Constrained Weighted Bayesian Bootstrap

Sam Rosen[1] **Jason Xu**[2]

[1]Department of Statistical Science, Duke University
[2]Department of Biostatistics, University of California, Los Angeles

Abstract

We prove the weighted Bayesian bootstrap, a method for approximate sampling of a posterior distribution, can be extended to sample from general constrained posterior distributions under mild assumptions. The method entails a simple algorithm that can take advantage of fast tools from convex optimization. Under regularity conditions, we show the asymptotic distribution of samples from the constrained weighted Bayesian bootstrap has a covariance matching the restricted maximum likelihood estimator, an efficient estimator. We assess the method empirically on a variety of constrained Bayesian problems, demonstrating broad applicability of the method as well as advantages over existing peer methods. The constrained weighted Bayesian bootstrap quickly samples from constrained posteriors, providing adequate uncertainty quantification for problems typically solved via optimization methods designed to deliver only a point estimate. As a case study, using constraints required in European-style option prices, uncertainty estimates of an option pricing surface are derived with constrained weighted Bayesian bootstrap.

1 INTRODUCTION

The canonical formulation of statistical learning tasks from an optimization lens seeks to maximize a measure of fit subject to a constraint or regularization term imposing desired structure. Accordingly, there is a vast literature with classical theory and efficient algorithms for solving constrained problems. Generally, the solutions delivered by these methods provide a point estimate of the parameters of interest. Constructing confidence intervals can quickly become complicated by common constraints such as sparsity, order or shape restrictions. Uncertainty estimates could then necessitate post-selection inference procedures, deviate from standard asymptotic theory, or otherwise require case-by-case treatment depending on the constraint structure [Lee et al., 2016, Panigrahi and Taylor, 2018].

Alternatively, Bayesian methods yield straightforward uncertainty quantification via the posterior distribution, but the computation and sampling procedure itself is often complicated by the presence of the constraints. When fitting a probabilistic model to data, the underlying parameters may require restriction to a subspace, $\tilde{\Theta} \subset \Theta$, of a larger parameter space. Exact sampling of a posterior distribution with support restricted to a general set $\tilde{\Theta}$ becomes difficult, even when the unconstrained posterior is well-defined. Popular generic sampling methods such as Hamiltonian Monte Carlo (HMC) cannot be easily used when imposing the constraint results in a lack of smoothness, and variants for special cases such as restrictions along manifolds entail higher computational overhead [Girolami and Calderhead, 2011]. Sampling from the unconstrained posterior and then discarding samples outside the constraint set *post hoc* will give valid samples, but can be extremely inefficient; in many cases $\tilde{\Theta}$ may be a low-probability or even zero-measure subspace of the posterior [Xu and Duan, 2023].

Existing literature for sampling a posterior distribution with restricted support varies in the underlying mechanism. As a reasonable heuristic, transformation-based approaches sample from unconstrained posteriors and transform the samples to the constrained space *ex post facto*, such as with an orthogonal projection [Astfalck et al., 2024], or an oblique projection determined from the posterior [Dunson and Neelon, 2003, Everink et al., 2023]. Orthogonal projections may ignore the geometry of the posterior and in cases where the posterior covariance or Fisher information must be estimated, oblique projections can be sensitive to this uncertainty. Specific prior forms encourage θ samples to be in $\tilde{\Theta}$ by heavily down-weighting the prior outside the constraint set with a hyperparameter controlling the amount of relaxation [Duan et al., 2020, Presman and Xu, 2023]; however,

Proceedings of the 42nd Conference on Uncertainty in Artificial Intelligence (UAI 2026), PMLR 337:5740–5763.

this may lead to samples that are close to, but not exactly in the constraint set. Although inexact, these methods are designed to interface nicely with gradient-based sampling methods. Some specialized priors take advantage of the robustness of the proximal mapping used in constrained optimization [Zhou et al., 2024, Xu et al., 2024]. Application of this principled approach is limited to constraint sets with the ability to repeatedly calculate proximal mappings quickly. Altogether, constrained sampling with these methods has various trade-offs, and many share a connection to an idea from convex optimization.

