Distance-to-Set Priors and Constrained Bayesian Inference

Rick Presman
Duke University

Jason Xu
Duke University

Abstract

Constrained learning is prevalent in many statistical tasks. Recent work proposes distance-to-set penalties to derive estimators under general constraints that can be specified as sets, but focuses on obtaining point estimates that do not come with corresponding measures of uncertainty. To remedy this, we approach distance-to-set regularization from a Bayesian lens. We consider a class of smooth *distance-to-set priors*, showing that they yield well-defined posteriors toward quantifying uncertainty for constrained learning problems. We discuss relationships and advantages over prior work on Bayesian constraint relaxation. Moreover, we prove that our approach is optimal in an information geometric-sense for finite penalty parameters ρ, and enjoys favorable statistical properties when $\rho \to \infty$. The method is designed to perform effectively within gradient-based MCMC samplers, as illustrated on a suite of simulated and real data applications.

1 INTRODUCTION

Constrained learning is ubiquitous in statistical tasks when seeking to impose desired structure on solutions. Concretely, consider the task of estimating a parameter $\mathbf{x} \in \mathbb{R}^d$ by minimizing some loss function $f(\mathbf{x})$ where $\mathbf{x}$ needs to satisfy a set of constraints encoded by a set $\mathcal{C}$. Then we seek:

$$\min_{\mathbf{x}} \; f(\mathbf{x}) \quad \text{s.t.} \quad \mathbf{x} \in \mathcal{C} \tag{1}$$

A simple but powerful observation that will make this amenable to effective algorithms is to equivalently express the restriction in terms of the Euclidean distance between the point $\mathbf{x}$ and the constraint set as $\mathsf{dist}(\mathbf{x}, \mathcal{C}) = 0$. In many instances, it is enough to only approximately satisfy the constraints. A recent framework accomplishes this kind of constraint relaxation through distance-to-set penalization (Chi et al., 2014; Xu et al., 2017): for $\rho \in (0, \infty)$,

$$\mathbf{x}^* \in \operatorname{argmin}_{\mathbf{x}} \left[f(x) + \frac{\rho}{2} \mathsf{dist}(\mathbf{x}, \mathcal{C})^2 \right].$$

Solutions to this problem can be obtained using a majorization-minimization (MM) scheme known as the proximal distance algorithm (Keys et al., 2019; Xu and Lange, 2022; Landeros et al., 2022), and it is so called because the iterative updates are defined via proximal operators (Parikh and Boyd, 2014). However, despite its ability to deliver point estimates effectively, it is very difficult to derive measures of uncertainty, and so a general theory of inference is difficult to obtain. Toward filling this methodological gap, we recast the optimization problem in a constrained Bayesian setting by an analog of these penalties that we term *distance-to-set priors*.

Our approach draws previously underexplored connections between this optimization framework and the broader constrained Bayesian inference literature (Ghosh, 1992; Gramacy et al., 2016), an area that continues to grow with exciting recent ideas. We focus on a tradition of sampling through gradient-based samplers such as Hamiltonian Monte Carlo, or HMC (Neal et al., 2011; Betancourt and Girolami, 2015). Lan et al. (2014) utilize a spherical HMC, mapping constraints that can be written as norms onto the hypersphere. Related recent work uses a Riemannian HMC under a manifold setup (Kook et al., 2022), extending a line of work pioneered by Byrne and Girolami (2013). Duan et al. (2020) replace a support constraint with a term that decays exponentially outside of the support, while Sen et al. (2018) project sample draws from unconstrained posteriors to the constraint set to approximate the original posterior. Recently, Xu et al. (2021) propose using priors based on proximal mappings related to the constraint sets. Concurrent work in Zhou et al. (2022) propose using the Moreau-Yosida envelope in a more general manner, using a class of epigraph priors toward regularized and constrained problems suited for proximal MCMC (Pereyra, 2016).

Distance-to-set priors build upon this line of inquiry, providing an effective, practical way to consider constrained inference problems. The framework is more general than many of the previous methods in that it essentially only requires that the constraint can be written as a set, and

Proceedings of the 26th International Conference on Artificial Intelligence and Statistics (AISTATS) 2023, Valencia, Spain. PMLR: Volume 206.

that projection onto that set is feasible. These priors are then easy to evaluate, and work well within gradient-based samplers due to our smooth formulation. This improves computational stability compared to previous approaches using posterior sampling algorithms such as HMC, as we investigate in an empirical study. Moreover from a theoretical perspective, this class of priors admits posteriors that converge in distribution to the original constrained problem along with their maximum a posteriori (MAP) estimates as we increase the parameter governing the degree of constraint enforcement. Finally, we draw a connection between Bayesian constraint relaxation and information geometry, revealing how distance-to-set priors are optimal in a certain sense, while simultaneously yielding a way to select the regularization parameter ρ systematically.

2 THEORY AND METHODS

We begin by briefly reviewing distance-to-set penalties and some of their key properties.

Distance-to-Set Penalties Let $\mathcal{C} \subset \mathbb{R}^n$ be convex, and let $f : \mathcal{C} \to \mathbb{R}$ be a convex function. Many constrained programming problems of the form (1) may be intractable in their original form, but can be converted to a sequence of simpler subproblems. To make progress, denote the Euclidean distance from any $\mathbf{x} \in \mathbb{R}^n$ to $\mathcal{C}$ by $\mathsf{dist}(\mathbf{x}, \mathcal{C}) := \inf_{\mathbf{y}\in\mathcal{C}} \|\mathbf{x} - \mathbf{y}\|_2$: then the condition $\mathbf{x} \in \mathcal{C}$ can be equivalently written as $\mathsf{dist}(x, \mathcal{C}) = 0$. Note that while the distance operator is not necessarily smooth, its square is differentiable as long as the projection of $\mathbf{x}$ onto $\mathcal{C}$, denoted $P_{\mathcal{C}}(\mathbf{x}) := \arg\min_{\mathbf{y}\in\mathcal{C}} \|\mathbf{x} - \mathbf{y}\|_2$, is single-valued (Lange (2016)). Thus, we may reformulate the problem by instead considering the *smooth* unconstrained optimization task:

$$\min_{\mathbf{x}} \left[f(\mathbf{x}) + \frac{\rho}{2}\mathsf{dist}(\mathbf{x}, \mathcal{C})^2 \right],$$

where $\rho > 0$ is a penalty parameter. To solve the resulting problem, Lange (2016) propose a method termed the proximal distance algorithm which makes use of the MM principle to create surrogate functions based on distance majorization (Chi et al., 2014). Its namesake derives from the fact that the minimization of the surrogate functions

$$g_\rho(\mathbf{x} \mid \mathbf{x}_k) = f(\mathbf{x}) + \frac{\rho}{2}\|\mathbf{x} - P_{\mathcal{C}}(\mathbf{x}_k)\|^2$$

is related to the proximal operator of f (Parikh and Boyd (2014)): recall for a function f, the proximal mapping with parameter λ is defined

$$\text{prox}_{\lambda f}(\mathbf{y}) \equiv \underset{\mathbf{x}}{\text{argmin}} \left[f(\mathbf{x}) + \frac{1}{2\lambda}\|\mathbf{x} - \mathbf{y}\|_2^2 \right],$$

which relates to our problem with $\mathbf{y}$ the projection at iterate k and $\lambda = \rho^{-1}$. Under this formulation, to recover the solution to the original optimization problem, it is necessary for $\rho \to \infty$ at some appropriate rate (Wright et al. (1999)). Conversely, fixing a finite ρ results in a solution where $\mathbf{x}$ is close to $\mathcal{C}$, but not strictly inside of the set. Both cases may be of interest depending on the modeling context. We will discuss primarily the latter in this paper but also establish theoretical relationships to the former. As our primary setting is statistical, we may think of $f(\mathbf{x})$ as a convex loss function.

2.1 Distance-to-Set Priors

The proximal distance algorithm mentioned above provides a method for obtaining point estimates under distance-to-set penalization. However, to the best of our knowledge, the current literature does not provide results pertaining to uncertainty quantification for these estimators. As a step toward understanding their uncertainty properties, our first contribution is to link these ideas to a Bayesian constraint relaxation framework. Identifying a penalized estimation problem with a Bayesian problem has been done at least as early as the seminal LASSO paper (Tibshirani, 1996). Consider data $\mathbf{y} \mid \boldsymbol{\theta} \in \mathbb{R}^n$ that has likelihood $L(\boldsymbol{\theta} \mid \mathbf{y})$ and is parameterized by some parameter $\boldsymbol{\theta}$ with prior $\pi(\boldsymbol{\theta})$ that is absolutely continuous with respect to Lebesgue measure, with support $\mathbb{R}^d$ but constrained to $\boldsymbol{\Theta} \subset \mathbb{R}^d$. Since $\boldsymbol{\theta}$ is constrained to $\boldsymbol{\Theta}$, Bayes' Theorem gives the posterior for $\boldsymbol{\theta}$:

$$\begin{aligned}\overline{\pi}(\boldsymbol{\theta} \mid \mathbf{y}) &\propto L(\boldsymbol{\theta} \mid \mathbf{y})\pi(\boldsymbol{\theta})\mathbf{1}_{\boldsymbol{\theta}\in\boldsymbol{\Theta}} \\ &\propto L(\boldsymbol{\theta} \mid \mathbf{y})\pi(\boldsymbol{\theta})\mathbf{1}_{\mathsf{dist}(\boldsymbol{\theta},\boldsymbol{\Theta})=0}\end{aligned}$$

where the second line follows from the discussion of distance-to-set penalties. Since sampling from a posterior sharply constrained on $\boldsymbol{\Theta}$ may be difficult, we can replace the indicator representing the constraint with $\exp\left(-\frac{\rho}{2}\mathsf{dist}(\boldsymbol{\theta}, \boldsymbol{\Theta})^2\right)$. An illustration is given in Figure 1. This term is equal to the indicator on $\boldsymbol{\Theta}$ and rapidly decays to zero as the distance from $\boldsymbol{\theta}$ to $\boldsymbol{\Theta}$ grows larger. We choose to square the distance-to-set operator to align with the distance-to-set optimization and to improve sampling performance, which we discuss in Section 2.4.

We propose to use *distance-to-set priors*, defined as follows:

$$\widetilde{\pi}(\boldsymbol{\theta}) := \pi(\boldsymbol{\theta}) \exp\left(-\frac{\rho}{2}\mathsf{dist}(\boldsymbol{\theta}, \boldsymbol{\Theta})^2\right),$$

where $\rho > 0$ is a hyperparameter in our treatment.

To bridge this with the optimization setting, we can define $f(\boldsymbol{\theta}) := -\log(L(\boldsymbol{\theta} \mid \mathbf{y})\pi(\boldsymbol{\theta}))$. However, it should be noted that although every likelihood function gives us a loss by taking the negative-log, the converse is not always true. Thus, one could generalize our approach by considering more general loss functions and incorporating them into the Bayesian framework via Gibbs posteriors (Bissiri et al., 2016); see also Jacob et al. (2017). Ignoring the constraint for a moment, we obtained the *unconstrained posterior*:

$$\pi(\boldsymbol{\theta} \mid \mathbf{y}) \propto L(\boldsymbol{\theta} \mid \mathbf{y})\pi(\boldsymbol{\theta}). \tag{2}$$

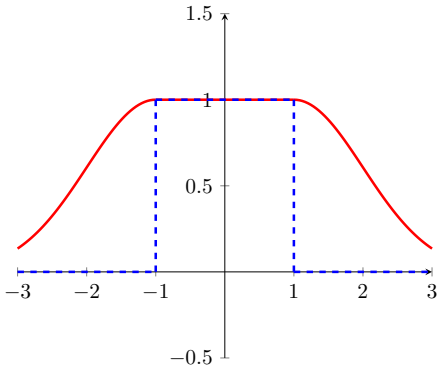


Figure 1: Schematic of distance-to-set prior for the set $\boldsymbol{\Theta} = [-1, 1]$. The blue dashed line represents the indicator $\mathbf{1}_{\boldsymbol{\Theta}}$, and the red solid line represents the relaxation of $\mathbf{1}_{\boldsymbol{\Theta}}$ that we consider in this paper.

