Bayesian constraint relaxation

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

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, *sharply* constrained priors are (a) unrealistic in many settings; and (b) tend to limit modeling scope to a narrow set of distributions that are tractable computationally. We propose to solve both of these problems via a general class of Bayesian *constraint relaxation* methods. The key idea is to replace the sharp indicator function of the constraint holding with an exponential kernel. This kernel decays with distance from the constrained space at a rate depending on a relaxation hyperparameter. By avoiding the sharp constraint, we enable use of off-the-shelf posterior sampling algorithms, such as Hamiltonian Monte Carlo, facilitating automatic computation in broad models. We study the constrained and relaxed distributions under multiple settings, and theoretically quantify their differences. We illustrate the method through multiple novel modeling examples.

KEY WORDS: Constrained Bayes, Constraint functions, Parameter restrictions; 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. A common approach relies on Lagrange and Karush-Kuhn-Tucker multipliers (**?**). However, simply producing a point estimate is often insufficient, as uncertainty quantification (UQ) is a key component of most statistical analyses. Usual large sample asymptotic theory, for example showing asymptotic normality of statistical estimators, tends to break down in constrained inference problems. Instead, limiting distributions may have a complex form that needs to be 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 (**?**). Although this strategy appears conceptually simple, there are two clear limitations in practice. First, it is often too restrictive to assume that the parameter exactly satisfies the constraint, and one may want to allow slight deviations. Second, it is in general very difficult to develop tractable posterior sampling algorithms except in special cases. For example, one may be forced to focus on particular forms for the prior and likelihood function to gain tractability and/or may need to develop specially tailored algorithms on a case-by-case basis.

To overcome the first problem, one can attempt to tune the parameters in an unconstrained prior to place high probability close to the constrained space. However, this approach is limited in scope and often cannot be used. An alternative strategy is to partially enforce the constraint. For example, in considering monotone function estimation, one may impose monotonicity only at a subset of points (REF). However, it is typically difficult to decide exactly which constraints to remove, and outside of specialized cases, this strategy is not appropriate. Hence, to our knowledge, there is essentially no solution to the general problem of how to choose a prior that is concentrated *close* to a particular constrained space.

There is a richer literature on the second problem - posterior sampling under constraints. A common strategy is to reparameterize to reduce the number of constraints and/or simplify the constraints. Examples include reparameterizations of positive semidefinite covariance matrices, probability vectors on the simplex, and spherical parameters. Unfortunately, convenient reparameterizations often are not available and/or may lead to challenges in prior elicitation. A simple and intuitive prior in the original parameterization may induce a complex prior in the new parameterization, motivating the use of *black-box* convenience priors that may conflict with prior knowledge. There are several alternatives focused on particular models for particular

constraints. These include both specialized distributions on specific manifolds (e.g., von Mises-Fisher and extensions (**??**)) and projection-based approaches (**?**). There is also a recent literature developing algorithms for certain classes of manifolds; for example **?** develop a geodesic Monte Carlo algorithm.

The primary contribution of this article is to propose a broad class of Bayesian priors that are formally close to a constrained space and can effectively solve both of our problems simultaneously. ~~Importantly, the constrained subset of the original parameter space can be very small or even have measure zero, leading to some technical complications.~~ The proposed class is very broad and acts to modify an initial unconstrained prior, having an arbitrary form, to be concentrated near the constrained space to an extent controlled by a hyperparameter. In addition, due to the simple form and lack of any sharp parameter constraints, general off-the-shelf sampling algorithms can be applied directly. Bypassing the need to develop or code complex and specialized algorithms is a major practical advantage.

The remainder of this article is organized as follows. Section 2 presents the framework for constructing the new class of Bayesian priors, introduces relevant notation, and supplies a general approach for constructing the relaxed posteriors when the constrained spaces have positive and zero measure. Representative examples are included. Section 3 contains a rigorous discussion of the theoretical aspects of our new class of relaxed posteriors. Section 4 reviews an approach for efficient posterior computation within the context of this article. Sections 5 and 6 contain detailed examples of new modeling attainable through out framework, and Section 7 contains a discussion.

2 Constraint Relaxation Methodology

2.1 Notation and Framework

Assume that $\theta \in \mathcal{D} \subset \mathcal{R}$ is an unknown continuous parameter, with $\dim(\mathcal{R}) = r < \infty$. The constrained sample space $\mathcal{D}$ is embedded in the r-dimensional Euclidean space $\mathcal{R}$, and can have either zero or positive Lesbesgue measure on $\mathcal{R}$. The traditional Bayesian approach to including constraints requires a 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 then

$$\pi_{\mathcal{D}}(\theta \mid Y) \quad \propto \quad \pi_{\mathcal{D}}(\theta)\mathcal{L}(\theta; Y), \tag{1}$$

where $\mathcal{L}(\theta; Y)$ is the likelihood function. We assume in the sequel that the restricted prior $\pi_{\mathcal{D}}(\theta) \propto \pi_{\mathcal{R}}(\theta)\mathbb{1}_{\mathcal{D}}(\theta)$, with $\pi_{\mathcal{R}}(\theta)$ an unconstrained distribution on $\mathcal{R}$ and $\mathbb{1}_{\mathcal{D}}(\theta)$ an indicator function that the constraint is satisfied.

As noted in Section 1, there are two primary problems motivating this article. The first is that it is often too restrictive to assume that θ is *exactly* within $\mathcal{D}$ *a priori*, and often is more plausible to assume that θ has high probability of falling within a small neighborhood of $\mathcal{D}$. The second is that the difficulty of

posterior sampling from (1) has greatly limited the scope of modeling, and there is a critical need for general algorithms that are tractable for a broad variety of choices of prior, likelihood and constraint.

In attempting to address these problems, we propose to replace (1) with the following *COnstraint RElaxed* (CORE) posterior density:

$$\tilde{\pi}_\lambda(\theta) \propto \mathcal{L}(\theta; Y)\pi_{\mathcal{R}}(\theta) \exp\Big(-\frac{1}{\lambda}\|v_{\mathcal{D}}(\theta)\|\Big), \tag{2}$$

where we repress the conditioning on data Y in $\tilde{\pi}_\lambda(\theta)$ for concise notation and use $\|v_{\mathcal{D}}(\theta)\|$ as a 'distance' from θ to the constrained space. We assume $\pi_{\mathcal{R}}(\theta)$ is proper and absolutely continuous with respect to Lesbesgue measure $\mu_{\mathcal{R}}$ on $\mathcal{R}$. As such, the constraint relaxed posterior $\tilde{\pi}_\lambda(\theta)$ corresponds to a coherent Bayesian probability model.

The hyperparameter $\lambda > 0$ controls how concentrated the prior is around $\mathcal{D}$, and as $\lambda \to 0$ the kernel $\exp\big(-\lambda^{-1}\|v_{\mathcal{D}}(\theta)\|\big)$ converges to $\mathbb{1}_{\mathcal{D}}(\theta)$ in a pointwise manner, ~~excluding $\theta \in \partial\mathcal{D}$ on the boundary of $\mathcal{D}$.~~ The statement about the boundary is 1) irrelevant when $\mathcal{D}$ has a positive measure and 2) not true if $\mathcal{D}$ is a manifold of lower dimension — the boundary of the manifold would be itself relative to the ambient space. For all $\lambda > 0$, $\tilde{\pi}_\lambda(\theta)$ introduces support outside of $\mathcal{D}$, creating a relaxation of the constraint. Both the value of λ and the choice of $\|v_{\mathcal{D}}(\theta)\|$ are important in controlling the concentration of the prior around $\mathcal{D}$. In the next subsection, we discuss the choice of distance. In the subsequent subsections, we discuss details specific to the cases in which $\mathcal{D}$ has positive and zero measure, respectively.

2.2 Distance to Constrained Space

In choosing $\|v_{\mathcal{D}}(\theta)\|$, a minimal condition is that $\|v_{\mathcal{D}}(\theta)\|$ is zero for $\theta \in \mathcal{D}$ and positive for $\theta \notin \mathcal{D}$. In addition, it is typically appealing for $\|v_{\mathcal{D}}(\theta)\|$ to be increasing as θ moves 'further away' from the restricted region $\mathcal{D}$. It then remains to characterize how we quantify 'further away'. In this regard, we focus on two types of distances - (I) *Direct Distances* that measure how far θ is from $\mathcal{D}$; and (II) *Indirect Distances* that measure how far certain functions of θ are from their constrained values when $\theta \in \mathcal{D}$.

(I) *Direct Distances*: Let $\|v_{\mathcal{D}}(\theta)\|$ correspond to the distance from θ to the closest location in $\mathcal{D}$. We focus in particular on the simple choice:

$$\|v_{\mathcal{D}}(\theta)\| = \inf_{x \in \mathcal{D}} \|\theta - x\|_k, \tag{3}$$

where $\|\cdot\|_k$ denotes a k-norm distance, typically $k = 1$ or 2.

(II) *Indirect Distances*: Let $\|v_{\mathcal{D}}(\theta)\|$ correspond to

$$\|v_{\mathcal{D}}(\theta)\| = \inf_{x \in \mathcal{D}} \|f(\theta) - f(x)\|_k, \tag{4}$$

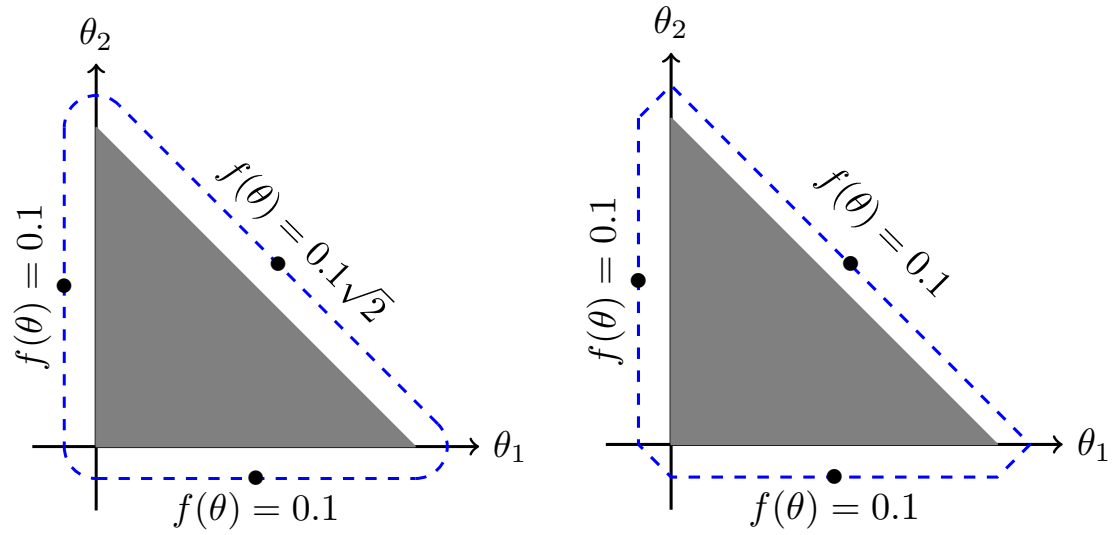
which reduces to (3) when $f(\theta) = \theta$. For example, the constraint may correspond to an upper bound on the function $f(\theta)$, so that it is natural to choose $\|v_{\mathcal{D}}(\theta)\|$ to correspond to how far $f(\theta)$ exceeds this upper bound.