In this work, we expand upon the weighted likelihood bootstrap first introduced in Newton and Raftery [1994], which transforms a posterior sampling problem into an optimization problem (see Section 2 for details). The weighted likelihood bootstrap has been generalized to a variety of contexts with success such as multimodal posteriors [Fong et al., 2019], nonparametric learning [Lyddon et al., 2019], and latent variable models [Han and Yang, 2019]. In more recent works, where the prior form is explicitly considered, it is referred to as the weighted Bayesian bootstrap [Newton et al., 2021]. The weighted Bayesian bootstrap is effective with lasso-based estimators for both sparse normal mean and high-dimensional regression problems with compelling theoretical properties such as concentration towards the true underlying parameters [Ng and Newton, 2022]. The work of Nie and Ročková [2023] produces samples of the spike-and-slab lasso posterior [Ročková and George, 2018] using the weighted Bayesian bootstrap while proving a minimax optimal rate of concentration. Similar ideas lead to an expansion to handle binary responses in Menacher et al. [2024].

Given the weighted likelihood/Bayesian bootstrap replaces standard posterior sampling using optimization steps, we propose a simple adaptation to enable sampling from posteriors with constrained support. This is especially salient as constraints can significantly hinder sampling-based posterior computation, while there is rich literature on constrained optimization. The weighted Bayesian bootstrap has shown robust efficacy both empirically and theoretically in a variety of domains, and this article will additionally show that many of these properties carry over to the constrained regime. Furthermore, the ease of implementation is shown for several examples, along with the difficulties of sampling in constrained spaces for other modern Bayesian posterior sampling methods[1].

2 METHODS

We consider data generated from a probability density function f_θ, with $\theta \in \Theta$ and prior beliefs encoded in $\pi(\theta)$. The standard form for the posterior distribution of θ given the underlying data follows

$$\pi(\theta \mid x) \propto f_\theta(x)\pi(\theta). \tag{1}$$

By restricting $\pi(\theta)$ to have positive support only on a set $\tilde{\Theta} \subset \Theta$, the *constrained* posterior distribution can be written up to proportionality

$$\tilde{\pi}(\theta \mid x) \propto \pi(\theta \mid x) \times \mathbf{1}(\theta \in \tilde{\Theta}), \tag{2}$$

where $\mathbf{1}$ denotes the 0–1 indicator function.

As surveyed in the introduction, it is not straightforward to account for the indicator constraint in generality. To make progress, we revisit an idea introduced in Newton and Raftery [1994] that performs approximate sampling efficiently when the *maximum a posteriori* estimate of (1) is available. They propose the weighted likelihood bootstrap, where samples are produced by first generating weights $w_n \sim n \times \text{Dirichlet}(1, \ldots, 1)$, and then maximizing a weighted log-likelihood function,

$$\arg\max_{\theta \in \Theta} \sum_{i=1}^{n} w_{n,i}\ell(x_i \mid \theta), \tag{3}$$

where $\ell(x_i \mid \theta) = \log f_\theta(x_i)$. These samples are shown to be first-order correct, sharing an asymptotic conditional distribution with the posterior distribution under a variety of priors. This may also be called the weighted Bayesian bootstrap when prior terms (also possibly weighted) are appended to (3) [Newton et al., 2021].

Newton and Raftery [1994] posits repeated perturbations of the maximum likelihood equation on a single dataset create a path toward sampling. Motivated by the well developed literature and breadth of tools for constrained problems in the optimization literature, we devise the constrained weighted Bayesian bootstrap (CWBB), a variant of the weighted likelihood bootstrap designed to address (2). To incorporate the prior constrained to $\tilde{\Theta}$, we add the log-prior term to (3) and write an equivalent form with an equality constraint

$$\arg\max_{\theta \in \Theta} \sum_{i=1}^{n} w_{n,i}\ell(x_i \mid \theta) + \log \pi(\theta), \quad h(\theta) = 0, \tag{4}$$

where $h \colon \mathbb{R}^p \mapsto \mathbb{R}^r$, $h(\theta) = 0$ if and only if $\theta \in \tilde{\Theta}$, and h satisfies regularity conditions such as continuous second derivatives in a neighborhood of θ_0. CWBB with this optimization problem is summarized in Algorithm 1.

Algorithm 1. *Constrained weighted Bayesian bootstrap.*

For $t = 1$ *to number of desired samples*
  Sample $w_n \sim n \times$ *Dirichlet*$(1, \ldots, 1)$.
  Set $\theta^{(t)}$ *according to* (4).
Output $\theta^{(\cdot)}$

[1] All code used to produce the experimental results and figures is found at `https://github.com/SamGRosen/CWBB`.