Combining this with constraint relaxation, we obtained what we call the *constraint relaxed posterior*:

$$\widetilde{\pi}(\boldsymbol{\theta} \mid \mathbf{y}) \propto L(\boldsymbol{\theta} \mid \mathbf{y})\pi(\boldsymbol{\theta}) \exp\left(-\frac{\rho}{2}\mathsf{dist}(\boldsymbol{\theta}, \boldsymbol{\Theta})^2\right) \tag{3}$$

The relaxed form avoids the discontinuity implied by the indicator function in the sharply *constrained posterior* :

$$\overline{\pi}(\boldsymbol{\theta} \mid \mathbf{y}) \propto L(\boldsymbol{\theta} \mid \mathbf{y})\pi(\boldsymbol{\theta})\mathbf{1}_{\boldsymbol{\theta}\in\boldsymbol{\Theta}} \tag{4}$$

For the remainder of this paper, we make the following assumptions:

Assumption 1. All probability measures are absolutely continuous with respect to d-dimensional Lebesgue measure with densities supported in $\mathbb{R}^d$.

Assumption 2. The unconstrained posterior $\pi(\boldsymbol{\theta} \mid \mathbf{y})$ is proper; that is, $\displaystyle\int_{\mathbb{R}^d} L(\boldsymbol{\theta} \mid \mathbf{y})\pi(\boldsymbol{\theta})\, d\boldsymbol{\theta} < \infty$.

Assumption 3. Unless stated otherwise, the support $\varnothing \neq \boldsymbol{\Theta} \subset \mathbb{R}^d$ is a closed and convex set.

We make Assumption 1 because we will have an interest in the performance of samplers, like HMC, that are designed to perform on continuous distributions and to simplify the setting. Assumption 2 guarantees the original posterior is not ill-posed, and ensures we are sampling from a well-defined distribution. Assumption 3 plays an integral role toward the smoothness properties of the constraint relaxation; they lead to continuity of the projection as well as a unique gradient of the squared distance.

These natural conditions asure that the object of interest is well-defined. The following proposition, as well as all theorems in the following section, are proven in the Supplement.

Proposition 1. *Under Assumptions 1 and 2, the constraint relaxed posterior $\widetilde{\pi}(\boldsymbol{\theta} \mid \mathbf{y})$ (Equation 3) is a proper density.*

2.2 Statistical Properties

Distance-to-set regularization and the underlying constrained problem are inextricably linked, so one would naturally hope that the constraint relaxed posterior behaves approximately like the constrained posterior when ρ is large. Fortunately, this is the case as we formalize in the guarantees below. Our first result shows that this class of distance-to-set priors also possess the desirable property that the sequence of MAP estimators of the relaxed posterior (indexed by the penalty parameter ρ) converge to the the MAP estimator of the non-relaxed problem as ρ grows large when the posterior is log-concave.

Theorem 1. *Suppose the unconstrained posterior $\pi(\boldsymbol{\theta} \mid \mathbf{y})$ (Equation 2) is strictly log-concave. Let $\{\widetilde{\pi}_{\rho_k}(\boldsymbol{\theta} \mid \mathbf{y})\}_{k\in\mathbb{N}}$ (Equation 3) be a sequence of constraint-relaxed posterior distributions where $\rho_k \uparrow \infty$ as $k \to \infty$. Further, define the following MAP estimators*

$$\widehat{\boldsymbol{\theta}}^M = \operatorname{argmax}_{\boldsymbol{\theta}} \overline{\pi}(\boldsymbol{\theta} \mid \mathbf{y}), \quad \widehat{\boldsymbol{\theta}}^M_{\rho_k} = \operatorname{argmax}_{\boldsymbol{\theta}} \widetilde{\pi}_{\rho_k}(\boldsymbol{\theta} \mid \mathbf{y}).$$

Then the sequence $\widehat{\boldsymbol{\theta}}^M_{\rho_k} \to \widehat{\boldsymbol{\theta}}^M$ as $k \to \infty$.

In addition to convergence of a point estimate, we can say more about the behavior of the entire distribution.

Theorem 2. *Let $\overline{\Pi}$ be the constrained posterior distribution with density $\overline{\pi}(\boldsymbol{\theta} \mid \mathbf{y})$, and let $\{\widetilde{\Pi}_{\rho_k}\}_{k\in\mathbb{N}}$ be a sequence of constraint-relaxed posterior distributions with densities $\{\widetilde{\pi}_{\rho_k}(\boldsymbol{\theta} \mid \mathbf{y})\}_{k\in\mathbb{N}}$, respectively, where $\rho_k \uparrow \infty$ as $k \to \infty$. Then $\|\widetilde{\Pi}_{\rho_k} - \overline{\Pi}\|_{\mathsf{TV}} \to 0$ as $k \to \infty$. It follows that $\widetilde{\Pi}_{\rho_k} \xrightarrow{D} \overline{\Pi}$ as $k \to 0$.*

Theorem 2 is consistent with concurrent work by Zhou et al. (2022), who show convergence in total variation distance for posterior distributions under a more general class of epigraph priors.

Information Projection The preceding results primarily concern the limiting setting where ρ grows large, confirming that the relaxed posteriors under our priors tend to the sharply constrained posterior. However, a common modeling application in practice entails selecting a finite value $\rho < \infty$ to promote structure encoded in the constraint $\mathcal{C}$. The next contribution highlights a connection between constrained and constraint-relaxed posterior distributions from an information geometric perspective, with practical implications on this use case.

Consider the special case of the moment-constrained information projection problem, originally studied by Csiszár (1975):

$$p^*(\theta) = \arg\min_{p(\boldsymbol{\theta})} \int p(\boldsymbol{\theta}) \log\left(\frac{p(\boldsymbol{\theta})}{\pi(\boldsymbol{\theta} \mid \mathbf{y})}\right) d\boldsymbol{\theta} \tag{5}$$
$$\text{s.t.} \quad \mathbb{E}_{\boldsymbol{\theta}\sim p}\left[\frac{1}{2}\mathsf{dist}(\boldsymbol{\theta}, \boldsymbol{\Theta})^2\right] = D$$

Thus we are interested in finding the closest density $p^*(\boldsymbol{\theta})$ to the unconstrained posterior $\pi(\boldsymbol{\theta} \mid \mathbf{y})$ in terms of KL divergence such that the expected square distance of $\boldsymbol{\theta}$ to $\boldsymbol{\Theta}$ under $p^*(\boldsymbol{\theta})$ is equal to some given value D.

Theorem 3. *Suppose that* $\mathbb{E}_{\boldsymbol{\theta}\sim\pi(\boldsymbol{\theta}|\mathbf{y})}[\mathsf{dist}(\boldsymbol{\theta}, \boldsymbol{\Theta})^2/2] > D$. *Then the constraint-relaxed posterior distribution* $\widetilde{\pi}(\boldsymbol{\theta} \mid \mathbf{y})$ *(Equation 3) is the solution to the moment-constrained information projection problem (Equation 5):*

$$p^*(\boldsymbol{\theta}) \propto \pi(\boldsymbol{\theta} \mid \mathbf{y}) \exp\left(-\frac{\lambda}{2}\mathsf{dist}(\boldsymbol{\theta}, \boldsymbol{\Theta})^2\right),$$

where $\lambda > 0$ *is a Lagrange multiplier that satisfies the moment constraint under* $p^*(\boldsymbol{\theta})$.

The solution given in Theorem 3 is known as *exponential tilting* (West, 2020; Tallman and West, 2022). Observe that for $\lambda = 0$, $p^*(\boldsymbol{\theta}) = \pi(\boldsymbol{\theta} \mid \mathbf{y})$. Moreover, D and λ are inversely related (see Supplementary Material for additional details), so in particular, if $D \to 0$, then $\lambda \to \infty$ and $p^*(\boldsymbol{\theta}) \to \overline{\pi}(\boldsymbol{\theta} \mid \mathbf{y})$. Exponential tilting therefore creates a spectrum of constraint relaxation with the unconstrained posterior on one end, the constrained posterior on the other end, and the constraint-relaxed posterior as the optimal choice in the sense that it is the closest to $\pi(\boldsymbol{\theta} \mid \mathbf{y})$ while maintaining a specified distance from $\boldsymbol{\Theta}$ in expectation.

This perspective has practical implications. A common challenge in regularization problems involves specifying the penalty parameter when it does not have an interpretable scale. Theorem 3 suggests a systematic solution by identifying the Lagrange multiplier λ from the information projection with the penalty ρ. We can solve for λ given a value for D, and then use that value as the corresponding value for ρ in the distance-to-set regularization or the corresponding Bayesian constraint relaxation. Thought it appears we've simply swapped specifying ρ with D, it's important to note that D is often interpretable in practice as it lives on the same scale as $\boldsymbol{\theta}$, so we can choose the level of relaxation based on real-world inputs or units in application.

2.3 Prior work on Bayesian Constraint Relaxation

The task our contributions address is closely related to the Bayesian constraint relaxation work by Duan et al. (2020). There, the authors also consider relaxing a sharply constrained prior by quantifying the distance to the desired constraint, with particular attention to the case when the constraint sets which they denote D lie in a lower dimensional subspace of the full space $\mathbb{R}^d$. They construct posteriors of the form $\widetilde{\pi}_\lambda \propto \ell(\boldsymbol{\theta}; Y)\pi_R(\boldsymbol{\theta})\exp\{-\lambda^{-1}\|\nu_D(\boldsymbol{\theta})\|\}$, where $s < d$ denotes the dimension of the constraint set D, which is represented algebraically as a solution to the system of equations $\{\nu_j(\boldsymbol{\theta}) = 0\}_{j=1}^s$. Duan et al. (2020) choose to measure the constraint violation explicitly using the function $\|\nu_D(\boldsymbol{\theta})\| = \sum_{j=1}^s |\nu_j(\boldsymbol{\theta})|$.

The authors briefly comment that users may flexibly choose a measure of constraint violation: along this line, our method not only shows how the squared Euclidean distance is preferable in many ways over their choice of $\|\nu_D(\boldsymbol{\theta})\|$, but makes a departure from defining constraints algebraically and component-wise by grounding in a *projection-based* framework. That is, even when a constraint set D has measure zero in $\mathbb{R}^d$, for any point $\mathbf{x} \in \mathbb{R}^d$, its projection $P_D(\mathbf{x}) \in \mathbb{R}^d$ also lives in the ambient space. By exploiting the projection-based characterization of the distance from points to sets, our formulation handles constraints *implicitly*, yielding effective algorithms that stay in the original space. Not only does this avoid having to explicitly write constraints algebraically, but obviates technical geometric measure theoretic arguments by avoiding the need to operate directly in the subspace containing D and resolve the mismatch in dimension when mapping back into $\mathbb{R}^d$. The next section provides transparent intuition for why simply squaring the penalty makes a big difference in practice on the performance of gradient-based samplers.

Our work shares a connection with recent work that proposes a class of nondifferentiable priors called epigraph priors, presented in the context of Bayesian trend filtering (Heng et al., 2023). Though connections to proximal distance algorithms and distance majorization are not explicitly referenced by the authors, projection onto the epigraph of a regularization function g depends on the proximal mapping of g (Xu et al., 2021), and the success of their framework hinges on the same algorithmic primitives and known projection operators or proximal maps that make computation attractive in our case. Indeed, the proximal map of an indicator function $1_C(x)$ of a set C is given by the projection $P_C(x)$ onto C. From another perspective, the Moreau-Yosida envelope of $1_C(x)$ is given by the squared distance between x and C. While neither of these discusses the Bayesian constraint framework of Duan et al. (2020), they can be seen as a general context for this work, and concurrent work by Zhou et al. (2022) also remarks on the connection between such epigraph prior approaches and distance regularization from the optimization perspective.

2.4 Sampling via Hamiltonian Monte Carlo

Having established some of its properties, we now discuss how to draw samples from the posterior distribution effectively. We advocate Hamiltonian Monte Carlo (HMC) (Neal et al., 2011; Betancourt and Girolami, 2015), a popular gradient-based MCMC algorithm that leverages Hamiltonian dynamics to generate informed parameter proposals.