To obtain insight into how the prior concentrates around $\mathcal{D}$, and the impact of the choice of distance, it is useful to introduce the concept of d-expansion of $\mathcal{D}$. In particular, the d-expansion of $\mathcal{D}$ with respect to $\|v_{\mathcal{D}}(\theta)\|$ is denoted as

$$\mathcal{D}_{\|v_{\mathcal{D}}(\theta)\|}(d) = \{\theta \in \mathcal{R} : \|v_{\mathcal{D}}(\theta)\| \leq d\},$$

which expands the restricted region $\mathcal{D}$ to add a 'halo' of size d of θ values that are within d of $\mathcal{D}$. Modifying the choice of distance will lead to some changes in the shape of $\mathcal{D}_{\|v_{\mathcal{D}}(\theta)\|}(d)$.

To illustrate, we consider a constrained space under 3 inequalities $\mathcal{D} = \{(\theta_1, \theta_2) : \theta_1 > 0, \theta_2 > 0, \theta_1 + \theta_2 < 1\}$. For type-I distance, we use 2-norm $\inf_{x \in \mathcal{D}} \|\theta - x\|_2$. This creates a space expansion of θ equally along the boundary $\partial\mathcal{D}$ (Figure 1(a)). On the other hand, one might be interested in the total violation to the inequalities, as represented in a function $f(\theta) = (-\theta_1)_+ + (-\theta_2)_+ + (\theta_1 + \theta_2 - 1)_+$ where $(x)_+ = x$ if $x > 0$ and 0 otherwise. The associated type-II distance is $\|v_{\mathcal{D}}(\theta)\| = f(\theta)$, and the expanded neighborhood is shown in Figure 1(b).



(a) Expanded neighborhood using type-I distance (b) Expanded neighborhood using type-II distance

Figure 1: The boundary of the neighborhood (blue dashed line) along $\|v_{\mathcal{D}}(\theta)\| = 0.1$ formed by two types of distances, (a) type-I (direct distance) and (b) type-II (indirect distance), around a constrained space formed by three inequalities (gray area).

The distance can be chosen based on prior belief about how the probability should decrease outside of $\mathcal{D}$. In the absence of prior knowledge supporting one choice over another, one can potentially try several different choices, while assessing sensitivity of the results. Even though the precise shape of the d-expanded regions, and hence the tails of the prior density outside of $\mathcal{D}$, can depend on the choice of distance, results are

typically essentially indistinguishable practically for different choices. Subtle differences in the d-expanded regions, such as those shown in Figure 1, tend to lead to very minimal differences in posterior inferences in our experience.

2.3 Constrained Space with Positive Measure

We start with the case when $\mathcal{D}$ is a subset of $\mathcal{R}$ with positive Lesbesgue measure, $\mu_{\mathcal{R}}(\mathcal{D}) > 0$. The sharply constrained density is then simply a truncated version of the unconstrained one, with

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

which is defined with respect to $\mu_{\mathcal{R}}$. For constraint relaxation, we replace the indicator with an exponential function of distance

$$\tilde{\pi}_{\lambda}(\theta) = \frac{\mathcal{L}(\theta; Y)\pi_{\mathcal{R}}(\theta)\exp\big(-\lambda^{-1}\|v_{\mathcal{D}}(\theta)\|\big)}{\int_{\mathcal{R}} \mathcal{L}(\theta; Y)\pi_{\mathcal{R}}(\theta)\exp\big(-\lambda^{-1}\|v_{\mathcal{D}}(\theta)\|\big)d\mu_{\mathcal{R}}(\theta)} \propto \mathcal{L}(\theta; Y)\pi_{\mathcal{R}}(\theta)\exp\big(-\lambda^{-1}\|v_{\mathcal{D}}(\theta)\|\big) \tag{5}$$

which is also absolutely continuous with respect to $\mu_{\mathcal{R}}$. Direct distances (Type-I) are often natural choices for $\|\nu_{\mathcal{D}}(\theta)\|$ in this setting.

Expression (5) replaces the function $\mathbb{1}_{\mathcal{D}}(\theta)$, which is equal to one for $\theta \in \mathcal{D}$ and zero for $\theta \notin \mathcal{D}$, with $\exp\big(-\lambda^{-1}\|v_{\mathcal{D}}(\theta)\|\big)$, which is still equal to one for $\theta \in \mathcal{D}$ but decreases exponentially as θ moves away from $\mathcal{D}$. The prior is effectively *shrinking* θ towards $\mathcal{D}$, with the exponential tails reminiscent of the double exponential (Laplace) prior that forms the basis of the widely used Lasso procedure. Potentially, we could allow a greater degree of robustness to the choice of $\mathcal{D}$ by choosing a heavier tailed function in place of the exponential; for example, using the kernel of a generalized double Pareto or t-density. However, such choices introduce an additional hyperparameter, and we focus on the exponential for simplicity.

Section 3.1 addresses the relationship between the strictly constrained posterior density and the relaxed posterior density, identifying cases where inferences become more similar as $\lambda \to 0$. For now, we consider the following example which illustrates how relaxation of the support can provide more realistic modeling.

Example: Gaussian with inequality constraints

As a simple illustrative example, we consider a Gaussian likelihood with inequality constraints on the mean. In particular, let

$$y_i \overset{iid}{\sim} \mathrm{No}(\theta, 1), \quad i = 1, \ldots, n, \qquad \pi_{\mathcal{R}}(\theta) = \mathrm{No}(\theta; 0, 1000).$$

Suppose there is prior knowledge that $\theta < 1$. The posterior under a sharply constrained model is

$$\pi_{\mathcal{D}}(\theta \mid Y) \propto \sigma^{-1}\phi\left(\frac{\theta - \mu}{\sigma}\right)\mathbb{1}_{\theta<1}, \quad \mu = \frac{\bar{y}n}{1/1000 + n}, \quad \sigma^2 = \frac{1}{1/1000 + n},$$

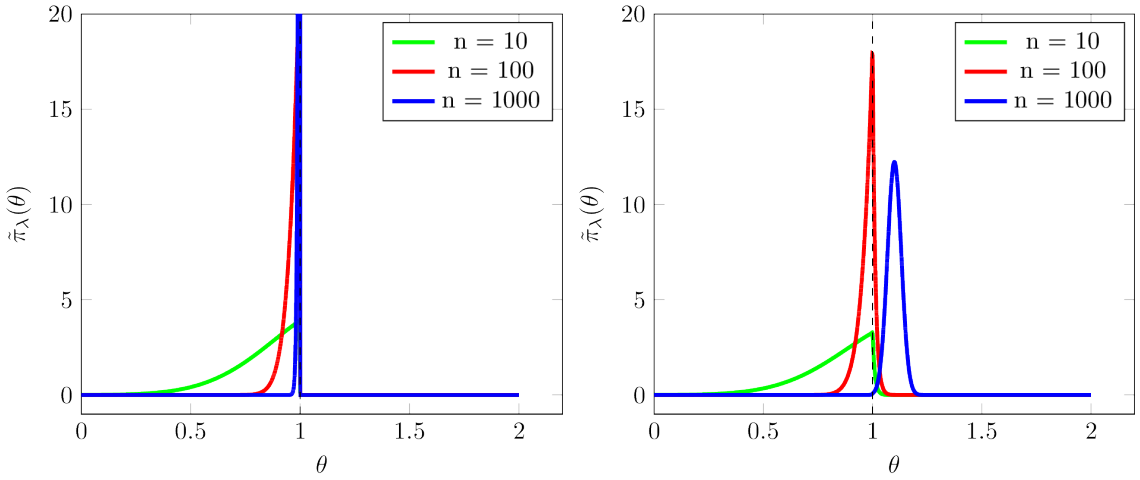
where ϕ denotes the density of the standard Gaussian. This posterior corresponds to $\text{No}_{(-\infty,1)}(\mu, \sigma^2)$, which is a $\text{No}(\mu, \sigma^2)$ distribution truncated to the region $\theta < 1$.

If θ is indeed less than one, incorporation of the constraint has the benefit of reducing uncertainty in the posterior distribution, leading to greater concentration around the true value. However, even slight misspecification of the constrained region can lead to biased inferences; for example, perhaps $\theta = 1.2$. In this case, as the sample size n increases, the sharply constrained posterior distribution $\text{No}_{(-\infty,1)}(\mu, \sigma^2)$ becomes more and more concentrated near the $\theta = 1$ boundary as illustrated in Figure 2(a).

The constraint relaxed (CORE) approach is well justified in this case as it allows some probability to be allocated to the $\theta > 1$ region under the posterior:

$$\tilde{\pi}_\lambda(\theta) \propto \sigma^{-1}\phi\left(\frac{\theta - \mu}{\sigma}\right) \exp\left(-\frac{(\theta - 1)_+}{\lambda}\right), \quad \mu = \frac{\bar{y}n}{1/1000 + n}, \quad \sigma^2 = \frac{1}{1/1000 + n}$$

where $(\theta - 1)_+$ is the direct distance to the constrained space. With a small value $\lambda = 10^{-2}$, representing high prior concentration very close to $\mathcal{D} = (-\infty, 1)$, the relaxed posterior is close to the sharply constrained one for small to moderate sample sizes, as illustrated in Figure 2(b). However, as n increases, the posterior becomes concentrated around the true θ value, even when it falls outside of the constrained space.



(a) Sharply constrained posterior

(b) Constraint relaxed posterior

Figure 2: Posterior densities for a Gaussian mean (θ) under sharp (panel (a)) and relaxed constraints (panel (b)). The constraint region is $\mathcal{D} = (-\infty, 1)$ and the true value is $\theta = 1.2$, which falls slightly outside $\mathcal{D}$ representing misspecification.

2.4 Constrained Space with Zero Measure

We now consider the case in which $\mathcal{D}$ is a measure zero subset of $\mathcal{R}$ with respect to $\mu_{\mathcal{R}}$, the Lebesgue measure on $\mathcal{R}$. Since $\mathcal{D}$ is measure zero, the sharply constrained posterior density can no longer be constructed by

truncating the unconstrained posterior on $\mathcal{D}$ and renormalizing. Rather, one must first use techniques from geometric measure theory to define a regular conditional probability for sharp constraints on $\mathcal{D}$. This additional step gives rise to a number of technical difficulties, discussed in more detail in Section 3. Most notably, the use of regular conditional probability alters the formulation of the distance function $\|\nu_{\mathcal{D}}(\theta)\|$ used in constraint relaxation.

As a motivating example, suppose the constrained space is the line $\mathcal{D} = \{\theta : \theta_1 + \theta_2 = 1\}$ and $\pi_{\mathcal{R}}(\theta) = 1\{\theta_1 \in (0,1), \theta_2 \in (0,1)\}$ is a uniform distribution on the unit square. As $\mu_{\mathcal{R}}(\mathcal{D}) = 0$, the sharply constrained posterior (5) is undefined. To circumvent this issue, one possibility is to replace (θ_1, θ_2) with $(\theta_1, 1-\theta_1)$, reducing the dimension of the problem. This approach is equivalent to building a regular conditional probability on $\mathcal{D}$ (see Section 3.2) which results in a posterior density defined with respect to the normalized 1-Hausdorff measure (arclength) on $\mathcal{D}$. However, in this reparameterized lower dimensional setting, the constraint is strictly enforced, eliminating any relaxation away from $\mathcal{D}$.