Illustrative example Consider the Bayesian linear regression where coefficients are restricted to the unit sphere in $\mathbb{R}^p$ after a linear transformation:

$$\begin{aligned} y \mid X, \beta &\sim N(X\beta, \tau_\epsilon^{-1} I_n), \\ \beta &\sim N(0, \tau_\beta^{-1} I_p) \times \mathbf{1}(\|A\beta\|_2^2 = 1), \end{aligned} \tag{5}$$

where $A \in \mathbb{R}^{p \times p}$. Denoting the ith row of X by x_i and ignoring proportionality constants, we can write (4) as

$$\underset{\|A\beta\|_2^2=1}{\arg\min} \frac{\tau_\epsilon}{2} \sum_{i=1}^{n} w_{n,i} \{y_i - x_i^\top \beta\}^2 + \frac{\tau_\beta}{2} \|\beta\|_2^2. \tag{6}$$

Let $\tilde{y}_i = w_{n,i}^{1/2} y_i$ and $\tilde{X} = \mathrm{diag}(w_n)^{1/2} X$. The Lagrangian of (6) and resulting gradient is

$$\begin{aligned} \mathcal{L}(\beta, \lambda) &= \frac{\tau_\epsilon}{2} \|\tilde{y} - \tilde{X}\beta\|_2^2 + \frac{\tau_\beta}{2} \|\beta\|_2^2 + \lambda(\|A\beta\|_2^2 - 1), \\ \nabla_\beta \mathcal{L}(\beta, \lambda) &= -\tau_\epsilon \tilde{X}^\top (\tilde{y} - \tilde{X}\beta) + \tau_\beta \beta + 2\lambda A^\top A \beta. \end{aligned} \tag{7}$$

Setting (7) to zero for stationarity and solving for β gives

$$\beta = \{\tilde{X}^\top \tilde{X} + \tau_\epsilon^{-1}(\tau_\beta I_p + 2\lambda A^\top A)\}^{-1} \tilde{X}^\top \tilde{y}. \tag{8}$$

Hence, the p-dimensional optimization problem is reduced to a one-dimensional root-finding problem for the Lagrange multiplier, λ, that satisfies both $\|A\beta\|_2^2 = 1$ and (8). In the unconstrained case, computation in the weighted Bayesian bootstrap reduces to calculating ridge regression solutions for randomly weighted data. In the constrained case, CWBB operates similarly, but the regularization parameter automatically adapts to debias the solutions towards the constraint set across iterations.

Figure 1 shows samples from several peer methods for sampling from the constrained posterior of (5) with $n = 100$, $p = 2$ and $A = I_2$. Exact constrained samples concentrate on the unit circle around the true value β_0. Both methods which relax the constraint lead to separate issues: the distance-to-set prior of Presman and Xu [2023] has samples very close to the constraint set, but still noticeably far despite a high regularization value of $\rho = 5000$; the constraint relaxation from Duan et al. [2020] generates samples far from regions of high posterior density induced by a local mode from the nonconvex constraint set. Ignoring the geometry of the posterior, orthogonally projected samples may project to low density regions of the posterior. This is ameliorated in the obliquely projected samples since the posterior covariance is known. Meanwhile, CWBB respects the constraint set, and we see that its samples most closely resemble the exact posterior density.

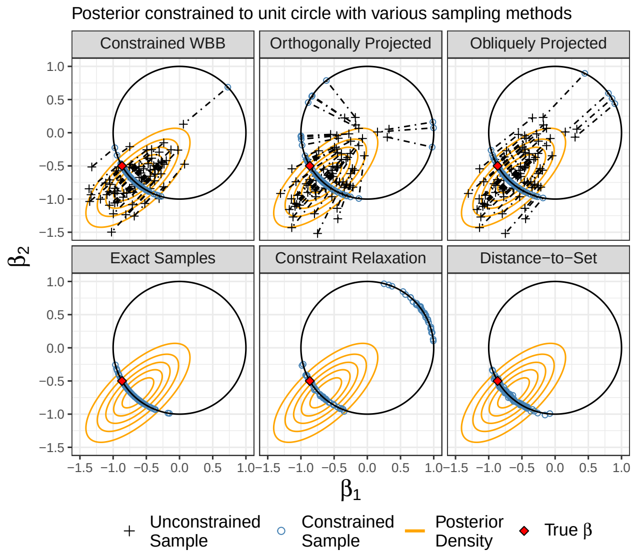


Figure 1: Graphs showing 75 samples of β restricted to the unit circle under six constrained sampling methods.