We briefly review the HMC framework: to sample from a posterior $\pi(\boldsymbol{\theta} \mid \mathbf{y}) \propto L(\boldsymbol{\theta} \mid \mathbf{y})\pi(\boldsymbol{\theta})$, where the posterior has support on $\mathbb{R}^d$, HMC begins by embedding $\boldsymbol{\theta}$ into $\mathbb{R}^{2d}$ via the introduction of an independent, auxiliary *momentum* parameter $\mathbf{p} \in \mathbb{R}^d$. The parameter of interest $\boldsymbol{\theta}$ plays the

role of the *position* vector; the sampler then explores their joint posterior: $\pi(\boldsymbol{\theta}, \mathbf{p} \mid \mathbf{y})$. Define the Hamiltonian function $H : \mathbb{R}^{2d} \to \mathbb{R}$ by $H(\boldsymbol{\theta}, \mathbf{p}) := -\log \pi(\boldsymbol{\theta}, \mathbf{p})$. By the independence of $\boldsymbol{\theta}$ and $\mathbf{p}$, we can write

$$H(\boldsymbol{\theta}, \mathbf{p}) = K(\mathbf{p}) + U(\boldsymbol{\theta}),$$

where one can take the kinetic energy to take the form $K(\mathbf{p}) := \frac{1}{2}\mathbf{p}^\intercal \mathbf{M}^{-1}\mathbf{p} + C$ for some constant C and mass matrix $\mathbf{M}$, and the potential energy $U(\boldsymbol{\theta}) := -\log \pi(\boldsymbol{\theta} \mid \mathbf{y})$. The Hamiltonian dynamics that describe how the parameters evolve over "time" impose structure on the manifold containing $(\boldsymbol{\theta}, \mathbf{p})$:

$$\begin{cases} \frac{d\boldsymbol{\theta}}{dt} = \nabla_{\boldsymbol{\theta}} H(\boldsymbol{\theta}, \mathbf{p}) = \nabla_{\boldsymbol{\theta}} \log \pi(\boldsymbol{\theta} \mid \mathbf{y}) \\ \frac{d\mathbf{p}}{dt} = -\nabla_{\mathbf{p}} H(\boldsymbol{\theta}, \mathbf{p}) = -\mathbf{M}^{-1}\mathbf{p} \end{cases}$$

Generally, there is no analytical tractable solution for this PDE, so we rely on what is known as the leap-frog integrator to discretize the PDE as follows. Given some step size ε and a number of steps L, we iterate for $l = 1, \ldots, L$:

1. $\mathbf{p}_{t+\varepsilon/2} = \mathbf{p}_t - \frac{\varepsilon}{2}\nabla_{\boldsymbol{\theta}} \log \pi(\boldsymbol{\theta} \mid \mathbf{y})\big|_{\boldsymbol{\theta}=\boldsymbol{\theta}_t}$
2. $\boldsymbol{\theta}_{t+\epsilon} = \boldsymbol{\theta}_t + \varepsilon \mathbf{M}^{-1}\mathbf{p}_{t+\varepsilon/2}$
3. $\mathbf{p}_{t+\varepsilon} = \mathbf{p}_{t+\varepsilon/2} - \frac{\varepsilon}{2}\nabla_{\boldsymbol{\theta}} \log \pi(\boldsymbol{\theta} \mid \mathbf{y})\big|_{\boldsymbol{\theta}=\boldsymbol{\theta}_{t+\varepsilon}}$

To incorporate this into a sampling algorithm, suppose we start with a current parameter draw $\boldsymbol{\theta}^{(s)}$. Draw $\mathbf{p}^0 \sim N_d(\mathbf{0}, \mathbf{M})$. Perform the leap-frog integrator to obtain a proposal $(\boldsymbol{\theta}^{(s+1)}, \mathbf{p}^*)$. After reversing the direction of momentum $-\mathbf{p}^* \mapsto \mathbf{p}^*$, we perform an accept-reject step to correct discretization error: accept $(\boldsymbol{\theta}^*, \mathbf{p}^*)$ with probability

$$\alpha = \min\left\{1, \frac{e^{-H(\boldsymbol{\theta}^{(s+1)}, \mathbf{p}^*)}}{e^{-H(\boldsymbol{\theta}^{(s)}, \mathbf{p}^0)}}\right\}.$$

Computational Advantages Duan et al. (2020) report instability in the HMC algorithm, particularly when constraints are tightly enforced (i.e., ρ is large) under their Bayesian constraint relaxation formulation. This section provides a simple explanation for this behavior by examining the gradients under each approach, and also reveals how our formulation avoids these pitfalls by yielding continuously differentiable gradients. In doing so, we greatly improve stability in HMC implementations so that adequate mixing is not restricted to narrow parameter ranges.

Proposition 2. *The log constraint-relaxed posterior* $\log \widetilde{\pi}(\boldsymbol{\theta} \mid \mathbf{y})$ *(Equation 3) is continuously differentiable as long as the log-posterior* $\log \pi(\boldsymbol{\theta} \mid \mathbf{y})$ *(Equation 2) is continuously differentiable in* $\boldsymbol{\theta}$.

The proof is detailed in the supplement, but follows readily from continuity and uniqueness of the projection, which are given by convexity. In particular, we see that the gradient

$$\nabla_{\boldsymbol{\theta}}\left[\frac{1}{2}\mathsf{dist}(\boldsymbol{\theta}, \boldsymbol{\Theta})^2\right] = \boldsymbol{\theta} - P_{\boldsymbol{\Theta}}(\boldsymbol{\theta})$$

converges continuously to 0 on the boundary of the constraint as desired.

To better understand advantages over prior work, we examine how the gradient would behave had we relaxed the constraint without squaring a distance-to-set penalty, akin to the focus on ℓ_1 approaches in (Duan et al., 2020). The log-posterior, denoted by $\widehat{\pi}(\theta)$ would be of the form:

$$\log \widehat{\pi}(\boldsymbol{\theta} \mid \mathbf{y}) = \log L(\mathbf{y} \mid \boldsymbol{\theta})\pi(\boldsymbol{\theta}) - \frac{\rho}{2}\mathsf{dist}(\boldsymbol{\theta}, \boldsymbol{\Theta}),$$

which is not smooth in general. In particular, examining the subdifferential with respect to $\boldsymbol{\theta}$ yields

$$\partial_{\boldsymbol{\theta}} \log \widehat{\pi}(\boldsymbol{\theta}) = \partial_{\boldsymbol{\theta}} \log L(\boldsymbol{\theta} \mid \mathbf{y})\pi(\boldsymbol{\theta}) - \begin{cases} \frac{\boldsymbol{\theta} - P_{\boldsymbol{\Theta}}(\boldsymbol{\theta})}{\|\boldsymbol{\theta} - P_{\boldsymbol{\Theta}}(\boldsymbol{\theta})\|_2}, & \boldsymbol{\theta} \notin \boldsymbol{\Theta} \\ 0, & \boldsymbol{\theta} \in \boldsymbol{\Theta} \end{cases}$$

Observe that the $\|\nabla_{\boldsymbol{\theta}}\mathsf{dist}(\boldsymbol{\theta}, \boldsymbol{\Theta})\|_2 = 1$ for $\boldsymbol{\theta} \notin \boldsymbol{\Theta}$, and 0 otherwise: the distance fails to be continuously differentiable at the boundary, instead *sharply transitioning* at a jump discontinuity. Computationally, this manifests as instability and poor mixing when the sampler is close to the constraint, as whenever $\boldsymbol{\theta} \approx P_{\boldsymbol{\Theta}}(\boldsymbol{\theta})$, the denominator becomes numerically close to 0. This agrees with empirical findings reported by Duan et al. (2020) and their remarks on instability in the Supplemental Materials.

Remark. We may weaken Assumption 3 so that $\boldsymbol{\Theta}$ is closed but not necessarily convex. In this case, Proposition 7 of Keys et al. (2019) assures that for a nonempty closed subset of $\mathbb{R}^n$, the projection operator is multi-valued on a set of measure zero, so the gradient formula for the squared distance function holds and is uniquely defined almost surely.

3 RESULTS AND PERFORMANCE

In this section, we investigate the performance of distance-to-set priors on increasingly more involved empirical studies. We find that our priors result in improved sampling performance relative to existing constraint relaxation methods.

Regression over the ℓ_2-Ball We illustrate our approach to measuring the uncertainty of distance-to-set penalization by considering a simple constrained formulation of the ridge regression problem. Here the constraint set $\mathcal{C} = B_2(0, 1)$ is the Euclidean unit ℓ_2-ball. From an estimation perspective, the proximal distance algorithm aims to solve the problem

$$\min_{\boldsymbol{\beta} \in \mathcal{C}} \|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|_2^2,$$

where $\mathbf{y} \in \mathbb{R}^n$, $\mathbf{X} \in \mathbb{R}^{n \times p}$, and $\boldsymbol{\beta} \in \mathbb{R}^p$, by considering a relaxed version. For a fixed $\rho \in (0, \infty)$,

$$\min_{\boldsymbol{\beta} \in \mathbb{R}^p} \|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|_2^2 + \frac{\rho}{2}\mathsf{dist}(\boldsymbol{\beta}, \mathcal{C})^2$$

Applying the iterations from the proximal distance algorithm without taking $\rho \uparrow \infty$ will solve this problem, the solution of which we denote by $\widehat{\boldsymbol{\beta}}$. To obtain corresponding uncertainty as measured by a posterior, consider the Gaussian model: $\mathbf{y} \mid \boldsymbol{\beta} \sim N(\mathbf{X}\boldsymbol{\beta}, \sigma^2\mathbf{I})$ with a flat prior $\pi(\boldsymbol{\beta}) \propto 1$. Then the constraint relaxed posterior is given by

$$\widetilde{\pi}(\boldsymbol{\beta} \mid \mathbf{y}) \propto \exp\left(-\frac{1}{2\sigma^2}\|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|_2^2\right) \exp\left(-\frac{\rho}{2}\mathsf{dist}(\boldsymbol{\beta}, \mathcal{C})^2\right)$$

Clearly, the MAP estimator $\widehat{\boldsymbol{\beta}}_{\text{MAP}}$ is equal to $\widehat{\boldsymbol{\beta}}$. Since we have a fully-specified posterior, we can supplement the estimator $\widehat{\boldsymbol{\beta}}$ with uncertainty quantification. Moreover, a more subtle point is that we can use the proximal distance algorithm to compute MAP estimates for the corresponding Bayesian model, which would normally be very difficult to obtain simply from drawing samples from the posterior.

We examine the performance in this model on simulated data. In this case, there is a simple closed-form expression for the projection:

$$P_{\mathcal{C}}(\boldsymbol{\beta}) = \begin{cases} \boldsymbol{\beta}/\|\boldsymbol{\beta}\|_2, & \boldsymbol{\beta} \notin \mathcal{C} \\ 0, & \boldsymbol{\beta} \in \mathcal{C} \end{cases}.$$

We choose $p = 2$ for easy of visualization, and generate the true $\boldsymbol{\beta} = (-1.295, -0.532)$ to lie outside of $\mathcal{C}$. We then draw $n = 100$ observations from a linear model under $\boldsymbol{\beta}$. To sample from the posterior distribution, we use the `stan` functionality in `R` that leverages NUTS-HMC (Hoffman et al., 2014), and set the hyperparameter $\rho = 10^3$ to tightly enforce the constraint.

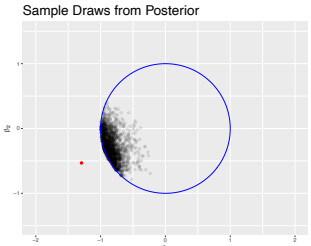


Figure 2: Draws from relaxed posterior, ridge regression.

Figure 2 displays the sample draws. At a glance, one can see that the posterior posterior distribution is concentrated near the boundary in the bottom-left quadrant since the true $\boldsymbol{\beta}$, denoted as a red point, lies in that direction. The posterior samples allow one to conduct inference. For instance, the 95% equi-tailed credible intervals for β_1 and β_2 are $(-0.99, -0.54)$ and $(-0.65, 0.11)$, respectively.

We further examine the impact of squaring the distance-to-set operator on posterior sampling performance in this context. Figure 3 depicts the trace plots and autocorrelation function (ACF) under a squared and unsquared distance-to-set term in the prior. The trace plots suggest better mixing and slightly less stickiness in the sampling trajectories. Moreover, there is a noticeable reduction in dependence between sample draws when using the squared distance-to-set priors based on the ACF plots. Overall, this is a relatively simple example—the dimension of the constraint set matches the dimensions of the ambient space within which it is embedded. In fact, both squared and unsquared priors perform well, and the samples from the latter look essentially the same as Figure 2, though we already see computational improvements by examining properties of the chain. These differences become more pronounced as we consider more challenging settings, such as a lower dimensional constraint set in the next example.

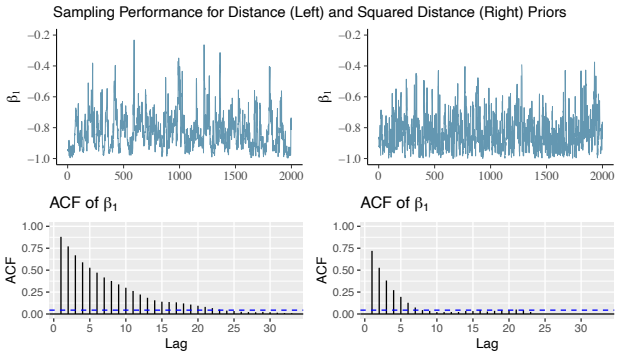


Figure 3: (Regression) Trace plots and ACF plots for both the unsquared (left) and squared (right) distance-to-set priors for β_1. Plots for β_2 look similar and are omitted.