Alternatively, one can create a relaxed posterior in the following manner. Motivated by the original constraint, $\theta_1 + \theta_2 = 1$, we set $\nu_{\mathcal{D}}(\theta) = \theta_1 + \theta_2 - 1$ so that $\|\nu_{\mathcal{D}}(\theta)\| = 0$ when $\theta \in \mathcal{D}$ and is positive otherwise. With this distance function, we define the relaxed posterior as

$$\tilde{\pi}_\lambda(\theta) \propto \mathcal{L}(\theta; Y)\pi_{\mathcal{R}}(\theta) \exp\left(-\frac{\|\nu_{\mathcal{D}}(\theta)\|}{\lambda}\right) = \mathcal{L}(\theta; Y) \exp\left(-\frac{\|\theta_1 + \theta_2 - 1\|}{\lambda}\right) \mathbb{1}\{(\theta_1 \in (0,1), \theta_2 \in (0,1)\}.$$

This density, defined with respect to Lebesgue measure on the plane, makes use of the original constraint, $\theta_1 + \theta_1 = 1$, to define a distance function which in turn allows for relaxation away from the constrained space.

In the preceding example, the constrained space was defined in terms of an equation involving the components of the parameter θ. Many other constraints, such as simplex or Stiefel manifold constraints, can be expressed similarly. Thus, it is natural to restrict ourselves to the setting in which $\mathcal{D}$ can be represented implicitly as the unique solution set of a consistent system of equations $\{v_j(\theta) = 0\}_{j=1}^s$. The constraint functions, $\{v_j\}_{j=1}^s$, must satisfy additional assumptions as stated in Section 3.2. For the moment, we highlight their use in defining a distance to the constrained space and constructing the sharply constrained posterior.

Given a set of constraints functions, let $\nu_{\mathcal{D}}(\theta) = [v_1(\theta), \ldots, v_s(\theta)]^T$ be a vector valued function from $\mathcal{R}$ to the s-dimensional Euclidean space $\mathbb{R}^s$. Note that $\nu_{\mathcal{D}}$ need not be onto $\mathbb{R}^s$. Throughout this subsection, we then define the distance function as $\|v_{\mathcal{D}}(\theta)\| = \|\nu_{\mathcal{D}}(\theta)\|_1 = \sum_{j=1}^s |v_j(\theta)|$. The terms $|v_j(\theta)|$ typically take the form of type-II distances as discussed in Section 2.1. Under some mild assumptions on the constraint functions, the preimages, $\nu_{\mathcal{D}}^{(-1)}(x)$, will be $(r-s)$-dimensional submanifolds of $\mathcal{R}$ for $\mu_{\mathbb{R}^s}$-almost every x in the range of $\nu_{\mathcal{D}}$. In particular, $\nu_{\mathcal{D}}^{(-1)}(0) = \mathcal{D}$. Observe that in the motivating example, the pre-images $\nu_{\mathcal{D}}^{(-1)}(x)$ are the lines $\theta_1 + \theta_2 = x + 1$ which are also one-dimensional sub-manifolds of $\mathcal{R}$ similar to $\mathcal{D}$.

More generally, while $\mathcal{D}$ has zero r-dimensional Lebesgue measure, corresponding to zero volume within $\mathcal{R}$, it will have non-zero $(r-s)$-dimensional surface area, corresponding to the normalized $(r-s)$-dimensional Hausdorff measure, denoted by $\bar{\mathcal{H}}^{(r-s)}$. To construct the sharply constrained posterior density, we renormalize the fully constrained density by its integral with respect to the normalized $(r-s)$-dimensional Hausdorff measure yielding a valid probability on the constrained space also known as a regular conditional probability. Using the normalized Hausdorff measure, we take

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

where the Jacobian of $\nu_{\mathcal{D}}$, $J(v_{\mathcal{D}}(\theta)) = \sqrt{(D\nu_{\mathcal{D}})'(D\nu_{\mathcal{D}})}$, is assumed to be positive and arises from the co-area formula (**?**). The Jacobian, in part, accounts for the differences in dimension between $\mathcal{D}$ and $\mathcal{R}$.

As in the positive measure case, to relax the constraint we begin with (1) and replace the indicator function with $\exp(-\|v_{\mathcal{D}}(\theta)\|/\lambda)$, adding support for $\|v_{\mathcal{D}}(\theta)\| > 0$. Therefore, the relaxed density is

$$\tilde{\pi}_{\lambda}(\theta) \propto \mathcal{L}(\theta; Y)\pi_{\mathcal{R}}(\theta)\exp(-\frac{1}{\lambda}\|v_{\mathcal{D}}(\theta)\|) = \mathcal{L}(\theta; Y)\pi_{\mathcal{R}}(\theta)\exp\left(-\frac{1}{\lambda}\sum_{j=1}^{s}\|v_j(\theta)\|\right). \tag{6}$$

Unlike the sharply constrained density however, the relaxed density is supported on $\mathcal{R}$ and is defined with respect to $\mu_{\mathcal{R}}$. As such, the Jacobian of $\nu_{\mathcal{D}}$ does not appear. This statement "As such..." is not at all a clear explanation of what is going on. This is a notable difference from the positive measure case, as the sharply constrained posterior density is no longer a pointwise limit of the relaxed density in general. I think we need to explain & justify the approach more carefully here. It is hard to argue that the (lack of) Jacobian term is an advantage but maybe we can argue that it is just a feature of our approach.

To illustrate the construction of constraint relaxed posterior densities in the measure zero subset case, we consider an example of a torus in $\mathbb{R}^3$.

Example: Support expansion near a curved torus

Let $\mathcal{R} = \mathbb{R}^3$ and consider a curved torus

$$\mathcal{D} = \left\{\theta : (\theta_1, \theta_2, \theta_3) = \big((1 + 0.5\cos\alpha_1)\cos\alpha_2, (1 + 0.5\cos\alpha_1)\sin\alpha_2, 0.5\sin\alpha_1\big), (\alpha_1, \alpha_2) \in [0, 2\pi)^2\right\}.$$

which has intrinsic dimension two and zero 3-dimensional Lesbesgue measure, $\mu_{\mathcal{R}}(\mathcal{D}) = 0$. **?** previously considered a uniform Hausdorff density $\pi_{\mathcal{D}}(\theta \mid Y) \propto 1$ over this compact manifold and utilized the transformed Lesbesgue density based on (α_1, α_2) for sampling strictly on the manifold.

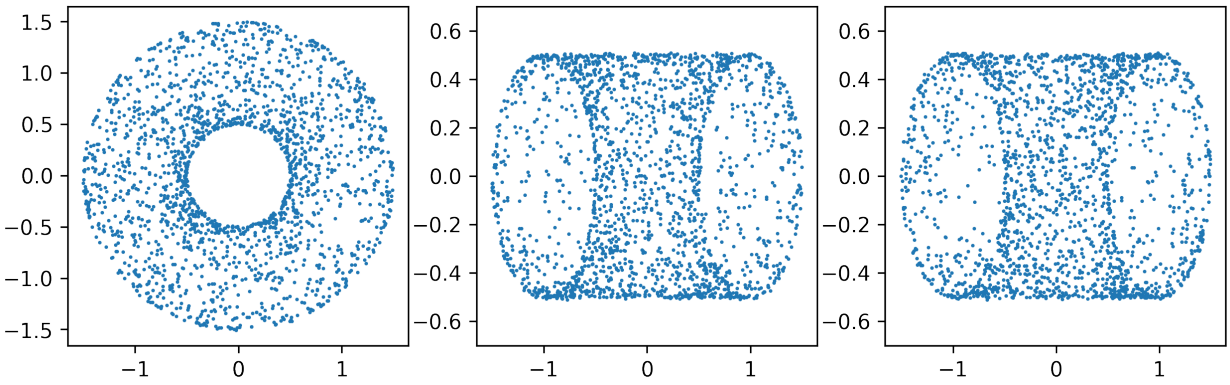
We now consider a different task by constructing a relaxed distribution near the torus. Here, $\mathcal{D}$ can be defined implicitly as the solution set to the equation

$$\nu_{\mathcal{D}}(\theta) = \left(1 - \sqrt{\theta_1^2 + \theta_2^2}\right)^2 + \theta_3^2 - \frac{1}{4} = 0.$$

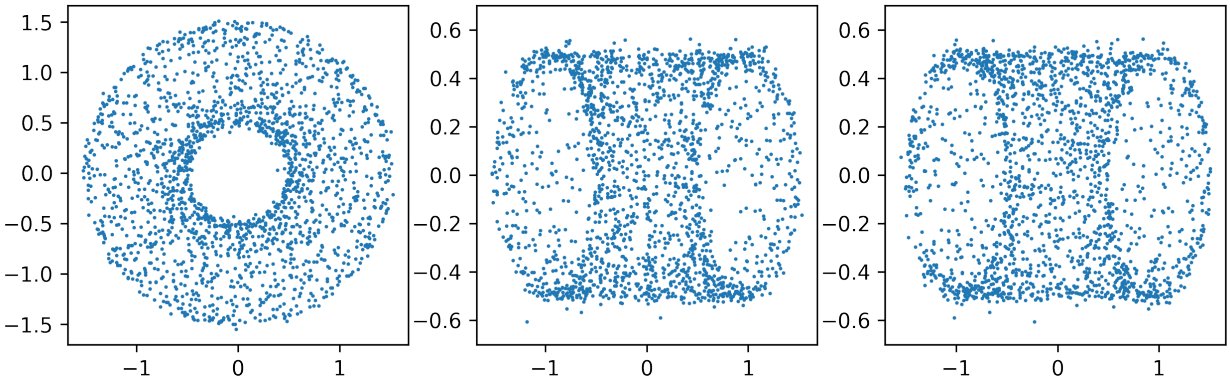
Using this, we define the relaxed posterior density

$$\begin{aligned}\tilde{\pi}_\lambda(\theta) &\propto J(\nu_{\mathcal{D}}(\theta)) \exp(-\lambda^{-1}\|v_{\mathcal{D}}(\theta)\|) \\ &= 2\sqrt{(1-\sqrt{\theta_1^2+\theta_2^2})^2+\theta_3^2}\exp\left\{-\lambda^{-1}\left|\left(1-\sqrt{\theta_1^2+\theta_2^2}\right)^2+\theta_3^2-\frac{1}{4}\right|\right\}\end{aligned} \tag{7}$$

which is defined with respect to 3-dimensional Lesbesgue measure. Here we are throwing the Jacobian term on the relaxed distribution (which is absolutely necessary for uniform sampling) but this is inconsistent with our presentation of the relaxed density without Jacobian. Figure 3 plots random samples from constraint relaxed posteriors under two different values of λ, corresponding to different degrees of relaxation. The samples were obtained using random walk Monte Carlo applied to (7).



(a) Constraint relaxed posterior with $\lambda = 0.01$



(b) Constraint relaxed posterior with $\lambda = 0.1$

Figure 3: Support expansion near a curved torus. Small λ controls the samples to be very close to the manifold (upper panel), while large λ adds support farther away from the constrained space (lower panel). Both densities correspond to more conventional Lesbesgue measure, instead of Hausdorff measure.

3 Theory

In this section, we present theoretic details pertaining to the behavior of the relaxed posterior. The primary focus will be on differences in the posterior expectation, $E[g(\theta)]$, under the sharply versus relaxed posteriors. Theorems bounding these differences in λ are presented for a suitable class of functions. All proofs are in the appendix.

Section 3.1 addresses the case in which $\mathcal{D}$ has positive measure, while Section 3.2 considers the more challenging measure zero case. Additionally, in Section 3.2, we offer a deeper exposition on the geometric measure theory concepts used in the construction of the constrained posterior, present additional properties required of the constraint functions, $\{v_j\}_{j=1}^s$, and state suitable choices of constraints for some common examples.

