Constraint Relaxation for Bayesian Modeling with Parameter Constraints

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

September 20, 2017

2 Motivation and Methods

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

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

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

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

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

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

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

would hope that samples from (1) could be easily generated and that they would behave as if drawn from the fully conditioned distribution when λ is sufficiently small. We consider this approach when adapted to a number of settings, but generally we refer to it as constraint relaxation (**CORE**).

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

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

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

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

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 (based on density proportional to $\mathcal{L}(\theta; Y)\pi_{\mathcal{R}}(\theta)$). 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 subsection, we will focus on the case where $\mathcal{D}$ is a measure zero submanifold of $\mathcal{R}$, ($\int_{\mathcal{D}} \mathcal{L}(\theta; Y)\pi_{\mathcal{R}} d\mu_{\mathcal{R}}(\theta) = 0$). 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 (changed 'one must' to 'we propose to') construct a *regular conditional probability* (r.c.p.) which is consistent with unconstrained probability density $\pi_{\mathcal{R}}$. A complete definition of the r.c.p. is given in the appendix. (briefly explain how r.c.p differs from conventional conditional probability) In this section, we develop a framework for the construction of the r.c.p. and attempt to offer some geometric intuition. Prior to formulating 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 defined implicitly as the solution set to a system of s equations, $\{\nu_j(\theta) = 0\}_{j=1}^{s}$, where

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

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

(c) for $k = 1, \ldots, s$, 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) $\nu_j^{(-1)}(0)$ and $\nu_k^{(-1)}(0)$ intersect transversally for $1 \leq j < k \leq s$.

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

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

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

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

In the more difficult situation where equality constraints are not given, finding $\{\nu_j\}_{j=1}^s$ may be very difficult. For the moment, we will assume that one can construct sufficiently accurate numerical approximations perhaps through cubic-splines or Fourier series, so that we may ignore the issue of existence.

With regards to uniqueness, unfortunately, we note that the constraints cannot be unique in any case. For example, given any constraints $\{\nu_j\}_{j=1}^s$ which satisfy (a)-(d) and non-zero constants $\{\lambda_j\}_{j=1}^s$, the constraints $\{\lambda_j \nu_j\}_{j=1}^s$ will also satisfy (a)-(d). It is then natural to wonder if one can find an optimal choice of the constraints. An optimal choice will depend largely on the properties of the constrained distribution that one wishes to estimate making the choice of $\{\nu_j\}_{j=1}^s$ context dependent. As such, we will address this issue in the examples contained in later sections. For now, let us proceed assuming that a suitable choice of $\{\nu_j(\theta)\}_{j=1}^s$ has been made.

To construct a density constrained to $\mathcal{D}$, we will make use of the normalized $(r-s)$-dimensional Hausdorff measure, $\bar{\mathcal{H}}^{r-s}$, ('a standard tool in geometric measure theory.') A more detailed discussion of the Hausdorff measure is contained in the appendix. For the purposes here, it is sufficient to remember that $\bar{\mathcal{H}}^{r-s}$ coincides with the usual interpretation of surface area, length, etc. of $\mathcal{D}$. In fact, if $\mathcal{D}$ is a smooth, compact submanifold of $\mathcal{R}$, then $\bar{\mathcal{H}}^{r-s}(\mathcal{D})$ is the $(r-s)$-dimensional surface area of $\mathcal{D}$.

Let $D(\nu(\theta))_{i,j} = \frac{\partial \nu_i}{\partial \theta_j}$. The s-dimensional Jacobian, $J(v(\theta)) = \sqrt{\det[(Dv)(Dv)']}$, can be interpreted geometrically as the maximum s-dimensional volume of the image of a unit cube s-dimensional cube in $\mathcal{R}$ under the map $\nu : \mathcal{R} \to \mathbb{R}^s$. Colloquially, $J(\nu(\theta))$ accounts for the change of the $(r-s)$-dimensional area of an infinitesimal set $\Delta\theta \subset \mathcal{D}$ when mapped to its image under ν. Thus, for $\Delta y = \nu(\Delta\theta)$,

$$\frac{1}{J(\nu(\theta))} \bar{\mathcal{H}}^{r-s}(\Delta\theta) \approx \mu_{\mathbb{R}^s}(\Delta y).$$

(Perhaps we don't need to put this interpretation here as it's too complicated, unless we would use

it to derive certain properties later?)

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

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

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

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

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

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

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

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

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

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

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

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 $p_z = r - s$. It is possible that $m^p(z) \in \{0, \infty\}$ for $p = 1, \ldots, r-1$. For example, 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 $m^{r-s}(z) = \infty$. See Diaconis et al. (2013) for additional discussion of this issue.