Sampling along a Lower-Dimensional Surface The von Mises-Fisher distribution $\mathsf{vMF}(\alpha, \mathbf{F})$ is supported on the sphere S^p with $\alpha \geq 0$ and $\mathbf{F} \in S^p$. When $\alpha = 0$, this reduces to the uniform distribution on the sphere (Fisher, 1953). One can then envision the von Mises-Fisher distribution as being a spherical distribution concentrated around a unit vector. Duan et al. (2020) observe that this distribution can be described as a multivariate normal with mean vector $\mathbf{F} \in \mathbb{R}^{p+1}$ constrained to the unit sphere. They further consider a generalization in which the multivariate normal likelihood is replaced with a multivariate Student-t distribution with m degrees of freedom, mean vector $\mathbf{F} \in \mathbb{R}^{p+1}$, and variance $\sigma^2\mathbf{I}_{p+1}$. Using distance-to-set priors, we revisit this setting and relax the constraint so that the points have to lie close to the constraint surface, targeting sampling from the following distribution:

$$\widetilde{\pi}(\boldsymbol{\theta} \mid \mathbf{y}) \propto \left(1 + \frac{\|\mathbf{F} - \boldsymbol{\theta}\|_2^2}{m\sigma^2}\right)^{-\frac{m+p}{2}} \exp\left(-\frac{\rho}{2}\mathsf{dist}(\boldsymbol{\theta}, S^p)^2\right)$$

Observe that S^2 has a smaller dimension that the space within which it is embedded, namely $\mathbb{R}^3$. We demonstrate how one would use distance-to-set constraint relaxation in this setting, we specify the the projection P_{S^p}, which maps $0 \neq \boldsymbol{\theta} \in \mathbb{R}^n$ to $\boldsymbol{\theta}/\|\boldsymbol{\theta}\|$. Thus, $\mathsf{dist}(\boldsymbol{\theta}, S^p) = \|\boldsymbol{\theta} - P_{S^p}(\boldsymbol{\theta})\|$. In Duan et al. (2020), the distance from the constraint is considered algebraically by $\nu(\boldsymbol{\theta}) = |\boldsymbol{\theta}^\intercal\boldsymbol{\theta} - 1|$. In essence, the distance from $\boldsymbol{\theta}$ to S^p is given by the distance in the level curve it lies on $\boldsymbol{\theta}^\intercal\boldsymbol{\theta}$ and the level curve defining S^p.

As such, we refer to this as the *level set relaxation* prior in comparisons reported here.

We now compare these two Bayesian constraint relaxation approaches using `stan`. For sampling using the relaxation $\nu(\boldsymbol{\theta})$, we use publicly accessible code obtainable in Duan. For our distance-to-set prior, we only need to update the constraint relaxation term. We summarize how these Bayesian constraint relaxation methods perform below for $p = 2$ (the sphere that forms the boundary of the Euclidean ℓ_2 unit ball in $\mathbb{R}^3$). Figure 4 plots results after drawing 2000 samples points using the peer algorithms, thinned by a factor of 10 for visual clarity. The distance-to-set prior mimics the theoretical draws from the vMF distribution, with some deviation due to a mild degree of constraint relaxation and slightly different tail behavior. In contrast, the level set relaxation prior leads to a chain that gets stuck during sampling, and the range of samples do not appear to be near the target constraint surface.

Table 1: Sampling Performance for Level Set Prior vs. Distance-to-Set Prior on Robust vMF Distribution

		Level Set Relaxation Prior				Distance-to-Set Prior			
ρ	Axis	Mean	2.5%	97.5%	ESS	Mean	2.5%	97.5%	ESS
1000	x	0.59	0.18	0.94	31.55	0.52	-0.02	0.93	853.22
	y	0.54	0.05	0.86	9.38	0.52	-0.08	0.93	736.91
	z	0.48	0.04	0.88	11.78	0.53	-0.11	0.96	728.31
10000	x	0.26	-0.10	0.57	1.35	0.51	-0.13	0.92	750.94
	y	0.41	-0.13	0.78	1.05	0.51	-0.14	0.93	650.81
	z	-0.75	-1.00	-0.47	1.02	0.51	-0.09	0.93	622.50
1e+05	x	0.27	0.07	0.46	1.00	0.50	-0.21	0.92	751.80
	y	0.06	-0.88	0.99	1.00	0.50	-0.29	0.94	600.63
	z	-0.01	-0.14	0.12	1.00	0.49	-0.22	0.92	702.80
1e+06	x	-0.38	-0.46	-0.30	1.00	0.52	-0.02	0.92	779.55
	y	0.51	0.06	0.95	1.00	0.51	-0.15	0.92	559.85
	z	-0.40	-0.89	0.08	1.00	0.53	-0.06	0.93	542.38

Figure 4 shows a cluster of sample draws for the level set relaxation prior, suggesting that sampler is sticky and does not explore the space well. On the other hand, our distance-to-set prior resembles draws from the theoretical distribution fairly well. Table 1 further reinforces this point. As we decrease λ (i.e., enforce the constraint more strictly), we see that the distribution concentrates away from the mean vector $(1/\sqrt{3}, 1/\sqrt{3}, 1/\sqrt{3})$, and the ESS (out of 1,000 post-warmup iterations per chain, 2 chains) is low. Our novel distance-to-set prior concentrates around the mean vector, and the ESS remains consistently high. Finally, the acceptance rate for our method ranges between 0.932—0.937, while it ranges between 0.766—0.815 for the level set relaxation prior. As it is desirable for acceptance rates to be close to 100% since Metropolis steps in HMC are meant to correct only for numerical error, this makes a strong case for the sampling performance under our approach.

3.1 Real Data Case Study

While the simulation studies highlight advantages of our approach, the final example considers a case study whose constraint is nontrivial to incorporate within prior methods. We apply our distance-to-set priors to constraints imposed on contingency tables imbued with isotonic constraints. We follow the design introduced by Agresti and Coull (2002) in which four treatment group doses were given (Placebo, Low dose, Medium dose, High dose) to patients with subarachnoid hemorrhage, and the outcomes were examined (Good recovery, Minor disability, Major disability, Vegetative state, and Death) to construct a dose-response curve. The data appears in Agresti and Coull (2002), summarized in the table below. The constraint on this table that is natural to assume is for the outcome to stochastically increase with respect to the treatment, which we formalize below.

A model for the order-based constraints on this contingency table is described by Sen et al. (2018). Following this treatment, suppose we have n observations exhaustively distributed over an $I \times J$ contingency table with entries n_{ij} for $i \in [I]$ and $j \in [J]$, where $[m] := \{1, \ldots, m\}$. We let the rows represent the doses, and the columns represent the outcomes. Suppose further that the probability of each observation ending up in the (i, j) cell is given by θ_{ij}. Let $n_{[i]} := \sum_{i \in [I]} n_{ij}$, and similarly, $\boldsymbol{\theta}_{[i]} := (\theta_{i1}, \ldots, \theta_{iJ})$:

	Recovery	Minor Disability	Major Disability	Vegetative State	Death
Placebo	θ_{11}	θ_{12}	θ_{13}	θ_{14}	θ_{15}
Low Dose	θ_{21}	θ_{22}	θ_{23}	θ_{24}	θ_{25}
Medium Dose	θ_{31}	θ_{32}	θ_{33}	θ_{34}	θ_{35}
High Dose	θ_{41}	θ_{42}	θ_{43}	θ_{44}	θ_{45}

Then we take the following model: for each $i \in [I]$ suppose

$$(n_{i1}, \ldots, n_{iJ}) \overset{\perp\!\!\!\perp}{\sim} \mathsf{Multi}(n_{[i]}, \boldsymbol{\theta}_{[i]}), \quad \boldsymbol{\theta}_{[i]} \overset{\perp\!\!\!\perp}{\sim} \mathsf{Dir}(\boldsymbol{\alpha}).$$

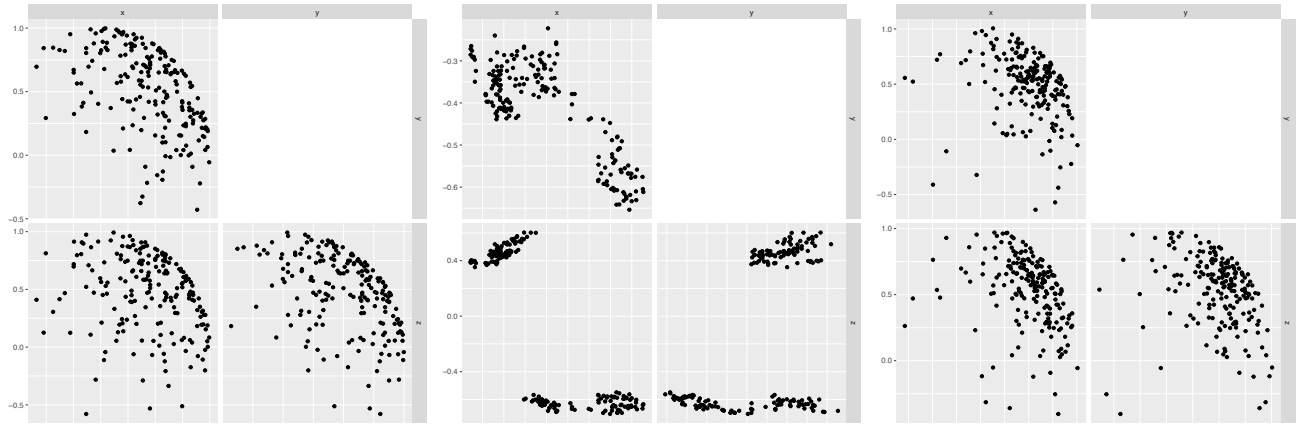
We impose the stochastic dominance constraint on the probabilities governing the contingency table as follows: for all $i \in [I]$, for all $j \in [J]$,

$$\sum_{k=1}^{j} \theta_{i+1,k} \geq \sum_{k=1}^{j} \theta_{ik}.$$

We may write the set of such probabilities obeying this stochastic dominance as the following isotonic constraint:

$$\Theta_{CT} := \left\{ (\theta_{ij})_{i \in I, j \in J} \,\middle|\, \sum_{k=1}^{j} \theta_{i+1,k} \geq \sum_{k=1}^{j} \theta_{ik} \text{ for } i \in [I], j \in [J] \right\}$$

As with any application of distance-to-set penalties or priors, a crucial subroutine requires computing the projection onto $\boldsymbol{\Theta}_{CT}$. It is not clear whether implementing the projection directly in a Stan file (Stan Development Team, 2022) is possible; instead, we implement our HMC-based sampler directly in `R`. We use a quadratic programming algorithm (Goldfarb and Idnani, 1982, 1983) available in the `quadprog` library (Turlach and Weingessel, 2013) to compute the projections that appear in the gradient of the log-posterior, and include the complete implementation details to the Supplement. It is worth noting that despite four seemingly independent multinomial-Dirichlet models, the stochastic dominance constraints entangle the distributions. The resulting constrained problem is complex, and distinct



(a) Theoretical Draws (vMF) (b) Level Set Relaxation (Robust vMF) (c) Distance-to-Set (Robust vMF)

Figure 4: Exact draws using the using the `rvmf` function in the `rFast` package compared to samples using the method of Duan et al. (2020) and our proposed method under $\rho = 10^5$ with $\mathbf{F} = (1/\sqrt{3}, 1/\sqrt{3}, 1/\sqrt{3})$, $\sigma^2 = 0.1$, and $m = 3$.

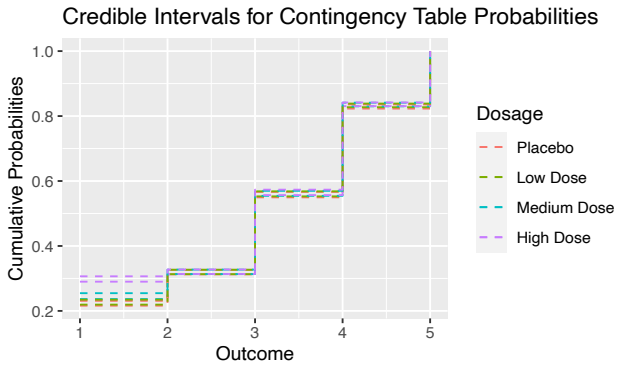


Figure 5: (Contingency Table) 95% posterior credible intervals under distance-to-set priors with $\rho = 7.5 \times 10^5$.