3.1 Constrained Space with Positive Measure

We focus on quantifying the difference between the sharply constrained and relaxed posterior distributions, both of which are absolutely continuous with respect to Lesbesgue measure on $\mathcal{R}$. The posterior expectation of g under the sharply constrained prior is

$$\mathbb{E}[g(\theta)|\theta \in \mathcal{D}] = \int_{\mathcal{D}} g(\theta)\pi_{\mathcal{D}}(\theta)d\mu_{\mathcal{R}} = \frac{\int_{\mathcal{D}} g(\theta)\mathcal{L}(\theta;Y)\pi_{\mathcal{R}}(\theta)d\mu_{\mathcal{R}}(\theta)}{\int_{\mathcal{D}} \mathcal{L}(\theta;Y)\pi_{\mathcal{R}}(\theta)d\mu_{\mathcal{R}}(\theta)}. \tag{8}$$

Similarly, the posterior expectation of g under the relaxed prior is

$$\mathbb{E}_{\tilde{\pi}_\lambda}[g(\theta)] = \int_{\mathcal{R}} g(\theta)\tilde{\pi}_\lambda(\theta)d\mu_{\mathcal{R}} = \frac{\int_{\mathcal{R}} g(\theta)\mathcal{L}(\theta;Y)\pi_{\mathcal{R}}(\theta)\exp(-\|v_{\mathcal{D}}(\theta)\|/\lambda)d\mu_{\mathcal{R}}(\theta)}{\int_{\mathcal{R}} \mathcal{L}(\theta;Y)\exp(-\|v_{\mathcal{D}}(\theta)\|/\lambda)\pi_{\mathcal{R}}(\theta)d\mu_{\mathcal{R}}(\theta)}. \tag{9}$$

A short calculation shows that the difference $\mathbb{E}[g(\theta)|\theta \in \mathcal{D}] - \mathbb{E}_{\tilde{\pi}_\lambda}[g(\theta)]$ depends on two quantities: the posterior probability of $\mathcal{D}$ under the constrained posterior, and the average magnitude of $|g(\theta)|$ over $\mathcal{R} \setminus \mathcal{D}$ with respect to the relaxed posterior. These results are summarized in Lemma 1.

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

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

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

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

While this Lemma indicates $\mathbb{E}_{\tilde{\pi}_\lambda}[g(\theta)] \to \mathbb{E}[g(\theta|\theta \in \mathcal{D}]$ as $\lambda \to 0^+$, for a fixed $\lambda > 0$, large differences can arise if (i) $|g(\theta)|\mathcal{L}(\theta; Y) \gg \exp(-\lambda^{-1}\|\nu_{\mathcal{D}}(\theta)\|)$ on average over a subset of $\mathcal{R} \setminus \mathcal{D}$ or (ii) the posterior probability of $\mathcal{D}$ is small with respect to the unconstrained posterior.

With regards to (i), consider the case where $\mathcal{F}$ is a measureable subset of $\mathcal{R}$ for which $\mathcal{F} \cap \mathcal{D} = \emptyset$ and let $g(\theta) = \mathbb{1}_{\mathcal{F}}(\theta)$. Then, $\mathbb{E}[g(\theta)|\theta \in \mathcal{D}] = 0$. However, $\mathbb{E}_{\tilde{\pi}_\lambda}[g(\theta)]$ may be large if $\mathcal{L}(\theta; Y) \gg \exp(-\lambda^{-1}\|\nu_{\mathcal{D}}(\theta)\|)$ for $\theta \in \mathcal{F}$. As such, over $\mathcal{F}$ the likelihood is dominating the relaxation allowing the CORE density to assign positive probability to $\mathcal{F}$ which is not possible for the sharply constrained density.

In the case of (ii), the error is inversely proportional to the square of the unconstrained posterior probability of $\mathcal{D}$. Thus, when $\theta \in \mathcal{D}$ is unlikely under the unconstrained model any relaxation away from the constraint is amplified. This effect in particular demonstrates the usefulness of constraint relaxation. If constraints are misspecified and $\theta \in \mathcal{D}$ is not supported by the data, the posterior estimates using the relaxed density can display a large sensitivity to the choice of λ indicating that the constraints themselves should be re-evaluated.

Turning to the issue of using CORE to estimate posterior expectation of g under sharp constraints, Lemma 1 indicates that 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 $\mathbb{E}[g(\theta)|\theta \in \mathcal{D}]$. The answer ultimately depends on the choice of distance function and its behavior on $\mathcal{R} \setminus \mathcal{D}$. We supply the following theorem when the distance function $\nu_{\mathcal{D}}(\theta) = \inf_{x \in \mathcal{D}} ||\theta - x||_2$. One can use the analysis contained in the proof of this theorem (Appendix A) as a guide to construct convergence rates for a different choice of $\|\nu_{\mathcal{D}}(\theta)\|$.

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 an 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. While this bound holds for general $\mathcal{D}$, one can attain a more intuitive bound depending on the surface area of $\mathcal{D}$ when it has a compact C^1 boundary. This case and more details regarding the coefficient in the error rate are contained in the proof but omitted here for brevity.

These results have some important implications analytically and numerically. First, let Π and $\tilde{\Pi}_\lambda$ be the measures induced by the sharply constrained and relaxed densities, respectively. The results from Lemma 1 and Theorem 1 allow us to bound the total variation between these measures.

Corollary 1. *For* $\lambda \ll 1$,

$$\|\Pi - \tilde{\Pi}_\lambda\|_{TV} = \sup_{A \in \mathcal{A}} |\Pi(A) - \tilde{\Pi}_\lambda(A)| = O(\sqrt{\lambda}),$$

where $\mathcal{A}$ is the Borel sigma-algebra, and we may conclude that $\|\Pi - \tilde{\Pi}_\lambda\|_{TV} \to 0$ as $\lambda \to 0^+$.

3.2 Constrained Space with Zero Measure

We begin with a review of some important concepts of geometric measure theory. In addition to supporting the analysis, we are reviewing these topics to offer insight into the behavior of the relaxed posterior. We begin with the definition of the d-dimensional Hausdorff measure.

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

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

We denote the normalized d-dimensional Hausdorff measure as $\bar{\mathcal{H}}^d(A) = \frac{\Gamma(\frac{1}{2})^d}{2^d \Gamma(\frac{d}{2}+1)} \mathcal{H}^d(A)$. When $d = r$, Lesbesgue and normalized Hausdorff measures coincide $\mu_{\mathbb{R}^m}(A) = \bar{\mathcal{H}}^d(A)$ (**?**). Additionally, for a subset $\mathcal{D}$, there exists a unique, critical value d such that $\bar{\mathcal{H}}^s(\mathcal{D}) = 0$ for $s > d$ and ∞ for $s < d$. The critical value, d, is referred to as the Hausdorff dimension of $\mathcal{D}$, which agrees with the usual notion of dimension when $\mathcal{D}$ is a piecewise smooth manifold. In fact, when $\mathcal{D}$ is a compact, d-dimensional submanifold of $\mathbb{R}^m$, it will have Hausdorff dimension d so that $\bar{\mathcal{H}}^d(\mathcal{D})$ is the d-dimensional surface area of A. In the zero measure setting, as discussed in Section 2, we are focusing on the case where $\mathcal{D}$ is an $(r-s)$-dimensional submanifold of $\mathcal{R}$. As such, it is natural to define the sharply constrained posterior with respect to $\bar{\mathcal{H}}^{r-s}$, which is referred to as a regular conditional probability, r.c.p **?**.

Defining the r.c.p on the measure zero constrained space $\mathcal{D}$ and the subsequent analysis requires the co-area formula.

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

$$\int_{\mathbb{R}^r} g(\theta) J[v(\theta)] d\mu_{\mathbb{R}^r}(\theta) = \int_{\mathbb{R}^s} \left(\int_{v^{-1}(y)} g(\theta) d\bar{\mathcal{H}}^{r-s}(\theta) \right) d\mu_{\mathbb{R}^s}(y). \tag{10}$$

Recall, we previously assumed that $\mathcal{D}$ can be defined implicitly as the solution set to a system of s equations, $\{v_j(\theta) = 0\}_{j=1}^s$, and we defined the map $\nu_{\mathcal{D}}(\theta) = [v_1(\theta), \ldots, v_s(\theta)]$ from our parameter space, $\mathcal{R}$, to the Euclidean space, $\mathbb{R}^s$. These constraint functions must adhere to some additional restrictions.

(a) $v_j : \mathcal{R} \to \mathbb{R}$ 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 $1 \leq j < k \leq s$.

Property (a) guarantees that $\nu_{\mathcal{D}}$ is itself Lipschitz so the co-area formula applies. The remaining properties (b)-(d) are constructed so that when $x \in \mathbb{R}^s$ is near zero, the preimage $\nu_{\mathcal{D}}^{(-1)}(x)$ is also an $(r-s)$-dimensional submanifold corresponding to a perturbation of the constrained space $\mathcal{D}$. In the remainder of this section, we assume that $\nu_{\mathcal{D}}^{(-1)}(x)$ is an $(r-s)$-dimensional submanifold of a.e x in the range of $\mathcal{D}$. While this is a very strong assumption, to attain relaxation near $\mathcal{D}$, the transversality condition (d) assures this for x near 0.

Criteria (a) - (d) may seem restrictive. However, many measure zero constraints can be defined implicitly to satisfy (b) - (d). In Table 1, we offer a few examples. An initial set of constraint functions can typically be modified to satisfy the Lipschitz condition by truncating the original parameter space $\mathcal{R}$ or by composing the constraints with bounded functions. The former choice was used for the Unit sphere and Stiefel manifold constraints in the table.

$\mathcal{R}$	$\mathcal{D}$	$\dim(\mathcal{R})$	$\dim(\mathcal{D})$	Constraint functions
$[0,1]^r$	Probability simplex, Δ^{r-1}	r	$r-1$	$v_1(\theta) = \sum(\theta) - 1$
$\mathbb{R}^r$	Line, span$\{\vec{u}\}$ $\vec{u} \neq \vec{0}$	r	1	$v_j(\vec{\theta}) = \vec{\theta}^T \vec{b}_j$ $\{\vec{b}_1, \ldots, \vec{b}_{r-1}\}$ a basis for span$\{\vec{u}\}^\perp$
$[-1,1]^r$	Unit sphere, $\mathbb{S}^{r-1}$	r	$r-1$	$v_1(\theta) = \|\theta\|^2 - 1$
$[-1,1]^{n\times k}$	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.

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

Theorem 3. *(?) Assume that $J(\nu_{\mathcal{D}}(\theta)) > 0$ and that for each $z \in \mathbb{R}^s$ there is a finite non-negative p_z such that,*

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

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

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

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

By construction, $\{\theta : \nu_{\mathcal{D}}(\theta) = z\}$ is an $(r-s)$ dimensional submanifold of $\mathcal{R}$ for $\mu_{\mathbb{R}^s}$-a.e. z in the range of $\nu_{\mathcal{D}}$. As such, it follows that one should take $p_z = r - s$. While it is still possible that $m^p(z) \in \{0, \infty\}$ for some z, we need not worry about these cases as they occur on a set of $\mu_{\mathbb{R}^s}$ measure zero . Most importantly, setting $z = 0$ allows us to define

$$\pi_{\mathcal{D}}(\theta|\theta \in \mathcal{D}, Y) = \frac{1}{m^{r-s}(0)} \frac{\mathcal{L}(\theta; Y)\pi_{\mathcal{R}}(\theta)\mathbb{1}_{\mathcal{D}}(\theta)}{J(\nu_{\mathcal{D}}(\theta))} \tag{11}$$

as the constrained posterior density as originally stated in Section 2.2.