3 THEORY

We establish several favorable properties of the posterior samples under CWBB. Under standard regularity conditions detailed in the Supplementary Material, the solution to (4) will exist and be conditionally consistent for θ_0, along almost every sample path asymptotically. The regularity conditions are a natural extension to constraints for those used in analysis of the weighted likelihood bootstrap [Newton and Raftery, 1994]. In addition, we establish CWBB has a clear asymptotic distribution that gives valid frequentist confidence intervals. The restricted maximum likelihood estimator from Aitchison and Silvey [1958],

$$\theta_n^* = \underset{\theta \in \Theta}{\arg\max} \sum_{i=1}^{n} \log f_\theta(x_i), \quad h(\theta) = 0, \tag{9}$$

is asymptotically efficient, achieving the Cramér-Rao lower bound under parametric constraints [Stoica and Ng, 1998]. We establish that CWBB samples converge in distribution to a multivariate normal centered on this estimator with a matching covariance expression. To this end, we assume that h has continuous bounded second derivatives in a neighborhood of θ_0, and full-rank Jacobian at θ_0. We note that these conditions do not preclude $h(\theta)$ taking the form of an inequality constraint, but we focus our treatment on equality constraints without loss of generality. Inactive inequality constraints have no effect on the asymptotic distribution while active inequality constraints can be subsumed into the equality constraints.

Theorem 1. *Let $X_{1:n}$ be independent and identically distributed observations with underlying density function f_{θ_0}. Let the Lagrangian of* (4) *be*

$$\mathcal{L}(\theta, \lambda) = \sum_{i=1}^{n} w_{n,i} \ell(x_i \mid \theta) + \log \pi(\theta) + \lambda^\top h(\theta).$$

Under regularity conditions, for all $\delta > 0$, there exists a *sequence* $\{\check{\theta}_n, \check{\lambda}_n\}$ *such that*

$$\mathrm{pr}\{\nabla \mathcal{L}(\check{\theta}_n, \check{\lambda}_n) = 0 \mid X_{1:n}\} \to 1 \quad a.s.[X_{1:\infty}],$$
$$\mathrm{pr}(\|\check{\theta}_n - \theta_0\|_2 < \delta \mid X_{1:n}) \to 1 \quad a.s.[X_{1:\infty}].$$

Theorem 2. *Let $X_{1:n}$ be independent and identically distributed observations with underlying density function f_{θ_0}. Let $\hat{\theta}_n$ be a strongly consistent estimator for θ_0, satisfying $h(\hat{\theta}_n) = 0$ and*

$$\left\| \frac{I - U(\hat{\theta}_n)}{n^{1/2}} \sum_{i=1}^{n} \nabla_\theta \log f_{\hat{\theta}_n}(X_i) \right\|_2 \to 0 \; a.s.[X_{1:\infty}] \quad (10)$$

where $D_h(\theta)$ is the Jacobian at θ for h,

$$J(\theta) = \mathbb{E}_{\theta_0}[\{\nabla_\theta \log f_\theta(X)\}\{\nabla_\theta \log f_\theta(X)\}^\top],$$
$$U(\theta) = D_h^\top(\theta)\{D_h(\theta_0)J(\theta_0)^{-1}D_h^\top(\theta)\}^{-1}D_h(\theta_0)J(\theta_0)^{-1}.$$

Let $\check{\theta}_n$ be a sample from Algorithm 1. Then, under regularity conditions, for any Borel set $A \in \mathbb{R}^p$,

$$\mathrm{pr}\big\{n^{1/2}(\check{\theta}_n - \hat{\theta}_n) \in A \mid X_{1:n}\big\} \to \mathrm{pr}(Z \in A) \; a.s.[X_{1:\infty}]$$

where $Z \sim N[0, J(\theta_0)^{-1}\{I - U(\theta_0)\}]$.

Note the restricted maximum likelihood estimator trivially satisfies (10) as demonstrated by the proof of Lemma 1 in Aitchison and Silvey [1958]; see the proof of Lemma 7 in the Supplementary Material for a similar proof.

Theorem 1 shows consistency of CWBB samples while Theorem 2 gives an asymptotic distribution for these samples. Theorem 2 additionally implies that the asymptotic covariance matrix is obtained by obliquely projecting the rows of the inverse Fisher information onto the kernel of $D_h(\theta_0)$, the Jacobian of the constraints at the true underlying parameter. This result is consistent with the asymptotic covariance of (9), giving the constrained form an explicit characterization for the uncertainty in approximate posterior samples. In addition, under these regularity conditions, the asymptotic distribution from Theorem 2 is invariant to the choice of h used to describe $\tilde{\Theta}$ (see Lemma 1 in the Supplementary Materials). Furthermore, we allow for $\tilde{\Theta}$ to be a lower dimensional subspace of Θ; this is not handled precisely in the analysis of constraint relaxation and is also lacking theoretical support in projecting the unconstrained posterior after sampling [Astfalck et al., 2024].