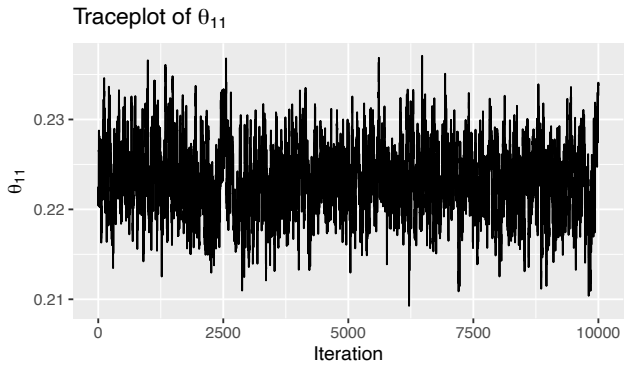


Figure 6: (Contingency Table) The trace plot of samples for the parameter $\theta_{1,1}$ in the contingency table.

from a standard setting with separate isotonic constraints (Chatterjee et al., 2015), which can be handled using the simpler pooled adjacent violators algorithm (PAVA).

Figure 5 displays the 95% credible intervals for the probabilities governing the contingency tables; detailed numerical values are tabulated in the Supplement. As we consider a large value of $\rho = 7.5 \times 10^5$, it is not surprising that the isotonic constraints are well-respected at the quantiles.

Despite the high degree of constraint enforcement, Figure 6 shows that our approach to constraint relaxation maintains strong performance despite a naïve implementation.

4 DISCUSSION

In this work, we revisit distance-to-set regularization from a Bayesian perspective, and advocate a flexible class of distance-to-set priors to allow uncertainty quantification under a well-defined posterior. Our class of priors is smooth, which is a crucial factor in the improved sampling performance under implementations such as HMC. Empirical results reflect this design, so that performance does not deteriorate even when the constraint set is a lower-dimensional manifold, the parameter ρ is large, or $P_{\boldsymbol{\Theta}}(\boldsymbol{\theta})$ cannot be expressed analytically in closed form. When $\rho \to \infty$, distance-to-set priors agree with the sharply constrained Bayesian inference problem, and when $\rho < \infty$, constraint-relaxed posteriors are optimal approximations in terms of KL divergence, while respecting the constraint to a specified amount in expectation.

There are a number of open extensions and inferential questions that remain interesting directions for future work in view of distance-to-set regularization. For instance, loss functions do not always originate from likelihoods. Although there is no longer a valid posterior distribution in such instances, Bissiri et al. (2016) proposes using loss functions in place of likelihoods to obtain Gibbs posteriors. Rigon et al. (2023) proposed using these for clustering more recently. We believe distance-to-set priors can similarly incorporate constraint information into these generalized

posteriors rather seamlessly.

We show how to incorporate a relaxed constraint into a prior in a way that is amenable to posterior computation, but have in a sense taken the original prior "for granted". The meaning of resulting analyses after imposing constraints on the prior may not be straightforward depending on the context (Stark, 2015), and unintuitive phenomena may arise as a result of mathematical artifacts; see for example the discussion in Section 3 of Gelman (1996). More statistical implications such as posterior concentration behavior under particular prior choices are also worthy of further study.

We use an ℓ_2 metric throughout, largely motivated by the smoothness of the distance-to-set operator. However, it may be more natural to consider non-Euclidean measures of nearness in many settings. For instance, when one seeks a prior to constrain a class of probability distributions to a subdomain of the probability simplex, KL divergence may be preferable. Bregman divergences (Bregman, 1967) more generally are a natural choice worth exploring further. A related and more general task is to more closely examine distance-to-set priors when the constraints comprise Riemannian manifolds. Algorithms such as RM-HMC (Girolami and Calderhead, 2011) allow the modeler to explore a highly non-Euclidean space using data locally encoded via Riemannian metrics, but it is not always easy to constrain samplers to operate exactly on the surface at each iteration. Incorporating distance-to-set priors into these more general modeling settings is an exciting direction to investigate.

A core computational subroutine in pursuing these extensions involves preserving the ability to efficiently compute projections. Convenient expressions are available for some constraint sets under these non-Euclidean settings, though we expect further investigation is critical to broaden the classes of projections and approximate projections we can consider in generality.

While we begin to examine the information geometric underpinnings along these lines via exponential tilting, these connections remain underexplored. There are likely deeper insights in these directions that may reveal further properties or motivate new classes of priors. We invite readers to consider future investigation into these promising research directions.

Acknowledgements

This work was partially supported by NSF grant DMS-2230074. We thank Alice Cima for stimulating discussions. We also thank Mike West and Emily Tallman for insightful discussions on exponential tilting and KL divergence.

References

Alan Agresti and Brent A Coull. The analysis of contingency tables under inequality constraints. *Journal of Statistical Planning and Inference*, 107(1-2):45–73, 2002.

Michael Betancourt. Cruising the simplex: Hamiltonian Monte Carlo and the Dirichlet distribution. In *AIP Conference Proceedings 31st*, volume 1443, pages 157–164. American Institute of Physics, 2012.

Michael Betancourt and Mark Girolami. Hamiltonian Monte Carlo for hierarchical models. *Current trends in Bayesian methodology with applications*, 79:2–4, 2015.

Patrick Billingsley. *Convergence of probability measures*. John Wiley & Sons, 2013.

Pier Giovanni Bissiri, Chris C Holmes, and Stephen G Walker. A general framework for updating belief distributions. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 78(5):1103–1130, 2016.

Stephen Boyd and Lieven Vandenberghe. *Convex optimization*. Cambridge university press, 2004.

Lev M Bregman. The relaxation method of finding the common point of convex sets and its application to the solution of problems in convex programming. *USSR computational mathematics and mathematical physics*, 7 (3):200–217, 1967.

Simon Byrne and Mark Girolami. Geodesic Monte Carlo on embedded manifolds. *Scandinavian Journal of Statistics*, 40(4):825–845, 2013.

Sabyasachi Chatterjee, Adityanand Guntuboyina, and Bodhisattva Sen. On risk bounds in isotonic and other shape restricted regression problems. *The Annals of Statistics*, 43(4):1774–1800, 2015.

Eric C Chi, Hua Zhou, and Kenneth Lange. Distance majorization and its applications. *Mathematical programming*, 146(1):409–436, 2014.

Imre Csiszár. I-divergence geometry of probability distributions and minimization problems. *The annals of probability*, pages 146–158, 1975.

Leo Duan. Bayescore. `https://github.com/leoduan/BayesCoRe`. Accessed: Sept. 7, 2022.

Leo L Duan, Alexander L Young, Akihiko Nishimura, and David B Dunson. Bayesian constraint relaxation. *Biometrika*, 107(1):191–204, 2020.

Rory A. Fisher. Dispersion on a sphere. *Proceedings of the Royal Society of London. Series A. Mathematical and Physical Sciences*, 217:295 – 305, 1953.

Andrew Gelman. Bayesian model-building by pure thought: some principles and examples. *Statistica Sinica*, pages 215–232, 1996.

Malay Ghosh. Constrained Bayes estimation with applications. *Journal of the American Statistical Association*, 87 (418):533–540, 1992.

Mark Girolami and Ben Calderhead. Riemann manifold Langevin and Hamiltonian Monte Carlo methods. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 73(2):123–214, 2011.

Donald Goldfarb and Ashok Idnani. Dual and primal-dual methods for solving strictly convex quadratic programs. In *Numerical analysis*, pages 226–239. Springer, 1982.

Donald Goldfarb and Ashok Idnani. A numerically stable dual method for solving strictly convex quadratic programs. *Mathematical programming*, 27(1):1–33, 1983.

Robert B Gramacy, Genetha A Gray, Sébastien Le Digabel, Herbert KH Lee, Pritam Ranjan, Garth Wells, and Stefan M Wild. Modeling an augmented lagrangian for blackbox constrained optimization. *Technometrics*, 58(1): 1–11, 2016.

Qiang Heng, Hua Zhou, and Eric C. Chi. Bayesian trend filtering via proximal Markov chain Monte Carlo. *Journal of Computational and Graphical Statistics, In press*, 2023. doi: 10.1080/10618600.2023.2170089.

Matthew D Hoffman, Andrew Gelman, et al. The No-U-Turn sampler: adaptively setting path lengths in Hamiltonian Monte Carlo. *J. Mach. Learn. Res.*, 15(1):1593–1623, 2014.

Pierre E Jacob, Lawrence M Murray, Chris C Holmes, and Christian P Robert. Better together? Statistical learning in models made of modules. *arXiv preprint arXiv:1708.08719*, 2017.

Kevin L Keys, Hua Zhou, and Kenneth Lange. Proximal distance algorithms: Theory and practice. *The Journal of Machine Learning Research*, 20(1):2384–2421, 2019.

Yunbum Kook, Yin Tat Lee, Ruoqi Shen, and Santosh S Vempala. Sampling with Riemannian Hamiltonian Monte Carlo in a constrained space. *arXiv preprint arXiv:2202.01908*, 2022.

Shiwei Lan, Bo Zhou, and Babak Shahbaba. Spherical Hamiltonian Monte Carlo for constrained target distributions. In *International Conference on Machine Learning*, pages 629–637. PMLR, 2014.

Alfonso Landeros, Oscar Hernan Madrid Padilla, Hua Zhou, and Kenneth Lange. Extensions to the proximal distance method of constrained optimization. *Journal of Machine Learning Research*, 23(182):1–45, 2022.

Kenneth Lange. *MM optimization algorithms*. SIAM, 2016.

James R Munkres. *Topology*, volume 2. Prentice Hall Upper Saddle River, 2000.

Radford M Neal et al. MCMC using Hamiltonian dynamics. *Handbook of Markov chain Monte Carlo*, 2(11):2, 2011.

Neal Parikh and Stephen Boyd. Proximal algorithms. *Foundations and Trends in optimization*, 1(3):127–239, 2014.

Marcelo Pereyra. Proximal Markov Chain Monte Carlo algorithms. *Statistics and Computing*, 26(4):745–760, 2016.

Tommaso Rigon, Amy H Herring, and David B Dunson. A generalized Bayes framework for probabilistic clustering. *Biometrika*, 2023. ISSN 1464-3510. doi: 10.1093/biomet/asad004.

John C Robertson, Ellis W Tallman, and Charles H Whiteman. Forecasting using relative entropy. *Journal of Money, Credit, and Banking*, 37(3):383–401, 2005.

Deborshee Sen, Sayan Patra, and David Dunson. Constrained inference through posterior projections. *arXiv preprint arXiv:1812.05741*, 2018.

Stan Development Team. RStan: the R interface to Stan, 2022. URL `https://mc-stan.org/`. R package version 2.21.5.

Philip B Stark. Constraints versus priors. *SIAM/ASA Journal on Uncertainty Quantification*, 3(1):586–598, 2015.

Emily Tallman and Mike West. On entropic tilting and predictive conditioning. *arXiv preprint arXiv:2207.10013*, 2022.

Robert Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Methodological)*, 58(1):267–288, 1996.

Berwin A Turlach and Andreas Weingessel. quadprog: Functions to solve quadratic programming problems. r package version 1.5-5, 2013.

Mike West. Perspectives on Bayesian decision analysis and constrained forecasting. *arXiv preprint arXiv:2007.11037*, 2020.

Stephen Wright, Jorge Nocedal, et al. Numerical optimization. *Springer Science*, 35(67-68):7, 1999.

Daniel Eliot Wulbert. Continuity of metric projections. *Transactions of the American Mathematical Society*, 134 (2):335–341, 1968.

Jason Xu and Kenneth Lange. A proximal distance algorithm for likelihood-based sparse covariance estimation. *Biometrika*, 109(4):1047–1066, 2022.

Jason Xu, Eric Chi, and Kenneth Lange. Generalized linear model regression under distance-to-set penalties. *Advances in Neural Information Processing Systems*, 30, 2017.

Maoran Xu, Hua Zhou, Yujie Hu, and Leo L Duan. Bayesian inference using the proximal mapping: Uncertainty quantification under varying dimensionality. *arXiv preprint arXiv:2108.04851*, 2021.

Xinkai Zhou, Eric C Chi, and Hua Zhou. Proximal MCMC for Bayesian inference of constrained and regularized estimation. *arXiv preprint arXiv:2205.07378*, 2022.

Supplementary Materials

A NUMERICAL DATA FOR STOCHASTIC ORDERING CASE STUDY

We provide numerical data supporting the credible intervals shown in Figure 5 of the paper.

