small enough so

that $m(x)$ is uniformly continuous on $B_1(0;\lambda|\log\lambda^{s+1}|)$. Then,

$$m_\lambda = O\Big(\lambda^{s+1}\Big) + \int_{B_1(0;\lambda|\log(\lambda^{s+1})|)} [m(0)+o(1)][1+O(\lambda^s)] d\mu_{\mathbb{R}^s}(x)$$

$$= O(\lambda^{s+1}) + [m(0)+o(1)][1+O(\lambda^s)] \underbrace{\frac{|2(s+1)\lambda\log\lambda|^s}{\Gamma(s+1)}}_{Vol(B_1(0;\lambda|\log(\lambda^{s+1})|)}$$

$$= m(0)\frac{|2(s+1)\lambda\log\lambda|^s}{\Gamma(s+1)} + o(|\lambda\log\lambda|^s)$$

at leading order as $\lambda \to 0^+$.

We know turn to the small λ behavior of $\tilde{\mathbb{E}}[g(\theta)]$. Again, we may choose λ sufficient small so that both

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

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

are continuous on $B_1(0;\lambda|\log\lambda^{s+1}|)$ and hence uniformly continuous at $x=0$.

Similar to the study of m_λ, separate the $\tilde{\mathbb{E}}[g(\theta)]$ into integrals over the interior and exterior of $B_1(0,\lambda|\log(\lambda)^{s+1}|)$. Again, we assume λ is taken to be sufficiently small so that both $m(x)$ and $G(x)$ are uniformly continuous on B_1. Then

$$\mathbb{E}_{\tilde{\Pi}}[g(\theta)] = \int_{\mathbb{R}^s} \frac{m(x)}{m_\lambda}\exp\Big(-\frac{||v(x)||_1}{\lambda}\Big)\mathbb{E}[g(\theta)|v(\theta)=x]d\mu_{\mathbb{R}^s}(x)$$

$$= \int_{\mathbb{R}^s} \frac{1}{m_\lambda}\exp\Big(-\frac{||v(x)||_1}{\lambda}\Big)\int_{v^{(-1)}(x)} \frac{\mathcal{L}(y;\theta)\pi_{\mathcal{R}}(\theta)}{J(v(\theta))} d\bar{\mathcal{H}}^{r-s}(\theta) d\mu_{\mathbb{R}^s}(x)$$

$$= \int_{\mathbb{R}^s\setminus B_1(0;\lambda|\log(\lambda^{s+1})|)} \frac{1}{m_\lambda}\exp\Big(-\frac{||x||_1}{\lambda}\Big)\int_{v^{(-1)}(x)} \frac{\mathcal{L}(y;\theta)\pi_{\mathcal{R}}(\theta)}{J(v(\theta))} d\bar{\mathcal{H}}^{r-s}(\theta) d\mu_{\mathbb{R}^s}(x)$$

$$+ \int_{B_1(0;\lambda|\log(\lambda^{s+1})|)} \frac{1}{m_\lambda}\exp\Big(-\frac{||x||_1}{\lambda}\Big)\int_{v^{(-1)}(x)} \frac{\mathcal{L}(y;\theta)\pi_{\mathcal{R}}(\theta)}{J(v(\theta))} d\bar{\mathcal{H}}^{r-s}(\theta) d\mu_{\mathbb{R}^s}(x)$$

$$= O\Big(\frac{\lambda^{s+1}}{m_\lambda}\Big) + \int_{B_1} \frac{m(0)+o(1)}{m_\lambda}(1+O(\lambda^s))\Big(\mathbb{E}[g(\theta)|v(\theta)=0] + o(1)\Big) d\mu_{\mathbb{R}^s}(x)$$

$$= O\Big(C\mathbb{E}|g|\frac{\lambda}{|\log\lambda|^s}\Big) + \mathbb{E}[g(\theta)|\theta\in\mathcal{D}] + o(1).$$

And we may conclude that

$$\Big|\mathbb{E}[g|\theta\in\mathcal{D} - \mathbb{E}_{\tilde{\Pi}}[g]\Big| \to O \text{ as } \lambda\to 0^+.$$

The proof of the corollary follows from changing the $o(1)$ correction within the integrals over $B_1(0;\lambda|\log\lambda^{s+1}|)$ with $O(\lambda|\log\lambda^{s+1}|)$ corrections. □

C Approximation Error of von–Mises Fisher distribution

We test $\lambda = 10^{-3}$, 10^{-4} and 10^{-5} for CORE-HMC. Table 3 shows the effective sample size per 1000 iterations, the effective 'violation' $|v(\theta)| = |\theta_1^2+\theta_2^2-1|$ and the $\left|\mathbb{E}_{\Pi}[\sum_j \theta_j]-\mathbb{E}_{\tilde{\Pi}}[\sum_j \theta_j]\right|$ as the approximation error. As the approximation error is numerically computed, to provide a baseline error, we also compare two independent samples from the same exact distribution. The approximation error 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.

	HMC based on CORE			Exact
	$\lambda = 10^{-3}$	$\lambda = 10^{-4}$	$\lambda = 10^{-5}$	
$\left\|\mathbb{E}_{\Pi}[\sum_j \theta_j] - \mathbb{E}_{\tilde{\Pi}}[\sum_j \theta_j]\right\|$	0.025 (0.014, 0.065)	0.016 (0.012, 0.019)	0.008 (0.006,0.015)	0.009 (0.007,0.015)
$\lvert v_{\mathcal{D}}(\theta)\rvert$	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
ESS /1000 Iterations	751.48	260.54	57.10	788.30

Table 3: Benchmark of constraint relaxation methods on sampling von–Mises Fisher distribution on a unit circle. For each CORE, average approximation error (with 95% credible interval, out of 10 repeated experiments) is computed, and numeric error is shown under column 'exact' as comparing two independent copies from the exact distribution. Effective sample size shows CORE with relatively large λ have high computing efficiency.