Alternate structure to section 2

September 11, 2017

---

2 Motivation and Methods

Suppose θ is an $\mathcal{R}$-valued random variable subject to some constraints which restrict it to a subset $\mathcal{D} \subset \mathcal{R}$. In the Bayesian setting, of principle interest here, θ is a parameter which is known to satisfy some constraints such that it resides in $\mathcal{D}$. In this case, a common approach is to choose a prior distribution with support $\mathcal{D}$. However, aside from some special cases, a suitable choice of prior may be limited. Moreover, sampling θ from the constrained space, when possible, may be difficult or computationally intractable.

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

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

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

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

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

Note that $\mathbb{1}_{\mathcal{D}}(\theta)$ is the pointwise limit of $\exp\big(-\nu_{\mathcal{D}}(\theta)/\lambda\big)$ (except perhaps on the boundary of $\mathcal{D}$) as $\lambda \to 0^+$. However, (1) has support $\mathcal{R}$ for all $\lambda > 0$, hence 'relaxing' the constraint. Since (1) is supported on $\mathcal{R}$ it is more suitable for off-the-shelf MCMC sampling strategies. Ideally, one would hope that samples from (1) could be easily generated and that they would behave as if drawn from the fully conditioned distribution when λ is sufficiently small. We consider this approach when adapted to a number of settings, but generally we refer to it as constraint relaxation (**CORE**).

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

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

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

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

The remainder of this section is organized as follows. In 2.1, we briefly discuss the construction of the fully constrained distribution $\theta | \theta \in \mathcal{D}$ when $\mathcal{D}$ is measure zero. Additionally, we suggest a general strategy for choosing the relaxed density. Relevant theorems comparing the relaxed and fully constrained distributions are given. In Section 2.2, we consider the simpler case where $\mathcal{D}$ has positive measure. Again, we suggest a general strategy for constructing the relaxed density and supply relevant theorems. For clarity, proofs of the theorems contained in 2.1 and 2.2 are supplied in the appendix. In Section 2.3, we discuss a number of examples which highlight the methods from 2.1 and 2.2.

2.1 CORE for submanifolds

In this setting, we will focus on the case where $\mathcal{D}$ is a measure zero submanifold of $\mathcal{R}$. As such, the construction of the conditional distribution of $\theta | \theta \in \mathcal{D}$ must be handled carefully as one cannot simply renormalize by a factor of $\left[\int_{\mathcal{D}} \mathcal{L}(\theta; Y) \pi_{\mathcal{R}} d\mu_{\mathcal{R}}(\theta) \right]^{-1}$. Instead, one must construct a *regular conditional probability* (r.c.p.) which is consistent with unconstrainted probability density $\pi_{\mathcal{R}}$. A complete definition of the r.c.p. is given in the Appendix. In this section, we develop a framework for the construction of the r.c.p. and attempt to offer some geometric intuition. Prior to discussing the formulation of the constrained density, we begin with a discussion on a few important properties of the constrained space.

We will assume $\mathcal{D}$ can be viewed as the solution set to a system of s equations, $\{\nu_j(\theta) = 0\}_{j=1}^s$, where

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

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

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

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

Henceafter, we refer to the functions $\nu_1, \ldots,_n u_k$ as constraint functions. In this case, if we let $\nu : \mathcal{R} \to \mathbb{R}^s$ be the vector-valued function $\nu(\theta) = [\nu_1(\theta), \ldots, \nu_s(\theta)]^T$, then $\mathcal{D} = \ker(v)$ is a comdimension s submanifold of $\mathcal{R}$ for $\mu_{\mathbb{R}^s}$-a.e. x the range of v. Furthemore, suppose the ambient space $\mathcal{R}$ is $r-$dimensional. Then $\mathcal{D}$ is a $(r-s)-$dimensional submanifold of $\mathcal{R}$, and it is natural to discuss the $(d-s)$-dimensional surface area of $\mathcal{D}$.

While the above requirements may appear to limit the potential constrained spaces which could be used, we note that many standard constraints can be formulated in the above framework. For example, when $\mathcal{D}$ is the probability simplex this construction is straightforward. However, a choice of constraints for the unit-sphere or the Stiefel manifold may be less clear.