2.5^{th} Percentile

	$j=1$	$j=2$	$j=3$	$j=4$	$j=5$
$i=1$	0.2161	0.3126	0.5496	0.8229	1.0000
$i=2$	0.2194	0.3131	0.5527	0.8272	1.0000
$i=3$	0.2365	0.3134	0.5544	0.8300	1.0000
$i=4$	0.2900	0.3137	0.5566	0.8306	1.0000

50^{th} Percentile

	$j=1$	$j=2$	$j=3$	$j=4$	$j=5$
$i=1$	0.2234	0.3194	0.5581	0.8297	1.0000
$i=2$	0.2266	0.3199	0.5605	0.8329	1.0000
$i=3$	0.2456	0.3202	0.5621	0.8356	1.0000
$i=4$	0.2983	0.3206	0.5646	0.8360	1.0000

97.5^{th} Percentile

	$j=1$	$j=2$	$j=3$	$j=4$	$j=5$
$i=1$	0.2310	0.3265	0.5663	0.8358	1.0000
$i=2$	0.2346	0.3268	0.5683	0.8384	1.0000
$i=3$	0.2546	0.3271	0.5698	0.8412	1.0000
$i=4$	0.3065	0.3277	0.5734	0.8418	1.0000

B PROOFS

B.1 Proof of Proposition 1

Proposition 1. *Under Assumptions 1 and 2, the constraint relaxed posterior $\widetilde{\pi}(\boldsymbol{\theta} \mid \mathbf{y})$ is a proper density.*

Proof. Noting that $\exp\left(-\frac{\rho}{2}\mathsf{dist}(\boldsymbol{\theta}, \boldsymbol{\Theta})^2\right) \leq 1$ for all $\boldsymbol{\theta} \in \mathbb{R}^d$, we immediately obtain from Assumption 2

$$\int_{\mathbb{R}^d} L(\boldsymbol{\theta} \mid \mathbf{y})\pi(\boldsymbol{\theta}) \exp\left(-\frac{\rho}{2}\mathsf{dist}(\boldsymbol{\theta}, \boldsymbol{\Theta})^2\right) \, d\boldsymbol{\theta} \leq \int_{\mathbb{R}^d} L(\boldsymbol{\theta} \mid \mathbf{y})\pi(\boldsymbol{\theta}) \, d\boldsymbol{\theta} < \infty.$$

□

B.2 Proof of Theorem 1

Theorem 1. *Suppose the unconstrained posterior $\pi(\boldsymbol{\theta} \mid \mathbf{y})$ is strictly log-concave. Let $\{\widetilde{\pi}_{\rho_k}(\boldsymbol{\theta} \mid \mathbf{y})\}_{k\in\mathbb{N}}$ be a sequence of constraint-relaxed posterior distributions where $\rho_k \uparrow \infty$ as $k \to \infty$. Further, define the following MAP estimators*

$$\widehat{\boldsymbol{\theta}}^M = \operatorname{argmax}_{\boldsymbol{\theta}} \overline{\pi}(\boldsymbol{\theta} \mid \mathbf{y}), \quad \widehat{\boldsymbol{\theta}}^M_{\rho_k} = \operatorname{argmax}_{\boldsymbol{\theta}} \widetilde{\pi}_{\rho_k}(\boldsymbol{\theta} \mid \mathbf{y}).$$

Then the sequence $\widehat{\boldsymbol{\theta}}^M_{\rho_k} \to \widehat{\boldsymbol{\theta}}^M$ as $k \to \infty$.

Proof. We closely follow the argument of (Wright et al., 1999, Theorem 17.1). Let $\boldsymbol{\theta}^*$ be a limit point of the sequence of MAP estimates $\left\{\widehat{\boldsymbol{\theta}}^M_{\rho_k}\right\}_{k\in\mathbb{N}}$ corresponding to the constraint-relaxed posterior distributions $\widetilde{\pi}_{\rho_k}(\boldsymbol{\theta} \mid \mathbf{y})$, and let $\widehat{\boldsymbol{\theta}}^M$ be the MAP estimate corresponding to the constrained posterior distribution $\overline{\pi}(\boldsymbol{\theta} \mid \mathbf{y})$. For the remainder of the proof, we drop the superscript M to ease notation.

We start with the basic inequality $\log \widetilde{\pi}_{\rho_k}\left(\widehat{\boldsymbol{\theta}} \mid \mathbf{y}\right) \leq \log \widetilde{\pi}_{\rho_k}\left(\widehat{\boldsymbol{\theta}}_{\rho_k} \mid \mathbf{y}\right)$ and expand:

$$\log \pi\left(\widehat{\boldsymbol{\theta}} \mid \mathbf{y}\right) - \frac{\rho}{2} \underbrace{\mathsf{dist}\left(\widehat{\boldsymbol{\theta}}, \boldsymbol{\Theta}\right)^2}_{=0} + \log C_{\rho_k} \leq \log \pi\left(\widehat{\boldsymbol{\theta}}_{\rho_k} \mid \mathbf{y}\right) - \frac{\rho_k}{2}\mathsf{dist}\left(\widehat{\boldsymbol{\theta}}_{\rho_k}, \boldsymbol{\Theta})^2\right) + \log C_{\rho_k},$$

where $C_{\rho_k} = \left(\int_{\mathbb{R}^d} \widetilde{\pi}_{\rho_k}(\boldsymbol{\theta} \mid \mathbf{y})\, d\boldsymbol{\theta}\right)^{-1}$ is the normalizing constant for $\widetilde{\pi}(\boldsymbol{\theta} \mid \mathbf{y})$. Simplifying the above expression gives

$$\log \pi\left(\widehat{\boldsymbol{\theta}} \mid \mathbf{y}\right) \leq \log \pi\left(\widehat{\boldsymbol{\theta}}_{\rho_k} \mid \mathbf{y}\right) - \frac{\rho_k}{2}\mathsf{dist}\left(\widehat{\boldsymbol{\theta}}_{\rho_k}, \boldsymbol{\Theta}\right)^2. \tag{*}$$

A bit of algebra shows that

$$0 \leq \mathsf{dist}\left(\widehat{\boldsymbol{\theta}}_{\rho_k}, \boldsymbol{\Theta}\right)^2 \leq \frac{2}{\rho_k}\left(\log \pi\left(\widehat{\boldsymbol{\theta}}_{\rho_k} \mid \mathbf{y}\right) - \log \pi\left(\widehat{\boldsymbol{\theta}} \mid \mathbf{y}\right)\right)$$

Since $\rho_k \uparrow \infty$ as $k \to \infty$ and $\log \pi(\cdot \mid \mathbf{y})$ is continuous, we have that

$$\frac{2}{\rho_k}\left(\log \pi\left(\widehat{\boldsymbol{\theta}}_{\rho_k} \mid \mathbf{y}\right) - \log \pi\left(\widehat{\boldsymbol{\theta}} \mid \mathbf{y}\right)\right) \overset{k\to\infty}{\longrightarrow} 0,$$

which implies that

$$\mathsf{dist}(\boldsymbol{\theta}^*, \boldsymbol{\Theta}) = \lim_{k\to\infty} \mathsf{dist}\left(\widehat{\boldsymbol{\theta}}_{\rho_k}, \boldsymbol{\Theta}\right) = 0.$$

The equality in the above expression follows from the continuity of $\mathsf{dist}(\cdot, \boldsymbol{\Theta})$ (Lange, 2016, Proposition 2.7.1(a)). Thus, $\boldsymbol{\theta}^* \in \boldsymbol{\Theta}$.

Next, since $\rho > 0$ and $\mathsf{dist}(\cdot, \boldsymbol{\Theta}) \geq 0$, we have for all $k \in \mathbb{N}$,

$$\log \pi\left(\boldsymbol{\theta}^* \mid \mathbf{y}\right) \geq \log \pi\left(\boldsymbol{\theta}^* \mid \mathbf{y}\right) - \frac{\rho_k}{2}\mathsf{dist}\left(\widehat{\boldsymbol{\theta}}_{\rho_k}, \boldsymbol{\Theta}\right)^2,$$

so we must have

$$\log \pi\left(\boldsymbol{\theta}^* \mid \mathbf{y}\right) \geq \lim_{k\to\infty}\left(\log \pi\left(\boldsymbol{\theta}^* \mid \mathbf{y}\right) - \frac{\rho_k}{2}\mathsf{dist}\left(\widehat{\boldsymbol{\theta}}_{\rho_k}, \boldsymbol{\Theta}\right)^2\right) = \log \pi\left(\boldsymbol{\theta}^* \mid \mathbf{y}\right) - \lim_{k\to\infty}\frac{\rho_k}{2}\mathsf{dist}\left(\widehat{\boldsymbol{\theta}}_{\rho_k}, \boldsymbol{\Theta}\right)^2 \tag{**}$$

Taking $k \to \infty$ on both sides of (*), we obtain

$$\log \pi\left(\widehat{\boldsymbol{\theta}} \mid \mathbf{y}\right) \leq \lim_{k\to\infty}\left(\log \pi\left(\boldsymbol{\theta}^* \mid \mathbf{y}\right) - \frac{\rho_k}{2}\mathsf{dist}\left(\widehat{\boldsymbol{\theta}}_{\rho_k}, \boldsymbol{\Theta}\right)^2\right) = \log \pi\left(\boldsymbol{\theta}^* \mid \mathbf{y}\right) - \lim_{k\to\infty}\frac{\rho_k}{2}\mathsf{dist}\left(\widehat{\boldsymbol{\theta}}_{\rho_k}, \boldsymbol{\Theta}\right)^2 \tag{***}$$

Combining (**) and (***), we obtain the inequality:

$$\log \pi\left(\boldsymbol{\theta}^* \mid \mathbf{y}\right) \geq \log \pi\left(\widehat{\boldsymbol{\theta}} \mid \mathbf{y}\right).$$

By the strict log-concavity of $\pi(\boldsymbol{\theta} \mid \mathbf{y})$ and the convexity of $\boldsymbol{\Theta}$ (Assumption 3), we have that $\widehat{\boldsymbol{\theta}}$ is the unique global maximum of $\overline{\pi}(\boldsymbol{\theta} \mid \mathbf{y})$. Therefore, $\boldsymbol{\theta}^* = \widehat{\boldsymbol{\theta}}$, and the conclusion follows. □

B.3 Proof of Theorem 2

Theorem 2. *Let $\overline{\Pi}$ be the constrained posterior distribution with density $\overline{\pi}(\boldsymbol{\theta} \mid \mathbf{y})$, and let $\{\widetilde{\Pi}_{\rho_k}\}_{k\in\mathbb{N}}$ be a sequence of constraint-relaxed posterior distributions with densities $\{\widetilde{\pi}_{\rho_k}(\boldsymbol{\theta} \mid \mathbf{y})\}_{k\in\mathbb{N}}$, respectively, where $\rho_k \uparrow \infty$ as $k \to \infty$. Then $\|\widetilde{\Pi}_{\rho_k} - \overline{\Pi}\|_{\mathsf{TV}} \to 0$ as $k \to \infty$. It follows that $\widetilde{\Pi}_{\rho_k} \xrightarrow{D} \overline{\Pi}$ as $k \to 0$.*