To understand the effects of constraint relaxation, consider a Borel subset, $\mathcal{F}$, of $\mathcal{R}$. Under the sharply constrained posterior,

$$\begin{aligned} P(\theta \in \mathcal{F}|Y) &= \int_{\mathcal{F}} \pi_{\mathcal{D}}(\theta|Y) d\bar{\mathcal{H}}^{r-s}(\theta) = \frac{\int_{\mathcal{F}} \mathcal{L}(\theta; Y)\pi_{\mathcal{R}}(\theta) J^{-1}(v_{\mathcal{D}}(\theta))\mathbb{1}_{\mathcal{D}}(\theta) d\bar{\mathcal{H}}^{r-s}(\theta)}{\int_{\mathcal{D}} \mathcal{L}(\theta; Y)\pi_{\mathcal{R}}(\theta) J^{-1}(v_{\mathcal{D}}(\theta))\mathbb{1}_{\mathcal{D}}(\theta) d\bar{\mathcal{H}}^{r-s}(\theta)} \\ &= \frac{\int_{\mathcal{F}\cap\mathcal{D}} \mathcal{L}(\theta; Y)\pi_{\mathcal{R}}(\theta) J^{-1}(v_{\mathcal{D}}(\theta)) d\bar{\mathcal{H}}^{r-s}(\theta)}{\int_{\mathcal{D}} \mathcal{L}(\theta; Y)\pi_{\mathcal{R}}(\theta) J^{-1}(v_{\mathcal{D}}(\theta)) d\bar{\mathcal{H}}^{r-s}(\theta)}. \end{aligned} \tag{12}$$

Alternatively, under the relaxed posterior,

$$P(\theta \in \mathcal{F}|Y) = \int_{\mathcal{F}} \tilde{\pi}_\lambda(\theta) d\mu_{\mathcal{R}}(\theta) = \frac{\int_{\mathcal{F}} \mathcal{L}(\theta; Y)\pi_{\mathcal{R}}(\theta) \exp\left(-\frac{1}{\lambda} \sum_{j=1}^s \|v_j(\theta)\| \right) d\mu_{\mathcal{R}}(\theta)}{\int_{\mathcal{R}} \mathcal{L}(\theta; Y)\pi_{\mathcal{R}}(\theta) \exp\left(-\frac{1}{\lambda} \sum_{j=1}^s \|v_j(\theta)\| \right) d\mu_{\mathcal{R}}(\theta)} \tag{13}$$

Making use of the behavior of the preimages of $\nu_{\mathcal{D}}$, we can reexpress (13) through the co-area formula as

$$\begin{aligned} P(\theta \in \mathcal{F}|Y) &= \frac{\int_{\mathbb{R}^s} \left[\int_{\mathcal{F}\cap\nu_{\mathcal{D}}^{(-1)}(x)} \mathcal{L}(\theta; Y)\pi_{\mathcal{R}}(\theta) J^{-1}(v_{\mathcal{D}}(\theta)) \exp\left(-\frac{1}{\lambda} \sum_{j=1}^s \|v_j(\theta)\| \right) d\bar{\mathcal{H}}^{r-s}(\theta) \right] d\mu_{\mathbb{R}^s}(x)}{\int_{\mathbb{R}^s} \left[\int_{\mathcal{R}\cap\nu_{\mathcal{D}}^{(-1)}(x)} \mathcal{L}(\theta; Y)\pi_{\mathcal{R}}(\theta) J^{-1}(v_{\mathcal{D}}(\theta)) \exp\left(-\frac{1}{\lambda} \sum_{j=1}^s \|v_j(\theta)\| \right) d\bar{\mathcal{H}}^{r-s}(\theta) \right] d\mu_{\mathbb{R}^s}(x)} \\ &= \frac{\int_{\mathbb{R}^s} \left[\int_{\mathcal{F}\cap\nu_{\mathcal{D}}^{(-1)}(x)} \mathcal{L}(\theta; Y)\pi_{\mathcal{R}}(\theta) J^{-1}(v_{\mathcal{D}}(\theta)) d\bar{\mathcal{H}}^{r-s}(\theta) \right] \exp\left(-\frac{1}{\lambda}\|x\|_1 \right) d\mu_{\mathbb{R}^s}(x)}{\int_{\mathbb{R}^s} \left[\int_{\mathcal{R}\cap\nu_{\mathcal{D}}^{(-1)}(x)} \mathcal{L}(\theta; Y)\pi_{\mathcal{R}}(\theta) J^{-1}(v_{\mathcal{D}}(\theta)) d\bar{\mathcal{H}}^{r-s}(\theta) \right] \exp\left(-\frac{1}{\lambda}\|x\|_1 \right) d\mu_{\mathbb{R}^s}(x)} \end{aligned} \tag{14}$$

In the last line of the previous expression momentarily ignore the outer integrals over $\mathbb{R}^s$ and consider only the ratio of the inner integrals. Setting $x = 0$ in this expression yields the posterior probability of $\mathcal{F}$ under the sharply constrained posterior. For a different choice of x, this expression gives a sharp posterior probability of $\mathcal{F}$ under the constraint $\theta \in \nu_{\mathcal{D}}^{(-1)}(x)$.

Thus, one can view the relaxed posterior as a weighted average of a collection of sharply constrained posteriors over a set of $(r-s)$ dimensional submanifolds in a neighborhood of $\mathcal{D}$. For the distribution over an individual submanifold $\nu_{\mathcal{D}}^{(-1)}(x)$, the weight depends on the distance, $\|v_{\mathcal{D}}(\theta)\|$, λ, and on the magnitude of the likelihood, $\mathcal{L}(\theta;Y)$, over the set. Analogous to the positive measure case, if $\mathcal{L}(\theta;Y) \gg \exp(-\lambda^{-1}\|\nu_{\mathcal{D}}(\theta)\|$, the relaxed posterior can shift probability away from $\mathcal{D}$ and onto $\nu_{\mathcal{D}}^{(-1)}(x)$ when justified by the data.

Similar to the positive measure case, we can also give statements regarding expectation of $g(\theta)$ under the constrained and relaxed posterior. The posterior expectation of $g(\theta)$ under the sharp constraint $\theta \in \mathcal{D}$ is

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

Using the definition of $\tilde{\pi}_\lambda$ from Section 2.2, the expected value of $g(\theta)$ with respect to the relaxed density, denoted $\mathbb{E}_{\tilde{\Pi}}[g(\theta)]$, is

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

where $m_\lambda = \int_{\mathcal{R}} \pi_{\mathcal{R}}(\theta)\mathcal{L}(\theta;Y)\exp(-\lambda^{-1}\|v(\theta)\|_1)d\mu_{\mathcal{R}}(\theta)$. The remaining 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*

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

$$G(x) = \int_{\nu_{\mathcal{D}}^{-1}(x)} g(\theta)\frac{\pi_{\mathcal{R}}(\theta)\mathcal{L}(\theta;Y)}{J\nu_{\mathcal{D}}(\theta))}d\bar{\mathcal{H}}^{r-s}(\theta).$$

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

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

Corollary 2. *In addition to the assumptions of Theorem 4, suppose that both m and G are differentiable at 0. Then*

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

as $\lambda \to 0^+$.

The continuity assumptions of Theorem 4 and differentiability assumptions of Corollary 2 have some important consequences. Recall, the pairwise transversal intersection requirement, (d), assures that $\nu_{\mathcal{D}}^{(-1)}(x)$

behaves like a small perturbation of $\mathcal{D}$ when x is near zero. Therefore, if the unconstrained posterior, $\mathcal{L}(\theta; Y)\pi_{\mathcal{R}}(\theta)$, the Jacobian, $J(\nu_{\mathcal{D}}(\theta))$, and g are all continuous on an open neighborhood containing $\mathcal{D}$, the continuity assumptions of Theorem 4 will follow.

Unfortunately, the continuity requirement does have one large ramification concerning the total variation between the measures induced by the sharply constrained and relaxed posteriors, which we again denote by Π and $\tilde{\Pi}_\lambda$. Observe that $\Pi(\mathcal{D}) = 1$ whereas $\tilde{\Pi}_\lambda(\mathcal{D}) = 0$. As a result, the total variation between the measures is unity for all λ. This is not precluded by our theory as $g(\theta) = \mathbb{1}_{\mathcal{D}}(\theta)$ is not continuous on an open neighborhood containing $\mathcal{D}$; the assumptions of Theorem 4 do not apply.

However, while these measures will not converge in total variation, one can still attain good estimates of $E[g(\theta)|\theta \in \mathcal{D}]$ for many functions of interest using the relaxed density. The convergence rates in Corollary 2 are sub-linear in λ much like the positive measure case. However, they are dimension dependent and we emphasize that they require additional assumptions about the local behavior near $\mathcal{D}$. As an investigation into this convergence, we assess the approximation error with different λ using the von Mises–Fisher distribution. The results of this study are provided in the appendix.

4 Posterior Computation

The constraint relaxed posterior 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 as a general algorithm that tends to have good performance in a variety of settings.

4.1 Hamiltonian Monte Carlo under Constraint Relaxation

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 samples 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, $H(\theta, p) = U(\theta) + K(p)$, with $U(\theta) = -\log \pi(\theta)$, $K(p) = p' M^{-1} p/2$, and $\pi(\theta)$ the unnormalized density in (5) or (6).

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

$$\begin{aligned} \frac{\partial \theta^{(t)}}{\partial t} &= \frac{\partial H(\theta, p)}{\partial p} = M^{-1} p, \\ \frac{\partial p^{(t)}}{\partial t} &= -\frac{\partial H(\theta, p)}{\partial \theta} = -\frac{\partial U(\theta)}{\partial \theta}. \end{aligned} \tag{15}$$

The exact solution to (15) is typically intractable but a valid Metropolis proposal can be generated by

numerically approximating (15) with a reversible and volume-preserving integrator (**?**). 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:

$$p \leftarrow p - \frac{\epsilon}{2}\frac{\partial U}{\partial \theta}, \quad \theta \leftarrow \theta + \epsilon M^{-1} p, \quad p \leftarrow p - \frac{\epsilon}{2}\frac{\partial U}{\partial \theta} \tag{16}$$

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

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. $\exp(-\sum_j \|v_j(\theta^*)\|\lambda_j^{-1})$. Typically, the stability limit of the leapfrog integrator is closely related to the largest eigenvalue $\xi_1(\theta)$ of the Hessian matrix $\mathbf{H}_U(\theta)$ of $U(\theta) = -\log \pi(\theta)$. In fact, 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 (**?**). Under the CORE framework, the

Hessian is given by

$$\mathbf{H}_U(\theta) = -\mathbf{H}_{\log(\mathcal{L}(\theta;y)\pi_{\mathcal{R}}(\theta))}(\theta) + \sum_j \lambda_j^{-1}\mathbf{H}\|v_j(\theta)\|\mathbb{1}_{\theta\notin\mathcal{D}_j}. \tag{17}$$

For $\theta \notin \mathcal{D}_j$, as we make relaxations tighter i.e. $\lambda_j \to 0$, the second term dominates the eigenvalue in the first term and the largest eigenvalue effectively becomes proportional to $\min_{j:\theta\notin\mathcal{D}_j} \lambda_j^{-1}$. In other words, if we think of the Hessian as representing a local covariance structure the target distribution, then the effect of constraints on the stability limit becomes significant roughly speaking when $\min_j \lambda_j^{-1}$ is chosen smaller than the variance of the distribution along $\mathcal{D}$.