As working example, suppose $\mathcal{R} = \mathbb{R}^3$ and $\mathcal{D}$ is the great circle where $||\theta||_2^2 = 1$ in the $\theta_1 = 0$ plane. Note that this great circle is a one-dimensional submanifold of $\mathbb{R}^3$ and the one-dimensional 'surface area' of $\mathcal{D}$ is the circumference of the circle, 2π. In this case, we must specify two constraints. A natural choice is

$$\nu_1(\theta) = \theta_1^2 + \theta_2^2 + \theta_3^2 - 1$$
$$\nu_2(\theta) = \theta_1$$

While these equations satisfy requirement (b)-(d), clearly ν_1 is not Lipschitz. However, one can replace ν_1 with $g \circ \nu_1$ for a Lipschitz $g : \mathbb{R} \to \mathbb{R}$ with bounded range such that $g^{(-1)}(0) = \{0\}$. For example, the constraints

$$\nu_1(\theta) = \arctan(\theta_1^2 + \theta_2^2 + \theta_3^2 - 1)$$
$$\nu_2(\theta) = \theta_1$$

are Lipschitz continuous, satisfy (b)-(d), and still identify $\mathcal{D}$ as the constrained space. This strategy can be generalized to other cases including Stiefel manifold.

As mentioned above, under criteria (a)-(d) the constrained space is a $(r-s)$-dimensional submanifold so that $\mu_{\mathcal{R}}(\mathcal{D}) = 0$. However, the normalized $(r-s)$-dimensional Hausdorff measure, $\bar{\mathcal{H}}^{r-s}$, of $\mathcal{D}$ is non-zero. A more detailed discussion the Hausdorff measure is contained in the Appendix. For the purposes here, it is sufficient to remember that the $(r-s)$-dimensional Hausdorff measure aligns with the usual interpretation surface area, length, etc. of $\mathcal{D}$. In fact, if $\mathcal{D}$ is contained in some compact subset of $\mathcal{R}$, then $\bar{\mathcal{H}}^{r-s}(\mathcal{D})$ is the $(r-s)$-dimensional area of $\mathcal{D}$.

In the working example above, $\bar{\mathcal{H}}^1(\mathcal{D}) = 2\pi$. More generally, for $\nu(\theta) = [\arctan(||\theta||_2^2 - 1), \theta_1]^T$, it follows that $\bar{\mathcal{H}}^1(\nu^{(-1)}(a, 0)) = 2\pi\sqrt{1 + \tan a}$ for $a > -\pi/4$. As an additional example, if $\mathcal{D} = \{\theta \in \mathcal{R}^r : ||\theta||_2^2 = 1\}$, then $\bar{\mathcal{H}}(\mathcal{D})$ is the surface of the unit sphere, $\mathbb{S}^{r-1}$.

Theorem 1. *Assume that $J(v(\theta)) > 0$ and that there is a finite non-negative integer p such that, for $z \in \mathbb{R}^s$*

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

then

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

is a valid regular conditional probability for $\theta \in \mathcal{D}$.

By construction, $\{\theta : v(\theta) = z\}$ is a $r - s$ dimensional submanifold of $\mathcal{R}$ for $\mu_{\mathbb{R}^s}$-a.e. z in the range of ν. As such, it follows that one should take $p = r - s$. To reiterate, when $\mathcal{D}$ is contained in a compact subset of $\mathcal{R}$, which is the case for the unit circle and the Stiefel manifold, it follows that $m^{r-s}(z) \in (0, \infty)$ for $\mu_{\mathbb{R}^s}$-a.e. z in the range of ν.

It is possible that $m^p(z) \in \{0, \infty\}$ for $p = 1, \ldots, r - 1$. If $\mathcal{D}$ is an unbounded subset of $\mathcal{R}$ and $\frac{\mathcal{L}(\theta;Y)\pi_{\mathcal{R}}(\theta)\mathbb{1}_{v(\theta)=z}}{J(v(\theta))}$ decays sufficiently slowly, then $m^{r-s}(z) = \infty$ for $\mu_{\mathbb{R}^s}$-a.e. z in the range of ν. See Diaconis et al. (2013) for additional discussion of this issue. However, in most practical applications we can use

$$\pi_{\mathcal{D}}(\theta|\theta \in \mathcal{D}, Y) = \frac{1}{m^{r-s}(\mathbf{0})} \frac{\mathcal{L}(\theta; Y)\pi_{\mathcal{R}}(\theta)\mathbb{1}_{v(\theta)=\mathbf{0}}}{J(v(\theta))} \tag{2}$$

as the fully constrained posterior density for θ with data Y and θ in the measure zero subset $\mathcal{D}$. This density is absolutely continuous with respect to the $r - s$ dimensional Hausdorff measure on $\mathcal{D}$ in the sense that

$$P(\theta \in F|\theta \in \mathcal{D}, Y) = \int_F \frac{1}{m^{r-s}(\mathbf{0})} \frac{\mathcal{L}(\theta; Y)\pi_{\mathcal{R}}(\theta)\mathbb{1}_{v(\theta)=\mathbf{0}}}{J(v(\theta))} d\bar{\mathcal{H}}^{r-s}(\theta)$$

for all measureable sets $\mathcal{F}$.