Proof. By Assumption 1, the constrained distribution $\overline{\Pi}$ and the constrained relaxed posterior distributions $\widetilde{\Pi}_{\rho_k}$, $k \in \mathbb{N}$, are absolutely continuous, so they have densities $\overline{\pi}(\boldsymbol{\theta} \mid \mathbf{y})$ and $\widetilde{\pi}_{\rho_k}(\boldsymbol{\theta} \mid \mathbf{y})$, respectively. Write these as

$$\begin{aligned}
\overline{\pi}(\boldsymbol{\theta} \mid \mathbf{y}) &= CL(\boldsymbol{\theta} \mid \mathbf{y})\pi(\boldsymbol{\theta})\mathbf{1}_{\boldsymbol{\theta}\in\boldsymbol{\Theta}} \\
\widetilde{\pi}_{\rho_k}(\boldsymbol{\theta} \mid \mathbf{y}) &= C_{\rho_k} L(\boldsymbol{\theta} \mid \mathbf{y})\pi(\boldsymbol{\theta}) \exp\left(-\frac{\rho_k}{2}\mathsf{dist}(\boldsymbol{\theta}, \boldsymbol{\Theta})^2\right),
\end{aligned}$$

where C and C_{ρ_k} are normalizing constants. Since $\mathbf{1}_{\boldsymbol{\Theta}} \le \exp\left(-\frac{\rho}{2}\mathsf{dist}(\boldsymbol{\theta}, \boldsymbol{\Theta})^2\right)$ for all $\boldsymbol{\theta} \in \mathbb{R}^d$ (with equality on $\boldsymbol{\Theta}$), we have

$$\begin{aligned}
C_{\rho_k}^{-1} &= \int_{\mathbb{R}^d} L(\boldsymbol{\theta} \mid \mathbf{y})\pi(\boldsymbol{\theta}) \exp\left(-\frac{\rho_k}{2}\mathsf{dist}(\boldsymbol{\theta}, \boldsymbol{\Theta})^2\right) d\,\boldsymbol{\theta} \\
&\ge \int_{\boldsymbol{\Theta}} L(\boldsymbol{\theta} \mid \mathbf{y})\pi(\boldsymbol{\theta}) \exp\left(-\frac{\rho_k}{2}\mathsf{dist}(\boldsymbol{\theta}, \boldsymbol{\Theta})^2\right) d\,\boldsymbol{\theta} \\
&= \int_{\boldsymbol{\Theta}} L(\boldsymbol{\theta} \mid \mathbf{y})\pi(\boldsymbol{\theta})\mathbf{1}_{\boldsymbol{\theta}\in\boldsymbol{\Theta}} d\,\boldsymbol{\theta} \\
&= C^{-1}
\end{aligned}$$

Thus, $C_{\rho_k} \le C$ for all $k \in \mathbb{N}$. By a similar calculation, we have

$$\begin{aligned}
C_{\rho_{k+1}}^{-1} &= \int_{\mathbb{R}^d} L(\boldsymbol{\theta} \mid \mathbf{y})\pi(\boldsymbol{\theta}) \exp\left(-\frac{\rho_{k+1}}{2}\mathsf{dist}(\boldsymbol{\theta}, \boldsymbol{\Theta})^2\right) d\,\boldsymbol{\theta} \\
&\ge \int_{\mathbb{R}^d} L(\boldsymbol{\theta} \mid \mathbf{y})\pi(\boldsymbol{\theta}) \exp\left(-\frac{\rho_k}{2}\mathsf{dist}(\boldsymbol{\theta}, \boldsymbol{\Theta})^2\right) d\,\boldsymbol{\theta} \\
&= C_{\rho_k}^{-1}.
\end{aligned}$$

We then have $C_{\rho_{k+1}} \le C_{\rho_k} \le C$ for all $k \in \mathbb{N}$. Therefore, $C_{\rho_k} \uparrow C$. Partition $\mathbb{R}^d = \boldsymbol{\Theta} \cup \boldsymbol{\Theta}^C$, and observe that for $\boldsymbol{\theta} \in \boldsymbol{\Theta}^C$,

$$\widetilde{\pi}_{\rho_k}(\boldsymbol{\theta} \mid \mathbf{y}) \ge 0 = \overline{\pi}(\boldsymbol{\theta} \mid \mathbf{y}).$$

For $\boldsymbol{\theta} \in \boldsymbol{\Theta}$, for all $k \in \mathbb{N}$,

$$\overline{\pi}(\boldsymbol{\theta} \mid \mathbf{y}) = CL(\boldsymbol{\theta} \mid \mathbf{y})\pi(\boldsymbol{\theta}) \ge C_{\rho_k} L(\boldsymbol{\theta} \mid \mathbf{y})\pi(\boldsymbol{\theta}) = \widetilde{\pi}_{\rho_k}(\boldsymbol{\theta} \mid \mathbf{y})$$

Thus, we have

$$\boldsymbol{\Theta}^C = \{\boldsymbol{\theta} : \widetilde{\pi}_{\rho_k}(\boldsymbol{\theta} \mid \mathbf{y}) \ge \pi(\boldsymbol{\theta} \mid \mathbf{y})\}, \qquad \boldsymbol{\Theta} = \{\boldsymbol{\theta} : \widetilde{\pi}_{\rho_k}(\boldsymbol{\theta} \mid \mathbf{y}) \le \pi(\boldsymbol{\theta} \mid \mathbf{y})\}, \quad k \in \mathbb{N}.$$

Using this fact about the partition, we examine the TV distance:

$$\begin{aligned}
\|\widetilde{\Pi}_{\rho_k} - \overline{\Pi}\|_{\mathsf{TV}} &= \frac{1}{2}\int_{\mathbb{R}^d} |\widetilde{\pi}_{\rho_k}(\boldsymbol{\theta} \mid \mathbf{y}) - \overline{\pi}(\boldsymbol{\theta} \mid \mathbf{y})|\, d\boldsymbol{\theta} \\
&= \frac{1}{2}\int_{\boldsymbol{\Theta}^C} \widetilde{\pi}_{\rho_k}(\boldsymbol{\theta} \mid \mathbf{y}) - \overline{\pi}(\boldsymbol{\theta} \mid \mathbf{y})\, d\boldsymbol{\theta} + \frac{1}{2}\int_{\boldsymbol{\Theta}} \overline{\pi}(\boldsymbol{\theta} \mid \mathbf{y}) - \widetilde{\pi}_{\rho_k}(\boldsymbol{\theta} \mid \mathbf{y})\, d\boldsymbol{\theta} \\
&= \underbrace{\frac{1}{2}\int_{\boldsymbol{\Theta}^C} \widetilde{\pi}_{\rho_k}(\boldsymbol{\theta} \mid \mathbf{y})\, d\boldsymbol{\theta}}_{(a)} + \underbrace{\frac{1}{2}(C - C_{\rho_k})\int_{\boldsymbol{\Theta}} L(\boldsymbol{\theta} \mid \mathbf{y})\pi(\boldsymbol{\theta})\, d\boldsymbol{\theta}}_{(b)}
\end{aligned}$$

For the first integral (a), observe that

$$\lim_{k\to\infty} \exp\left(-\frac{\rho_k}{2}\mathsf{dist}(\boldsymbol{\theta}, \boldsymbol{\Theta})^2\right) = \mathbf{1}_{\boldsymbol{\theta}\in\boldsymbol{\Theta}}$$

in a pointwise manner for all $\boldsymbol{\theta} \in \mathbb{R}^d$, so for all $\boldsymbol{\theta} \in \mathbb{R}^d$, $\widetilde{\pi}_{\rho_k}(\boldsymbol{\theta} \mid \mathbf{y}) \downarrow 0$ as $k \to \infty$. Note that

$$\widetilde{\pi}_{\rho_k}(\boldsymbol{\theta} \mid \mathbf{y}) = C_{\rho_k} L(\boldsymbol{\theta} \mid \mathbf{y}) \pi(\boldsymbol{\theta}) \exp\left(-\frac{\rho_k}{2} \mathsf{dist}(\boldsymbol{\theta}, \boldsymbol{\Theta})^2\right) \leq C L(\boldsymbol{\theta} \mid \mathbf{y}) \pi(\boldsymbol{\theta}) \in L^1$$

by Assumption 2. Therefore, by the Monotone Convergence Theorem, we have

$$\lim_{k \to \infty} \frac{1}{2} \int_{\boldsymbol{\Theta}^C} \widetilde{\pi}_{\rho_k}(\boldsymbol{\theta} \mid \mathbf{y}) \, d\boldsymbol{\theta} = \frac{1}{2} \int_{\boldsymbol{\Theta}^C} \lim_{k \to \infty} \widetilde{\pi}_{\rho_k}(\boldsymbol{\theta} \mid \mathbf{y}) \, d\boldsymbol{\theta} = 0.$$

For the second integral (b), we have:

$$\lim_{k \to \infty} \frac{1}{2}(C - C_{\rho_k}) \int_{\boldsymbol{\Theta}} L(\boldsymbol{\theta} \mid \mathbf{y}) \pi(\boldsymbol{\theta}) \, d\boldsymbol{\theta} = 0$$

Putting this all together, we obtain

$$\|\widetilde{\Pi}_{\rho_k} - \overline{\Pi}\|_{\mathsf{TV}} \to 0 \quad \text{as} \quad k \to \infty.$$

This shows that $\widetilde{\Pi}_{\rho_k}$ converges to Π in total variation distance. An equivalent way of writing this is to say that

$$\lim_{k \to \infty} \sup_{\mathcal{S} \in \mathcal{B}(\mathbb{R}^d)} |\widetilde{\Pi}_{\rho_k}(\mathcal{S}) - \overline{\Pi}(\mathcal{S})| = 0,$$

where we take the supremum over all Borel sets $\mathcal{B}(\mathbb{R}^d)$. Let $\mathcal{A} \subset \mathcal{B}(\mathbb{R}^n)$ be the collection of $\overline{\Pi}$-continuity sets (i.e., for all $A \in \mathcal{A}$, $\overline{\Pi}(\partial A) = 0$). Clearly,

$$\lim_{k \to \infty} \sup_{A \in \mathcal{A}} |\widetilde{\Pi}_{\rho_k}(A) - \overline{\Pi}(A)| \leq \lim_{k \to \infty} \sup_{A \in \mathcal{B}(\mathbb{R}^d)} |\widetilde{\Pi}_{\rho_k}(A) - \overline{\Pi}(A)| = 0.$$

Then by the Portmanteau Theorem (Billingsley, 2013, Theorem 2.1), we conclude that $\widetilde{\Pi}_{\rho_k} \xrightarrow{D} \overline{\Pi}$.

□

B.4 Proof of Theorem 3

Theorem 3. *Suppose that $\mathbb{E}_{\boldsymbol{\theta} \sim \pi(\boldsymbol{\theta} \mid \mathbf{y})}[\mathsf{dist}(\boldsymbol{\theta}, \boldsymbol{\Theta})^2/2] > D$. Then the constraint-relaxed posterior distribution $\widetilde{\pi}(\boldsymbol{\theta} \mid \mathbf{y})$ is the solution to the moment-constrained information projection problem:*

$$p^*(\boldsymbol{\theta}) \propto \pi(\boldsymbol{\theta} \mid \mathbf{y}) \exp\left(-\frac{\lambda}{2} \mathsf{dist}(\boldsymbol{\theta}, \boldsymbol{\Theta})^2\right),$$

where $\lambda > 0$ is a Lagrange multiplier that satisfies the moment constraint under $p^*(\boldsymbol{\theta})$.

Proof. The general form of the moment-constrained information projection problem can be stated as follows:

$$\begin{aligned} &\min_p \mathrm{KL}(p \,\|\, q) := \int p \log\left(\frac{p}{q}\right) \\ &\text{s.t. } \mathbb{E}_p[g(X)] = g_0, \end{aligned}$$

where p and q are density functions, $X \in \mathbb{R}^m$ is a random vector, $g : \mathbb{R}^m \to \mathbb{R}^d$ is a function of X, and $g_0 \in \mathbb{R}^d$ is some given vector. The optimal solution is given by West (2020) to be

$$p^*(x) \propto q(x) \exp\left(\lambda^\intercal g(x)\right).$$

See also Csiszár (1975) and Robertson et al. (2005). The optimal form of the solution is known as exponential tilting, and one can solve for the vector λ using the moment constraint:

$$\mathbb{E}_{p^*}[g(X)] = \int_{\mathbb{R}^m} g(x) q(x) \exp\left(\lambda^\intercal g(x)\right) \, dx = 0.$$

Specializing to our case, we take $x = \boldsymbol{\theta}$, $q(\boldsymbol{\theta}) = \pi(\boldsymbol{\theta} \mid \mathbf{y})$, $g(\boldsymbol{\theta}) = \frac{1}{2}\mathsf{dist}(\boldsymbol{\theta}, \boldsymbol{\Theta})^2$, and $g_0 = D$. This gives

$$p^*(\boldsymbol{\theta}) \propto \pi(\boldsymbol{\theta} \mid \mathbf{y}) \exp\left(\frac{\lambda}{2}\mathsf{dist}(\boldsymbol{\theta}, \boldsymbol{\Theta})^2\right),$$

where $\lambda \in \mathbb{R}$. Moreover, note that by (Tallman and West, 2022, Equation 5) and the remark immediately proceeding it, we have the relationship

$$\frac{\partial}{\partial \lambda}\mathbb{E}_{p^*}\left[\frac{1}{2}\mathsf{dist}(\boldsymbol{\theta}, \boldsymbol{\Theta})^2\right] > 0.$$

Moreover, observe that $\lambda = 0$ if and only if $p^*(\boldsymbol{\theta}) = \pi(\boldsymbol{\theta} \mid \mathbf{y})$. Thus, if we assume $\mathbb{E}_{\pi(\boldsymbol{\theta}|\mathbf{y})}[\mathsf{dist}(\boldsymbol{\theta}, \boldsymbol{\Theta})^2/2] > D$, then $\lambda < 0$. Reparameterizing, we have

$$p^*(\boldsymbol{\theta}) \propto \pi(\boldsymbol{\theta} \mid \mathbf{y}) \exp\left(-\frac{\lambda}{2}\mathsf{dist}(\boldsymbol{\theta}, \boldsymbol{\Theta})^2\right),$$

where $\lambda > 0$, and this gives the desired result. □

B.5 Proof of Proposition 2

Proposition 2. *The log constraint-relaxed posterior* $\log \widetilde{\pi}(\boldsymbol{\theta} \mid \mathbf{y})$ *is continuously differentiable as long as the log-posterior* $\log \pi(\boldsymbol{\theta} \mid \mathbf{y})$ *is continuously differentiable in* $\boldsymbol{\theta}$.