('By construction, $\mathcal{D} = \{\theta : v(\theta) = \mathbf{0}\}$ is a $(r-s)$ dimensional submanifold of $\mathcal{R}$ for $\mu_{\mathbb{R}^s}$-a.e.. We further limit the considered range of $z = v(\theta)$ to $\{z : m^s(z) \in (0, \infty)\}$')

However, in most practical applications Theorem (1) allows us to define

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

as the fully constrained posterior density as long as $\pi_{\mathcal{R}}$ decays sufficiently fast when $\mathcal{D}$ is an unbounded subset of $\mathcal{R}$ (I'm not sure what this condition is about?). Note that π_D 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 $\mathcal{F} \subset \mathcal{R}$. As a result, we can define the conditional expectation of $g(\theta)$ given $\theta \in \mathcal{D}$ as

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

A proof of Theorem 1, omitted in this section, is contained in the appendix. It follows the approach from Diaconis et al. (2013) and utilizes the co-area formula from Federer (2014). For the moment, we consider 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

$$\tilde{\pi}_\lambda(\theta|Y) \propto \mathcal{L}(\theta;Y)\pi_{\mathcal{R}}(\theta)\exp\left(-\frac{1}{\lambda}||\nu(\theta)||_1\right). \tag{3}$$

(Since the exact density is defined using Hausdorff, perhaps the approximation can be via co-area formula involving Hausdorff, showing it can be converted the Lesbesgue density like this) so that $\tilde{\pi}_\lambda$ converges point-wise to zero for $\mu_{\mathcal{R}}$-a.e. $\theta \in \mathcal{R}$ as $\lambda \to 0^+$. However, since $\tilde{\pi}_\lambda$ is supported on $\mathcal{R}$, it is a density with respect to $\mu_{\mathcal{R}}$ which is an important difference from $\pi_{\mathcal{D}}$.

****** Still finalizing 1-Wasserstein distance proof ******

****** Statement and discussion of convergence results to follow ******

Theorem 1. *The 1-Wasserstein distance, $W(\pi_{\mathcal{D}}, \tilde{\pi}_\lambda)$, of the measures with densities given in Equations (2) and (3), satisfies the bound*

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

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

2.2 CORE for positive measure subsets

In this subsection, we consider the case where $\mathcal{D}$ has positive measure. As such, the constrained posterior density, $\pi_{\mathcal{D}}$, for $\theta|\theta \in \mathcal{D}$ and data 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, this constrained density is absolutely continuous with respect to $\mu_{\mathcal{R}}$.

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

$$\tilde{\pi}_\lambda(\theta) = \mathcal{L}(\theta;Y)\frac{\mathcal{L}(\theta;Y)\pi_{\mathcal{R}}(\theta)\exp\left(-\frac{v_d(\theta)}{\lambda}\right)}{\int_{\mathcal{R}} \mathcal{L}(\theta;Y)\pi_{\mathcal{R}}(\theta)\exp\left(-\frac{v_d(\theta)}{\lambda}\right)d\mu_{\mathcal{R}}(\theta)} \propto \mathcal{L}(\theta;Y)\pi_{\mathcal{R}}(\theta)\exp\left(-\frac{v_d(\theta)}{\lambda}\right)$$

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

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

$$v_d(z) = \inf_{x \in \mathcal{D}} ||z - x||_k \tag{4}$$

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

$$v_d(z) = \inf_{x \in \mathcal{D}} \sqrt{(x - z)^T A (x - z)} \tag{5}$$

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

For most general choices of $v_d(\theta)$ 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 can be applied to larger class of functions.

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

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

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

Corollary 1. *Suppose $g \in \mathbb{L}^2(\mathcal{R}, \pi_{\mathcal{R}} d\mu_{\mathcal{R}})$, $\pi_{\mathcal{D}}$ and $\tilde{\pi}_\lambda$ are as above, and $v_d(\theta)$ has the form of Eq. (4) with $k = 2$. Then for $0 < \lambda \ll 1$,*

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

($\log(\lambda) < 0$*?)*

This corollary follows by applying the Cauchy-Schwartz inequality to the term in the numerator of the bound given in Theorem 1. Some care must be taken if $\mathcal{D}$ is a unbounded subset of $\mathcal{R}$, and these technical details are discussed in the appendix.

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 $v_d(\theta)$. As such, one avenue for mitigating numerical difficulties which may arise when $\lambda \ll 1$ is to use Eq. 5 to relax the density in directions where accuracy is less important.