A proof of Theorem 1, omitted in this section, is contained in the Appendix. It follows the approach from Diaconis et al. (2013). We now turn to the construction of the relaxed density.

Similar to the motivating example given initially, we seek to relax the indicator function $\mathbb{1}_{\mathcal{D}}(\theta) = \mathbb{1}_{\nu(\theta)=\mathbf{0}}$ to a function with support on unconstrained space. We propose the approximate, relaxed density

$$\tilde{\pi}_\lambda(\theta|Y) \propto \frac{\mathcal{L}(\theta; Y)\pi_{\mathcal{R}}(\theta) \exp(-\frac{1}{\lambda}\sum_{k=1}^s |\nu_k(\theta)|^2)}{J(v(\theta))}. \tag{3}$$

so that $\pi_{\mathcal{D}}$ is the point-wise limit of $\tilde{\pi}_\lambda$ for $\mu_{\mathcal{R}}$-a.e. $\theta \in \mathcal{R}$ as $\lambda \to 0^+$. However, $\tilde{\pi}_\lambda$ is a density with respect to $\mu_{\mathcal{R}}$ unlike $\pi_{\mathcal{D}}$. Observe that one is free to rescale each of the constraints, $\nu_k \mapsto \lambda_k \nu_k$, without breaking the Lipschitz constraint thereby enabling one to tune the strength of each of the s constraints.

We have elected to square the constraint functions for two reasons. First, it ensures that $\exp(-\frac{1}{\lambda}\sum_{k=1}^s |\nu_k(\theta)|^2)$ is strictly less than one for all $\theta \notin \mathcal{D}$. Furthermore, $\exp(-\frac{1}{\lambda}\sum_{k=1}^s |\nu_k(\theta)|^2)$ will be differentiable in θ for linear functions ν_k, which is a necessary requirement to sample using Hamiltonian Monte Carlo (see Section 4.5).

******** I STILL NEED TO REVIEW TECHNICAL DETAILS OF THEOREMS ********
******** MUST ADD DEFINITION/DISCUSSION OF 1-WASSERSTEIN DISTANCE ********

Theorem 1. *The 1-Wasserstein distance, $W(\pi_{\mathcal{D}}, \tilde{\pi}_\lambda)$, of the measures with densities given by* **??** *and* **??**, *satisfies the bound*

$$W(\pi_{\mathcal{D}}, \tilde{\pi}_\lambda) \leq \lambda \frac{k_1}{m(0)}\left(1 + \frac{k_3}{m(0)}\right) + \exp(-\lambda t)\left(\frac{k_1}{m^2(0)} + \frac{k_2}{m(0)}\right)$$

where t is the radius of a ball in $\mathbb{R}^s$.

Note that $W(\pi_{\mathcal{D}}, \tilde{\pi}_\lambda) \to 0$ as $\lambda \to 0^+$.

2.2 CORE for positive measure subsets

Suppose now that $\mathcal{D}$ has positive measure. Then the constrained posterior density, $\pi_{\mathcal{D}}$, for $\theta|\theta \in \mathcal{D}, Y$ is

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

Unlike the previous section, the constrained density is absolutely continuous with respect to $\mu_{\mathcal{R}}$.

Suppose we approximate $\pi_{\mathcal{D}}$ with an approximate, relaxed density

$$\tilde{\pi}_\lambda(z) = \mathcal{L}(\theta; Y)\frac{\pi_{\mathcal{R}}(z)\exp\big(-\frac{v(z)}{\lambda}\big)}{\int_{\mathcal{R}} \mathcal{L}(\theta; Y)\pi_{\mathcal{R}}(\theta)\exp\big(-\frac{v(\theta)}{\lambda}\big)d\theta} \propto \mathcal{L}(\theta; Y)\pi_{\mathcal{R}}(z)\exp\big(-\frac{v(z)}{\lambda}\big)$$

which is also absolutely continuous with respect to $\mu_{\mathcal{R}}$. Here $\lambda > 0$ and $v(\theta)$ measures the distance from θ to the constrained space $\mathcal{D}$, i.e. $v(\theta) = 0\ \forall z \in \mathcal{D}$ and is positive otherwise. Formally, as $\lambda \to 0^+$, $\exp(-v(\theta)/\lambda) \to \mathbb{1}_{\mathcal{D}}(\theta)$ pointwise. If $\mathcal{D}$ is an open subset of $\mathcal{R}$, this limit may not hold on the boundary of $\mathcal{D}$, denoted $\partial\mathcal{D}$. However, in general $\mu_{\mathcal{R}}(\partial\mathcal{D}) = 0$ and we are working with densities. Thus, we can ignore this issue.