The above discussion shows that a choice of extremely small λ_j — corresponding to very tight constraints — could create a computational bottleneck for HMC. Additionally, very tight constraints make it difficult for the no-U-turn criterion of **?** to appropriately calibrate the number of leapfrog steps because the U-turn condition may be met too early to adequately explore the parameter space. For this reason, it is in general best not to make constraints tighter than necessary for computational efficiency. On the other hand, when the leapfrog integrator requires a stepsize $\epsilon \ll \min_j \lambda_j^{-1/2}$ for an accurate approximation, one can safely make the constraint tighter as desired without affecting computational efficiency of HMC.

In our experience, a small number of experiments with different values of λ's were sufficient to find out when the constraint starts to become a bottleneck. Also, HMC usually achieved satisfactory sampling efficiency under reasonably tight constraints. In the appendix, we use a problem of sampling from the von Mises–Fisher distribution to illustrate how a choice of λ affects sampling efficiency.

5 Simulated Examples

CORE enables much greater flexibility for general modeling practice. We now illustrate two interesting case via simulated examples.

Example: Sphere t Distribution

In the first example, we consider modeling on a $(p-1)$-sphere $\mathcal{D} = \mathbb{S}^{p-1}$. Since von Mises–Fisher distribution (**?**) is the result of constraining a multivariate Gaussian $\theta \sim \mathrm{No}(F, I\sigma^2)$ with $F \in \mathcal{D}$ and $v(\theta) = \theta'\theta - 1$.

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

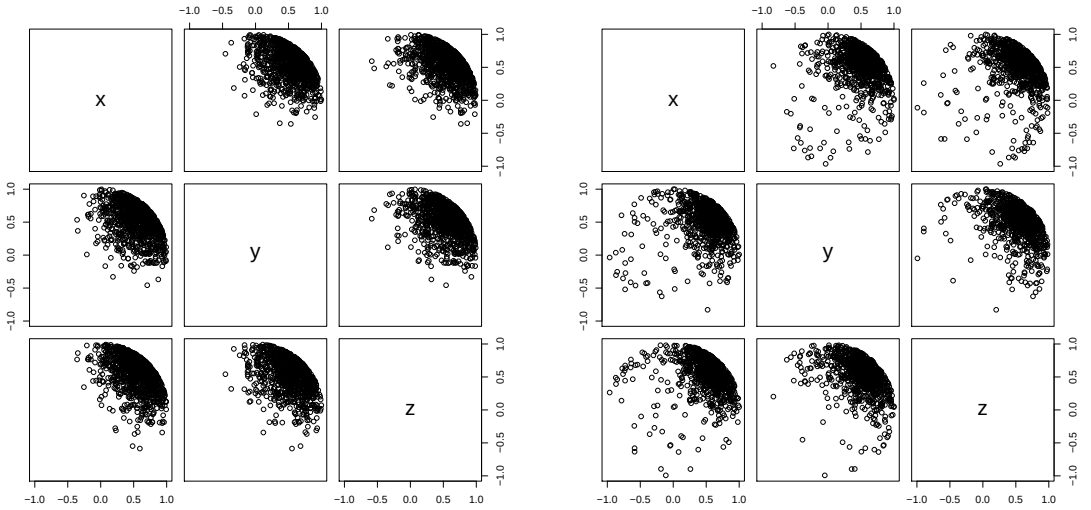
The behavior of this density can be largely explained by the form of constrained Gaussian: θ is symmetrically distributed around F, with density decaying exponentially as the quadratic form of Euclidean distance

$\|\theta - F\|^2$ increases.

CORE allows us to easily consider different 'parent' distribution, instead of Gaussian. Using $v(\theta) = \theta'\theta - 1$ to form a distance in CORE, we consider a t-density

$$\tilde{\pi}_\lambda(\theta) \propto (1 + \frac{\|F - \theta\|^2}{m\sigma^2})^{-\frac{(m+p)}{2}} \exp(-\lambda^{-1}\|\theta'\theta - 1\|)$$

with m degrees of freedom, mean $F \in \mathcal{D}$ and variance $I\sigma^2$. The density decays in polynomial rate as $\|F-\theta\|^2$ increases, as opposed to the exponential decay in Gaussian. Figure 4 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 4: 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

For the second example, we consider relaxing an ordered Dirichlet distribution, with density:

$$\pi_{\mathcal{D}}(\theta) \propto \prod_{j=1}^{J} \theta_j^{\alpha-1} \cdot \mathbb{1}_{\sum_{j=1}^{J} \theta_j = 1} \cdot \prod_{j=1}^{J-1} \mathbb{1}_{\theta_j \geq \theta_{j+1}} \,. \tag{18}$$

One motivation for adding order constraint among $\theta_\cdot$ is to reduce the chance of label-switching during posterior sampling (reviewed in **?**). Obviously, order constraint prevents the switch between large θ_j and small $\theta_{j'}$. Although tractable HMC sampling was proposed for estimation exactly under the simplex constraint

$\sum_{j=1}^{J} \theta_j = 1$ CITE Betancourt, the order constraint would break the algorithm. Therefore, we use CORE to relax the order constraint:

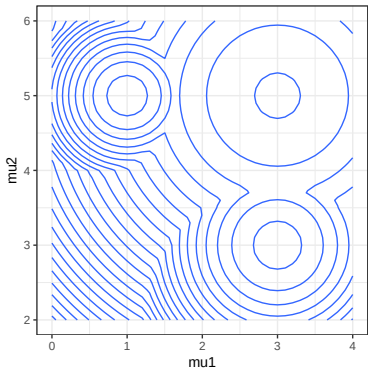
$$\tilde{\pi}_\lambda(\theta) \propto \prod_{j=1}^{J} \theta_j^{\alpha-1} \cdot \mathbb{1}_{\sum_{j=1}^{J} \theta_j = 1} \cdot \prod_{j=1}^{J-1} \exp\Big(- \lambda(\theta_{j+1} - \theta_j, 0)_+ \Big), \tag{19}$$

which can be sampled using simplex HMC CITE Betancourt.

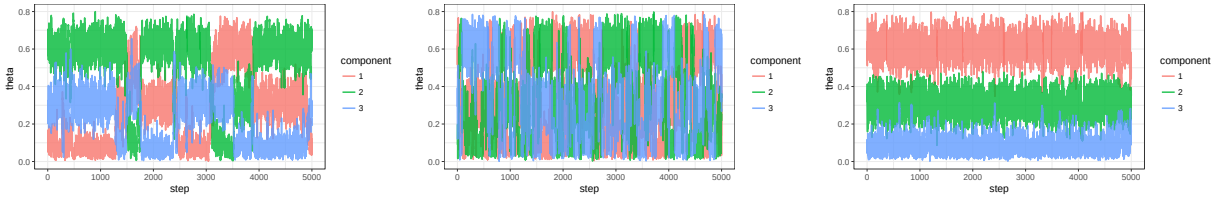
To illustrate the performance, we use the conventional Dirichlet prior without order constraint and the CORE prior described above with $\lambda = 10^{-6}$, as two different priors for a hierarchical normal distribution with a mixture mean, for data $y_i \in \mathbb{R}^2$ indexed by $i = 1, \ldots, n$:

$$y_i \overset{indep}{\sim} \text{No}(\mu_i, \Sigma), \qquad \mu_i \overset{iid}{\sim} G, \qquad G(.) = \sum_{j=1}^{3} \theta_j \delta_{\mu_j}(.),$$

We simulate $n = 100$ 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 5(a) shows the contour of posterior density of μ. The small sample size in each component leads to large overlap in the posterior, creating label-switching problem when using conventional Dirichlet prior (traceplots in Figure 5(b,c) for Gibbs sampling and HMC), whereas Dirichlet with mildly relaxed order constraint has almost no label-switching issue (Figure 5(d)).



(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 Dirichlet with mildly relaxed order constraint

Figure 5: Contour of the posterior density of mixture component means μ_1, μ_2, μ_3 and traceplot of the posterior sample for the component weights $\theta_1, \theta_2, \theta_3$ in a 3-component normal mixture model. Panel (a) shows that there is significant overlap among component means μ_1, μ_2, μ_3. Without ordering in θ, its traceplot shows label-switching issue in both Gibbs (b) and HMC (c) sampling of Dirichlet distribution. While strict ordering makes almost impossible to sample the posterior, slightly relaxed ordering via CORE allows tractable computation and substantially less label-switching (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 (**?**), 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 **?** 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 $U = \{u_{(k,l)}\}_{k=1,\ldots,V;r=1,\ldots,d}$ but let the loadings

$\{v_{(i,r)}\}_{r=1,\ldots,d}$ vary across subjects, specifically

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

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

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

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

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; each $v_{(i,r)} > 0$.

To obtain interpretable estimate of the shared factor U, it is important to have the algorithm converge. Therefore, it is common to have the factor matrix U on a Stiefel manifold $\mathcal{V}(n,d) = \{U : U'U = I_d\}$, in order to remove the free rotation or rescaling of U (**?**). However, this severely limits the choice of prior distribution one could use on U.

Therefore, we use a CORE prior as a relaxation from the Stiefel manifold $\exp(-\lambda^{-1}\|U'U - I\|)$ with $\lambda = 10^{-3}$, creating a near-orthonormal space. This relaxation allows us to consider global-local prior for obtain shrinkage on U, whose non-negligible elements that away from 0 might correspond to important regions of the brain. In this case, we utilize the Dirichlet-Laplace prior for shrinkage (**?**):

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

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

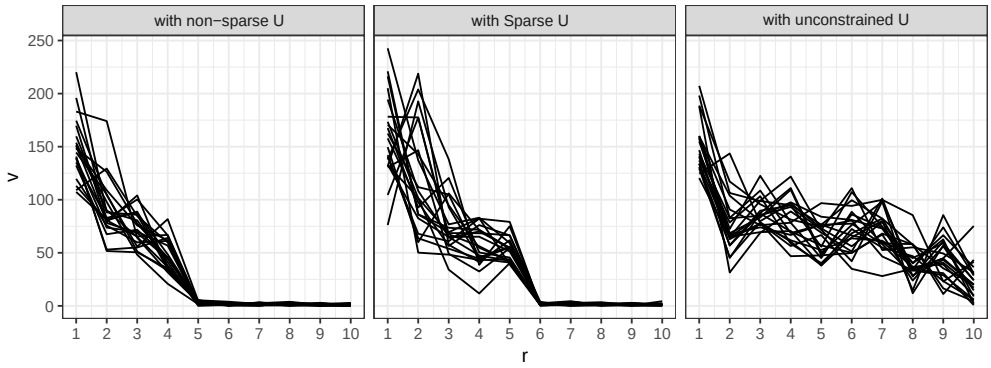
for $k = 1, \ldots, V$, Lap$(0,1)$ denotes the Laplace distribution centered at 0 with scale 1. To induce sparsity in each Dirichlet, we use $\alpha = 0.1$ as suggested by **?**.

We run the described model, along with other two baseline models with no shrinkage but simple normal prior $u_{(k,r)} \sim \text{No}(0,1)$: one under CORE prior for near-orthonormality and one completely unconstrained. 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 $d = 20$.