Proof. By Assumption 3, $\boldsymbol{\Theta}$ is convex, so $P_{\boldsymbol{\Theta}}(\boldsymbol{\theta})$ is single-valued. We then have $\nabla_{\boldsymbol{\theta}}\left[\frac{1}{2}\mathsf{dist}(\boldsymbol{\theta}, \boldsymbol{\Theta})^2\right] = \boldsymbol{\theta} - P_{\boldsymbol{\Theta}}(\boldsymbol{\theta})$ (Lange, 2016) for any $\boldsymbol{\theta}$.

It remains to show that the gradient is continuous. It suffices to show that the projection operator $P_{\boldsymbol{\Theta}}(\boldsymbol{\theta})$ is continuous. One way to see this is to observe that $P_{\boldsymbol{\theta}}(\boldsymbol{\theta})$ is firmly nonexpansive (i.e., 1-Lipschitz) when $\boldsymbol{\Theta}$ is closed and convex, and we immediately draw the conclusion.

However, we record here a slightly more general result that is useful for more general settings. For this we define the term Chebyshev set (Wulbert, 1968); a set C is Chebyshev if for all $x \in C$, $P_C(x)$ is a singleton. Theorem 3 of Wulbert (1968) provides one characterization of continuity which occurs for Chebyshev sets: if $C \subset X$ is a locally compact, Chebyshev set in a Banach space, then P_C is continuous if and only if C is convex.

$\boldsymbol{\Theta} \subset \mathbb{R}^n$ is convex, and so it is Chebyshev, and $\mathbb{R}^n$ is a Banach space. Moreover, it is easy to check that $\mathbb{R}^n$ is locally compact and Hausdorff. Since $\boldsymbol{\Theta}$ is closed, Corollary 29.3 of Munkres (2000) gives us that $\boldsymbol{\Theta}$ is locally compact. Thus, by the above theorem, we can conclude that $P_{\boldsymbol{\Theta}}$ is continuous.

Thus, for any point $\boldsymbol{\theta}_0 \subset \boldsymbol{\Theta}$,

$$\lim_{\boldsymbol{\theta} \to \boldsymbol{\theta}_0} \nabla_{\boldsymbol{\theta}}\left[\frac{1}{2}\mathsf{dist}(\boldsymbol{\theta}, \boldsymbol{\Theta})^2\right] = \lim_{\boldsymbol{\theta} \to \boldsymbol{\theta}_0} \left[\boldsymbol{\theta} - P_{\boldsymbol{\Theta}}(\boldsymbol{\theta})\right] = \boldsymbol{\theta}_0 - P_{\boldsymbol{\Theta}}(\boldsymbol{\theta}_0)$$

We also have that $\mathsf{dist}(\boldsymbol{\theta}, \boldsymbol{\Theta})$ is identically 0 on $\boldsymbol{\Theta}$, so it's gradient there will be 0. In particular, if $\boldsymbol{\theta}_0 \in \partial\boldsymbol{\Theta} \subset \boldsymbol{\Theta}$, then

$$\lim_{\boldsymbol{\theta} \to \boldsymbol{\theta}_0} \nabla_{\boldsymbol{\theta}}\left[\frac{1}{2}\mathsf{dist}(\boldsymbol{\theta}, \boldsymbol{\Theta})^2\right] = \boldsymbol{\theta}_0 - P_{\boldsymbol{\Theta}}(\boldsymbol{\theta}_0) = \boldsymbol{\theta}_0 - \boldsymbol{\theta}_0 = 0 = \nabla_{\boldsymbol{\theta}}\left[\frac{1}{2}\mathsf{dist}(\boldsymbol{\theta}, \boldsymbol{\Theta})^2\right]\Big|_{\boldsymbol{\theta} = \boldsymbol{\theta}_0 \in \partial\boldsymbol{\Theta}}$$

□

C HMC IMPLEMENTATION, STOCHASTIC ORDERING CASE STUDY

We outline the details of the HMC sampler used in the contingency table application. Suppose the prior distribution is

$$\theta_{[i]} \overset{\perp\!\!\!\perp}{\sim} \mathsf{Dir}(\alpha_{[i]}), \quad i \in [I],$$

where the entries of $\alpha_{[i]}$ are positive. By a small abuse of notation, let us also denote the matrix of θ_{ij} by θ. Also, let Δ denote the $(J+1)$-simplex. Then the posterior distribution with the distance-to-set prior is given by

$$\pi(\theta \mid n_{[1]}, \ldots, n_{[I]}) \propto \left(\prod_{i=1}^{I}\prod_{j=1}^{J}\theta_{ij}^{n_{ij}}\right)\left(\prod_{i=1}^{I}\prod_{j=1}^{J}\theta_{ij}^{\alpha_{ij}-1}\right)\exp\left(-\frac{\rho}{2}\mathsf{dist}(\theta,\Theta_{CT})\right)\mathbf{1}_{\bigcap_{i=1}^{I}\{\theta_{[i]}\in\Delta\}}$$
$$\propto \left(\prod_{i=1}^{I}\prod_{j=1}^{J}\theta_{ij}^{n_{ij}+\alpha_{ij}-1}\right)\exp\left(-\frac{\rho}{2}\|\theta - P_{\Theta_{CT}}(\theta)\|_F^2\right)\mathbf{1}_{\bigcap_{i=1}^{I}\{\theta_{[i]}\in\Delta\}}$$

As mentioned above, given some θ, $P_{\Theta_{CT}}(\theta)$ can be computed using the `quadprog` library in `R`. For the HMC algorithm, the (negative) potential energy is defined by the log-posterior, which is given by

$$\log\pi(\theta \mid n_{[1]}, \ldots, n_{[I]}) = C + \sum_{i=1}^{I}\sum_{j=1}^{J}(n_{ij}+\alpha_{ij}-1)\log\theta_{ij} - \frac{\rho}{2}\|\theta - P_{\Theta_{CT}}(\theta)\|_F^2 + \log\left(\mathbf{1}_{\bigcap_{i=1}^{I}\{\theta_{[i]}\in\Delta\}}\right),$$

where C is a constant that does not depend on θ. Additionally, we need to compute the gradient of the (negative) log-potential with respect to θ. We set $\theta_{iJ} = 1 - \sum_{j=1}^{J-1}\theta_{ij}$ for $i \in [I]$. Let us denote the matrix θ with the J^{th} column removed by $\widetilde{\theta}$ and the corresponding posterior by $\widetilde{\pi}$. Denote the space of $I \times (J-1)$ matrices $\widetilde{\theta}$ that satisfy the desired isotonic constraint by

$$\widetilde{\Theta}_{CT} := \left\{(\theta_{ij})_{i\in[I],j\in[J-1]} \,\middle|\, \sum_{k=1}^{j}\theta_{i+1,k} \geq \sum_{k=1}^{j}\theta_{ik} \text{ for } i \in [I], j \in [J-1]\right\}$$

Note that the entries of elements of $\widetilde{\Theta}_{CT}$ are non-negative. However, the rows do not sum to less than 1. Then the reduced (negative) log-potential becomes:

$$\log\widetilde{\pi}(\widetilde{\theta} \mid n_{[1]}, \ldots, n_{[I]}) = C + \underbrace{\sum_{i=1}^{I}\left((n_{iJ}+\alpha_{iJ}-1)\log\left(1-\sum_{j=1}^{J-1}\theta_{ij}\right) + \sum_{j=1}^{J-1}(n_{ij}+\alpha_{ij}-1)\log\theta_{ij}\right)}_{=:f(\widetilde{\theta})}$$
$$- \frac{\rho}{2}\|\widetilde{\theta} - P_{\widetilde{\Theta}_{CT}}(\widetilde{\theta})\|_F^2$$

We compute the gradient of the first term component-wise and obtain

$$\frac{\partial f(\widetilde{\theta})}{\partial\theta_{ij}} = \frac{n_{ij}+\alpha_{ij}-1}{\theta_{ij}} - \frac{n_{iJ}+\alpha_{iJ}-1}{1-\sum_{j=1}^{J-1}\theta_{ij}}$$

To compute the gradient of the second (penalty) term, we note that the projection may be trickier because of the above substitution for θ_{iJ}. However, we show this is not the case. Define $\widehat{\Theta}_{CT}$ to be the canonical embedding of $\widetilde{\Theta}_{CT}$ into the space of $I \times J$ matrices where the isotonic constraint only holds for the first $J-1$ columns in each row; more formally, $\widetilde{\Theta}_{CT} \hookrightarrow \widehat{\Theta}_{CT}$, and

$$\widehat{\Theta}_{CT} := \left\{(\theta_{ij})_{i\in[I],j\in[J]} \,\middle|\, \sum_{k=1}^{j}\theta_{i+1,k} \geq \sum_{k=1}^{j}\theta_{ik} \text{ for } i \in [I], j \in [J-1]\right\}$$

Note that since each row of θ is a probability distribution, we have $\sum_{k=1}^{J}\theta_{ik} = 1$ holds for all $i \in [I]$. Thus, the following condition holds trivially for all $i \in [I]$

$$\sum_{k=1}^{J}\theta_{i+1,k} \geq \sum_{k=1}^{J}\theta_{ik}.$$

Thus, we immediately obtain $\widehat{\Theta}_{CT} = \Theta_{CT}$. As a corollary, projecting onto $\widehat{\Theta}_{CT}$ is equivalent to simply projecting onto the original parameter space Θ_{CT} because the last column of θ, which we removed with our substitution, does not have an impact on the projection. This means we can project onto $\widetilde{\Theta}_{CT}$ by projecting onto Θ_{CT} and ignoring the last column of elements. Finally, we ensure that each row of θ sums to 1 manually once we have determined a proposal for the first $J-1$ columns as described above.The gradient of the second term (the penalty term) is obtained then as:

$$\nabla_\theta\left(\frac{\rho}{2}\mathsf{dist}(\widetilde{\theta},\widetilde{\Theta}_{CT}\right) = \nabla_\theta\left(\frac{\rho}{2}\mathsf{dist}(\theta,\Theta_{CT})^2\right) = \rho(\theta - P_{\Theta_{CT}}(\theta)),$$

where we ignore the J^{th} column of θ. We can use the potential function and the gradient to obtain proposals θ; we ensure that each row of θ sums to 1 manually once we have determined a proposal for the first $J-1$ columns using the procedure described above.

Next, it is worth noting that the first term of the gradient has properties akin to the log-barrier function technique (Boyd and Vandenberghe, 2004). In particular, note that

$$\lim_{\theta_{ij} \to 0^+} \frac{\partial f(\widetilde{\theta})}{\partial \theta_{ij}} = +\infty \qquad \text{and} \qquad \lim_{\theta_{ij} \to 1-} \frac{\partial f(\widetilde{\theta})}{\partial \theta_{ij}} = -\infty.$$

Thus, our sampler will deviate away from the boundary of the simplex that supports the Dirichlet prior. However, due to the fact that in practice we use a discretized leapfrog integrator to solve the Hamilton PDEs along with the fact that the gradients can be wildly different for each θ_{ij} but we have a constant scalar ϵ step-size, it may be that the behavior of $f(\widetilde{\theta})$ may overshoot the simplex when updating the position and momentum in an HMC proposal. As a result, we address this using a rather simple approach by rejecting any proposals that fall outside of the simplex. We also terminate any position-momentum updates that by chance fall outside of the simplex since the log-posterior is not defined outside of the simplex. While these solutions may offer some recourse with implementing this sampler, more sophisticated techniques, such as Spherical HMC Lan et al. (2014) or through spherical transformation of the simplex (Betancourt, 2012), may result in an improved implementation. However, the main purpose here is to demonstrate how one may incorporate a distance-to-set prior rather than construct the best-performing sampler for this particular problem.