There are many possible choices for v which can be selected for different reasons. Perhaps the simplest choice is to take

$$v(z) = \inf_{x \in \mathcal{D}} ||z - x||_k.$$

where $||\cdot||_k$ denotes the distance using the k-norm. Under this choice of v, the relaxation is isotrophic. More generally, one could use

$$v(z) = \inf_{x \in \mathcal{D}} [(x - z)^T A(x - z)] \tag{4}$$

for some positive definite matrix A. In this case, the relaxation is anistrophic, and can be viewed as a form of directional 'tempering.' This choice of distance, v, allows for a more detailed specification of the rates at which individual components of θ relax to $\mathcal{D}$.

For most general choices of $v(\theta)$ it follows that $\pi_{\mathcal{D}}$ is the pointwise limit of $\tilde{\pi}$ for $\mu_{\mathcal{R}}$ a.e. θ in $\mathcal{R}$. Furthermore, since both the constrained density, $\pi_{\mathcal{D}}$, and the relaxed density, $\tilde{\pi}$, are absolutely continuous with respect to $\mu_{\mathcal{R}}$, estimates of $E[g(\theta)|\theta \in \mathcal{D}]$ using the relaxed density are stronger.

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

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

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

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

$$\lim_{\lambda \to 0^+} \left| E[g(\theta)|\theta \in \mathcal{D}] - \tilde{E}[g(\theta)] \right| = 0.$$

This corollary follows immediately from Theorem 1 and the dominated convergence theorem. Here we use $(E|g(\theta)| + |g(\theta)|)\pi_{\mathcal{R}}(\theta)$ as the dominating function since $g \in \mathbb{L}^1(\mathcal{R}, \pi_{\mathcal{R}}d\mu_{\mathcal{R}})$ by assumption and $v(\theta) > 0$ for all θ.

Theorem 1 and Corollary 1 have some important implications both analytically and numerically. First, although the requirement that $\mathcal{D}$ has positive measure is much stronger than that considered in the previous section, one can use the relaxed density to approximate $E[g(\theta)|\theta \in \mathcal{D}]$ for a much larger class of functions than Lipschitz-1 functions only. In particular, in addition to point estimates, $E[\theta|\theta \in \mathcal{D}]$, it is possible to approximate probabilities $P(\theta \in \mathcal{F}|\theta \in \mathcal{D})$ and higher moments, e.g. $E[\Pi_j \theta_j^{k_j}|\theta \in \mathcal{D}]$, so long as these moments exist for the unconstrained density $\pi_{\mathcal{R}}$.

Secondly, these bounds demonstrate that the error in using the relaxed density to approximate $E[g(\theta)|\theta \in \mathcal{D}]$ is proportional to $\left[\int_{\mathcal{D}} \mathcal{L}(\theta; Y)\pi_{\mathcal{R}}(\theta)d\mu_{\mathcal{R}}(\theta)\right]^{-2}$. Therefore, in practice λ may need to be very small, particularly in the case where $0 < P(\theta \in \mathcal{D}) \ll 1$. Of course, specific details of the scaling of $\left|E[g(\theta)|\theta \in \mathcal{D}] - \tilde{E}[g(\theta)]\right|$ will depend upon the choice of $v(\theta)$. As such, one avenue for mitigating numerical difficulties which may arise when $\lambda \ll 1$ is to use Eq. 4 to relax the density in directions where accuracy is less important.

2.3 Examples

A Definitions and Proofs for Section 2.1

Definition 1. *Let* $(\Omega, \mathcal{F}, P)$ *be a probability spaces and* $\mathcal{C} \subseteq \mathcal{F}$ *a sub-sigma algebra. A function* $K(w, dw)$ *from* $(\Omega \times \mathcal{F})$ *to* $[0, 1]$ *is a regular conditional probability for* P *given* $\mathcal{C}$ *if*

(i) For each $w \in \Omega$, $K(w, \cdot)$ *is a probability measure on* $\mathcal{F}$.

(ii) For each $F \in \mathcal{F}$, *the function* $w \mapsto K(w, F)$ *is* $\mathcal{C}$ *measurable.*

(iii) For each $C \in \mathcal{C}$, $F \in \mathcal{F}$, $P(C \cap F) = \int_C K(w, F)P(dw)$.

B Proofs for Section 2.2

Proof. Proof of Theorem 1 asdf □