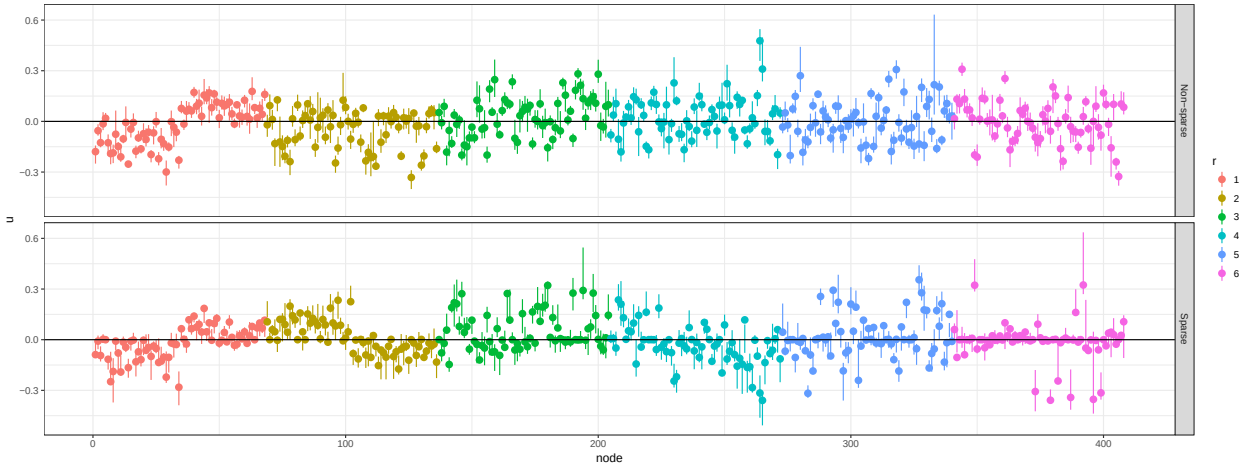
As anticipated, the completely unconstrained model fail to converge; while the other two models successfully converge even though we mildly relax the orthonormality constraint via CORE. Besides improving convergence, those two models have a much faster drop to 0 in loadings, compared to the unconstrained

model (Figure 6(a)), which indicates that near-orthogonal factors are a more efficient representation of the span of the latent space.

To compare the two models using CORE, Figure 6(b) plots the top 6 factors U_r under the normal and shrinkage priors. 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 6(a)), model under shrinkage prior detects more variability in the subject-specific loadings (represented by each line), especially over the second factor.



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



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

Figure 6: Factors and loadings estimates of the network models. Panel (a) compares the varying loadings of the subjects in three models. Panel (b) shows that shrinkage model shows difference starting from the second factor (model with unconstrained U is omitted due to non-convergence in the factor);

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 reserved rescans of the same subjects (prediction AUC). Table 2 lists the benchmark results. The two models under CORE show much better performance, especially in prediction. Although we do not see a

clear improved prediction by further using shrinkage prior, the sparse factors can be more useful for scientific interpretation.

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%

Table 2: Benchmark of 3 models for 21 brain networks. Models with near-orthonormality show much better performance in AUC.

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.

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

$$\begin{aligned}&\int_{\mathcal{R}} g(\theta)\mathcal{L}(\theta;Y)\pi_{\mathcal{R}}(\theta)\exp(-\|v_{\mathcal{D}}(\theta)\|/\lambda)d\mu_{\mathcal{R}}(\theta)\\ &=\int_{\mathcal{R}\setminus\mathcal{D}} g(\theta)\mathcal{L}(\theta;Y)\pi_{\mathcal{R}}(\theta)\exp(-\|v_{\mathcal{D}}(\theta)\|/\lambda)d\mu_{\mathcal{R}}(\theta)+\int_{\mathcal{D}} g(\theta)\mathcal{L}(\theta;Y)\pi_{\mathcal{R}}(\theta)\exp(-\|v_{\mathcal{D}}(\theta)\|/\lambda)d\mu_{\mathcal{R}}(\theta).\end{aligned} \tag{20}$$

For brevity, we let $f(\theta) = \mathcal{L}(\theta;Y)\pi_{\mathcal{R}}(\theta)$ and use $df(\theta) = \mathcal{L}(\theta;Y)\pi_{\mathcal{R}}(\theta)d\mu_{\mathcal{R}}(\theta)$ throughout the proof. Then,

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

where the second equality follows from combining the fractions and making use of (20). We can bound the denominator from below by $C_{\mathcal{D}}^2 = \left[\int_{\mathcal{D}} \mathcal{L}(\theta;Y)\pi_{\mathcal{R}}(\theta)d\mu_{\mathcal{R}}(\theta)\right]^2 > 0$ so that

$$\begin{aligned}
&\left|\mathbb{E}[g(\theta)|\theta\in\mathcal{D}] - \mathbb{E}_{\tilde{\pi}_\lambda}[g(\theta)]\right| \\
&\leq \frac{\left|\int_{\mathcal{R}\setminus\mathcal{D}} \exp(-\|v_{\mathcal{D}}(\theta)\|/\lambda)df(\theta)\cdot\int_{\mathcal{D}} g(\theta)df(\theta) - \int_{\mathcal{D}} df(\theta)\cdot\int_{\mathcal{R}\setminus\mathcal{D}} g(\theta)\exp(-\|v_{\mathcal{D}}(\theta)\|/\lambda)df(\theta)\right|}{C_{\mathcal{D}}^2}
\end{aligned}$$

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,

$$\begin{aligned}
&\left|\mathbb{E}[g(\theta)|\theta\in\mathcal{D}] - \mathbb{E}_{\tilde{\pi}_\lambda}[g(\theta)]\right| \\
&\leq \frac{\left|\int_{\mathcal{R}\setminus\mathcal{D}} \exp(-\|v_{\mathcal{D}}(\theta)\|/\lambda)df(\theta)\right|\cdot\left|\int_{\mathcal{D}} g(\theta)df(\theta) - \int_{\mathcal{R}\setminus\mathcal{D}} g(\theta)\exp(-\|v_{\mathcal{D}}(\theta)\|/\lambda)df(\theta)\right|}{C_{\mathcal{D}}^2} \\
&\qquad + \frac{\left|\int_{\mathcal{R}\setminus\mathcal{D}} g(\theta)\exp(-\|v_{\mathcal{D}}(\theta)\|/\lambda)df(\theta)\right|\cdot\left|\int_{\mathcal{D}} df(\theta) - \int_{\mathcal{R}\setminus\mathcal{D}} \exp(-\|v_{\mathcal{D}}(\theta)\|/\lambda)df(\theta)\right|}{C_{\mathcal{D}}^2}
\end{aligned}$$

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,

$$\begin{aligned}
&\left|\int_{\mathcal{R}\setminus\mathcal{D}} \exp(-\|v_{\mathcal{D}}(\theta)\|/\lambda)df(\theta)\right|\cdot\left|\int_{\mathcal{D}} g(\theta)df(\theta) - \int_{\mathcal{R}\setminus\mathcal{D}} g(\theta)\exp(-\|v_{\mathcal{D}}(\theta)\|/\lambda)df(\theta)\right| \\
&\leq \left|\int_{\mathcal{R}\setminus\mathcal{D}} \exp(-\|v_{\mathcal{D}}(\theta)\|/\lambda)df(\theta)\right|\cdot\left(\left|\int_{\mathcal{D}} g(\theta)df(\theta)\right| + \left|\int_{\mathcal{R}\setminus\mathcal{D}} g(\theta)\exp(-\|v_{\mathcal{D}}(\theta)\|/\lambda)df(\theta)\right|\right) \\
&\leq \int_{\mathcal{R}\setminus\mathcal{D}} \exp(-\|v_{\mathcal{D}}(\theta)\|/\lambda)df(\theta)\cdot\left(\int_{\mathcal{D}} |g(\theta)|df(\theta) + \int_{\mathcal{R}\setminus\mathcal{D}} |g(\theta)|\exp(-\|v_{\mathcal{D}}(\theta)\|/\lambda)df(\theta)\right) \\
&\leq \int_{\mathcal{R}\setminus\mathcal{D}} \exp(-\|v_{\mathcal{D}}(\theta)\|/\lambda)df(\theta)\cdot\int_{\mathcal{R}} |g(\theta)|df(\theta) = C_{\mathcal{R}}\mathbb{E}|g(x_i)|\int_{\mathcal{R}\setminus\mathcal{D}} \exp(-\|v_{\mathcal{D}}(\theta)\|/\lambda)df(\theta).
\end{aligned}$$

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 $\mathcal{L}(\theta; Y)\pi_{\mathcal{R}}(\theta)$. Secondly,

$$\begin{aligned}
&\left| \int_{\mathcal{R}\setminus\mathcal{D}} g(\theta) \exp(-\|v_{\mathcal{D}}(\theta)\|/\lambda) df(\theta) \right| \cdot \left| \int_{\mathcal{D}} df(\theta) - \int_{\mathcal{R}\setminus\mathcal{D}} \exp(-\|v_{\mathcal{D}}(\theta)\|/\lambda) df(\theta) \right| \\
&\leq \int_{\mathcal{R}\setminus\mathcal{D}} |g(\theta)| \exp(-\|v_{\mathcal{D}}(\theta)\|/\lambda) df(\theta) \cdot \left| \int_{\mathcal{D}} df(\theta) \right| + \left| \int_{\mathcal{R}\setminus\mathcal{D}} \exp(-\|v_{\mathcal{D}}(\theta)\|/\lambda) df(\theta) \right| \\
&\leq \int_{\mathcal{R}\setminus\mathcal{D}} |g(\theta)| \exp(-\|v_{\mathcal{D}}(\theta)\|/\lambda) df(\theta) \left(\int_{\mathcal{D}} df(\theta) + \int_{\mathcal{R}\setminus\mathcal{D}} df(\theta) \right) \\
&= C_{\mathcal{R}} \int_{\mathcal{R}\setminus\mathcal{D}} |g(\theta)| \exp(-\|v_{\mathcal{D}}(\theta)\|/\lambda) df(\theta).
\end{aligned}$$

Thus, we have the bounds specified by the theorem,

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

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 $(E|g(x_i)|+|g(x_i)|)\mathcal{L}(\theta; Y)\pi_{\mathcal{R}}(\theta)$ is a dominating function of $(E|g(\theta)|+|g(\theta)|)\mathcal{L}(\theta; Y)\pi_{\mathcal{R}}(\theta) \exp(-\|v_{\mathcal{D}}(\theta)\|/\lambda)$ 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.

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

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

$$\begin{aligned}
&C_{\mathcal{R}} \int_{\mathcal{R}\setminus\mathcal{D}} (\mathbb{E}|g(\theta)| + |g(\theta)|) \exp(-\|v_{\mathcal{D}}(\theta)\|/\lambda) df(\theta) \\
&\leq C_{\mathcal{R}} \left(\int_{\mathcal{R}\setminus\mathcal{D}} (\mathbb{E}|g(\theta)| + |g(\theta)|)^2 df(\theta) \right)^{1/2} \left(\int_{\mathcal{R}\setminus\mathcal{D}} \exp(-2\|v_{\mathcal{D}}(\theta)\|/\lambda) df(\theta) \right)^{1/2}
\end{aligned}$$

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

$$C_{\mathcal{R}} \int_{\mathcal{R}\setminus\mathcal{D}} (\mathbb{E}|g(\theta)| + |g(\theta)|) \exp(-\|v_{\mathcal{D}}(\theta)\|/\lambda) df(\theta)$$

$$= \underbrace{\left(3C_{\mathcal{R}}^2(\mathbb{E}|g|)^2 + C_{\mathcal{R}}\mathbb{E}[|g|^2]\right)^{1/2}}_{C_g} \left(\exp(-2\|v_{\mathcal{D}}(\theta)\|/\lambda)\right)^{1/2} = C_g\left(\int_{\mathcal{R}\setminus\mathcal{D}} \exp(-2\|v_{\mathcal{D}}(\theta)\|/\lambda) df(\theta)\right)^{1/2}$$

We separate the integral

$$\int_{\mathcal{R}\setminus\mathcal{D}} \exp(-2\|v_{\mathcal{D}}(\theta)\|/\lambda) df(\theta)$$

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

$$\int_{\mathcal{R}\setminus\mathcal{D}} \exp(-2\|v_{\mathcal{D}}(\theta)\|/\lambda) df(\theta)$$

$$= \int_{\{\theta:\|v_{\mathcal{D}}(\theta)\|>-\lambda\log\lambda\}} \exp(-2\|v_{\mathcal{D}}(\theta)\|/\lambda) df(\theta) + \int_{\{\theta:\,0<\|v_{\mathcal{D}}(\theta)\|<-\lambda\log\lambda\}} \exp(-2\|v_{\mathcal{D}}(\theta)\|/\lambda) df(\theta)$$

$$\leq \lambda^2 \int_{\{\theta:\|v_{\mathcal{D}}(\theta)\|>-\lambda\log\lambda\}} df(\theta) + \int_{\{\theta:\,0<\|v_{\mathcal{D}}(\theta)\|<-\lambda\log\lambda\}} \exp(-2\|v_{\mathcal{D}}(\theta)\|/\lambda) df(\theta)$$

$$\leq C_{\mathcal{R}}\lambda^2 + \int_{\{\theta:\,0<\|v_{\mathcal{D}}(\theta)\|<-\lambda\log\lambda\}} \exp(-2\|v_{\mathcal{D}}(\theta)\|/\lambda) df(\theta)$$

To review, to this point we have shown that

$$\left|\mathbb{E}[g(\theta)|\theta \in \mathcal{D}] - \mathbb{E}_{\tilde{\pi}_\lambda}[g(\theta)]\right| \leq \frac{C_g}{C_{\mathcal{D}}^2}\left(C_{\mathcal{R}}\lambda^2 + \int_{\{\theta:\,0<\|v_{\mathcal{D}}(\theta)\|<-\lambda\log\lambda\}} \exp(-2\|v_{\mathcal{D}}(\theta)\|/\lambda) df(\theta)\right)^{1/2} \tag{21}$$

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

$$\int_{V_\lambda} \exp(-2\|v_{\mathcal{D}}(\theta)\|/\lambda) df(\theta) \leq \sup_{V_\lambda} |\mathcal{L}(\theta; Y)\pi_{\mathcal{R}}(\theta)| \int_{V\lambda} \exp(-2\|v_{\mathcal{D}}(\theta)\|/\lambda) d\mu_R(\theta)$$

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

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

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 $\mathcal{R}$. 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 **??**

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

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,

$$\begin{aligned}
&\int_{\{\theta:\, 0<\|v_{\mathcal{D}}(\theta)\|<-\lambda\log\lambda\}} \exp(-2\|v_{\mathcal{D}}(\theta)\|/\lambda)\mathcal{L}(\theta; Y)\pi_{\mathcal{R}}(\theta)d\mu_{\mathcal{R}}(\theta) \\
&\leq \sup_{x\in[0,-\lambda\log\lambda]} \left(\int_{v_{\mathcal{D}}^{-1}(x)} \frac{1}{2}\mathcal{L}(\theta; Y)\pi_{\mathcal{R}}(\theta)d\bar{\mathcal{H}}^{r-1}(\theta) \right) \cdot \int_0^{-\lambda\log\lambda} e^{-\frac{x}{\lambda}} dx \\
&= \sup_{x\in[0,-\lambda\log\lambda]} \left(\int_{v_{\mathcal{D}}^{-1}(x)} \frac{1}{2}\mathcal{L}(\theta; Y)\pi_{\mathcal{R}}(\theta)d\bar{\mathcal{H}}^{r-1}(\theta) \right) \cdot (\lambda - \lambda^2) = O(\lambda)
\end{aligned}$$

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

$$\begin{aligned}
&\left|\mathbb{E}[g(\theta)|\theta \in \mathcal{D}] - \mathbb{E}_{\tilde{\pi}_\lambda}[g(\theta)]\right| \\
&\leq \frac{C_g}{C_{\mathcal{D}}^2}\left(C_{\mathcal{R}}\lambda^2 + \sup_{x\in[0,-\lambda\log\lambda]} \left(\int_{v_{\mathcal{D}}^{-1}(x)} \frac{1}{2}\mathcal{L}(\theta; Y)\pi_{\mathcal{R}}(\theta)d\bar{\mathcal{H}}^{r-1}(\theta) \right) \cdot (\lambda - \lambda^2) \right)^{1/2} \\
&= \frac{C_g}{C_{\mathcal{D}}^2} \cdot \sup_{x\in[0,-\lambda\log\lambda]} \left(\int_{v_{\mathcal{D}}^{-1}(x)} \frac{1}{2}\mathcal{L}(\theta; Y)\pi_{\mathcal{R}}(\theta)d\bar{\mathcal{H}}^{r-1}(\theta) \right) \cdot \sqrt{\lambda} + o(\sqrt{\lambda})
\end{aligned}$$

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, it follows that

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

as $\lambda \to 0^+$. □

Proof. Proof of Corollary 1 Note that for any Borel set A,

$$|\Pi(A) - \tilde{\Pi}_\lambda(A)| = |\mathbb{E}[\mathbb{1}_A(\theta)|\theta \in \mathcal{D}] - \mathbb{E}_{\tilde{\pi}_\lambda}[\mathbb{1}_A(\theta)]|$$

to which we may apply the bound from Lemma 1. Therefore,

$$|\Pi(A) - \tilde{\Pi}_\lambda(A)| \leq \frac{\int_{\mathcal{R}\setminus\mathcal{D}}(C_{\mathcal{R}}\mathbb{E}|\mathbb{1}_A(\theta)| + \mathbb{1}_A(\theta))\mathcal{L}(\theta;Y)\pi_{\mathcal{R}}(\theta)\exp(-\|v_{\mathcal{D}}(\theta)\|/\lambda)d\mu_{\mathcal{R}}(\theta)}{\left[\int_{\mathcal{D}}\mathcal{L}(\theta;Y)\pi_{\mathcal{R}}(\theta)d\mu_{\mathcal{R}}(\theta)\right]^2}$$

$$\leq \frac{\int_{\mathcal{R}\setminus\mathcal{D}}(C_{\mathcal{R}}+1)\mathcal{L}(\theta;Y)\pi_{\mathcal{R}}(\theta)\exp(-\|v_{\mathcal{D}}(\theta)\|/\lambda)d\mu_{\mathcal{R}}(\theta)}{\left[\int_{\mathcal{D}}\mathcal{L}(\theta;Y)\pi_{\mathcal{R}}(\theta)d\mu_{\mathcal{R}}(\theta)\right]^2}$$

From Theorem 1, the final statement in the preceding inequality if $O(\sqrt{\lambda})$ and holds for all A. Therefore, we have shown $\sup_{A:Borel}|\Pi(A) - \tilde{\Pi_\lambda}(A)| = O(\sqrt{\lambda})$ for small λ thereby completing the proof.

□

B Proofs from Section 3.2

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

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

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

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

Secondly, we have the relaxed distribution

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

where the normalizing constant is calculated w.r.t. Lesbesgue measure on $\mathcal{R}$, denoted by $\mu_{\mathcal{R}}$,

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

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

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

Let,

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

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

$$\mathbb{E}[g(\theta)|v(\theta) = x] = \frac{1}{m(x)} \int_{v^{-1}(x)} g(\theta) \frac{\pi_{\mathcal{R}}(\theta)\mathcal{L}(\theta;Y)}{J(v(\theta))} d\bar{\mathcal{H}}^{r-s}(\theta). \tag{22}$$

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

$$\mathbb{E}_{\tilde{\Pi}}[g(\theta)] = \int_{\mathbb{R}^s} \frac{m(x)}{m_\lambda} \exp\bigg(- \frac{||x||_1}{\lambda} \bigg) \mathbb{E}\big[g(\theta)|v(\theta) = x\big] d\mu_{\mathbb{R}^s}(x). \tag{23}$$

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.

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

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

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

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,

$$
\begin{aligned}
m_\lambda &= O\bigg(\lambda^{s+1}\bigg) + \int_{B_1(0;\lambda|\log(\lambda^{s+1})|)} [m(0) + o(1)][1 + O(\lambda^s)] d\mu_{\mathbb{R}^s}(x) \\
&= O(\lambda^{s+1}) + [m(0) + o(1)][1 + O(\lambda^s)] \underbrace{\frac{|2(s+1)\lambda\log\lambda|^s}{\Gamma(s+1)}}_{Vol(B_1(0;\lambda|\log(\lambda^{s+1})|)} \\
&= m(0)\frac{|2(s+1)\lambda\log\lambda|^s}{\Gamma(s+1)} + o(|\lambda\log\lambda|^s)
\end{aligned}
$$

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

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

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

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 $\mathbb{E}_{\tilde{\Pi}}[g(\theta)]$ 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

$$\begin{aligned}
\mathbb{E}_{\tilde{\Pi}}[g(\theta)] &= \int_{\mathbb{R}^s} \frac{m(x)}{m_\lambda} \exp\left(-\frac{||x||_1}{\lambda}\right) \mathbb{E}[g(\theta)|v(\theta)=x] d\mu_{\mathbb{R}^s}(x) \\
&= \int_{\mathbb{R}^s} \frac{1}{m_\lambda} \exp\left(-\frac{||x||_1}{\lambda}\right) \int_{v^{(-1)}(x)} \frac{\mathcal{L}(\theta;Y)\pi_{\mathcal{R}}(\theta)}{J(v(\theta))} d\bar{\mathcal{H}}^{r-s}(\theta) d\mu_{\mathbb{R}^s}(x) \\
&= \int_{\mathbb{R}^s \setminus B_1(0;\lambda|\log(\lambda^{s+1})|)} \frac{1}{m_\lambda} \exp\left(-\frac{||x||_1}{\lambda}\right) \int_{v^{(-1)}(x)} \frac{\mathcal{L}(\theta;Y)\pi_{\mathcal{R}}(\theta)}{J(v(\theta))} d\bar{\mathcal{H}}^{r-s}(\theta) d\mu_{\mathbb{R}^s}(x) \\
&\quad + \int_{B_1(0;\lambda|\log(\lambda^{s+1})|)} \frac{1}{m_\lambda} \exp\left(-\frac{||x||_1}{\lambda}\right) \int_{v^{(-1)}(x)} \frac{\mathcal{L}(\theta;Y)\pi_{\mathcal{R}}(\theta)}{J(v(\theta))} d\bar{\mathcal{H}}^{r-s}(\theta) d\mu_{\mathbb{R}^s}(x) \\
&= O\left(\frac{\lambda^{s+1}}{m_\lambda}\right) + \int_{B_1} \frac{m(0)+o(1)}{m_\lambda}(1+O(\lambda^s)]\Big(\mathbb{E}[g(\theta)|v(\theta)=0] + o(1)\Big) d\mu_{\mathbb{R}^s}(x) \\
&= O\left(C\mathbb{E}|g|\frac{\lambda}{|\log\lambda|^s}\right) + \mathbb{E}[g(\theta)|\theta\in\mathcal{D}] + o(1).
\end{aligned}$$

And we may conclude that

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

The proof of the corollary follows by changing the $o(1)$ correction within the integrals over $B_1(0;\lambda|\log\lambda^{s+1}|)$ to $O(\lambda|\log\lambda^{s+1}|)$ corrections. As a result, the leading order approximation error is then $O(\lambda|\log\lambda|^{-s})$ as $\lambda \to 0^+$. □

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
	$\lambda = 10^{-3}$	$\lambda = 10^{-4}$	$\lambda = 10^{-5}$	
$\left\lvert \mathbb{E}_{\Pi}[\sum_j \theta_j] - \mathbb{E}_{\widehat{\Pi}}[\sum_j \theta_j] \right\rvert$	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.