In particular, CWBB converges to a distribution with support only on the constraint set as the sample size grows. This is not the case for some other sampling methods proposed for (2), such as constraint-relaxed posteriors of the form

$$\pi_\rho(\theta \mid x_{1:n}) \propto \exp\left\{\sum_{i=1}^{n} \log f_\theta(x_i) - \rho q(\theta)\right\}, \quad (11)$$

where $q(\theta) = \mathrm{dist}^2(\theta, \tilde{\Theta})/2$ [Presman and Xu, 2023, Zhou et al., 2024] or more generally $q \geq 0$ such that $\theta \in \tilde{\Theta} \iff q(\theta) = 0$ [Duan et al., 2020]. While these approaches to posterior inference coincide with the constrained problem as the penalty parameter $\rho \to \infty$, this relationship with the sample size is not uniform, so it is difficult to establish the analogous posterior consistency theory. Asymptotically, under a fixed penalty $\rho > 0$, samples fall back to the standard Bernstein–von Mises theorem (see Chapter 10 of van der Vaart [1998]): if q is finite in a neighborhood of θ_0, then $n^{1/2}\theta \mid x_{1:n}$ converges to a distribution with variance $J(\theta_0)^{-1}$. This implies the constraints become less important to the posterior covariance and subsequent uncertainty estimation as more data is ingested, unless the relaxation adjusts to the sample size. In fact, using results on generalized posteriors from Miller [2021], Proposition 1 shows that if the relaxation strength changes linearly with n, then the asymptotic covariance is full-rank, regardless of the dimension of $\tilde{\Theta}$.

Proposition 1. *Suppose $\Theta \subset \mathbb{R}^p$, $\rho > 0$, and $q(\theta)$ is three-times continuously differentiable in a closed ball around θ_0. In addition, $f_\theta(x) = \exp\{\theta^\top s(x) - \kappa(\theta)\}$ is a full, regular, and identifiable member of the exponential family in natural form. Finally, the matrix $\Sigma = [-\mathbb{E}_{\theta_0}\{\nabla_\theta^2 \log f_{\theta_0}(X)\} + \rho \nabla_\theta^2 q(\theta_0)]^{-1}$ is positive-definite. Then if $\theta \sim \pi_{n\rho}$ from* (11) *there exists a strongly consistent estimator $\hat{\theta}_n$ such that*

$$d_{TV}\{n^{1/2}(\theta - \hat{\theta}_n), N(0, \Sigma)\} \to 0,$$

almost surely, where d_{TV} is the total variation distance.

For concreteness, Proposition 1 implies the constraint-relaxed posterior for many exponential family distributions with $\rho' = n\rho$ will have an asymptotic covariance slightly informed by the constraints. As an example, if $q(\theta) = \mathrm{dist}^2(\theta, \tilde{\Theta})/2$ where $\tilde{\Theta} = \{\theta \colon P\theta = \theta\}$ and P is an orthogonal projection matrix, then the asymptotic covariance is $\{J(\theta_0) + \rho(I - P)\}^{-1}$.

4 EXPERIMENTS

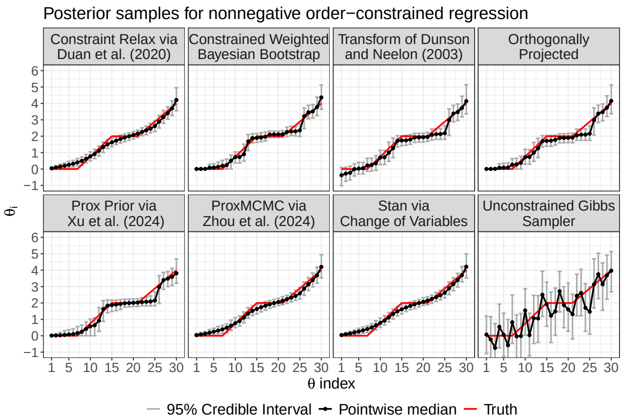
4.1 NONNEGATIVE ORDER-CONSTRAINED PARAMETERS IN REGRESSION

In Dunson and Neelon [2003], the authors describe a Bayesian linear model under the restriction where the learned coefficients are both nonnegative and nondecreasing according to underlying indices. This model is motivated by a biomedical study with ordered categorical predictors, e.g. increasing intensities of treatment regimes. This setting can be modeled with the Bayesian system

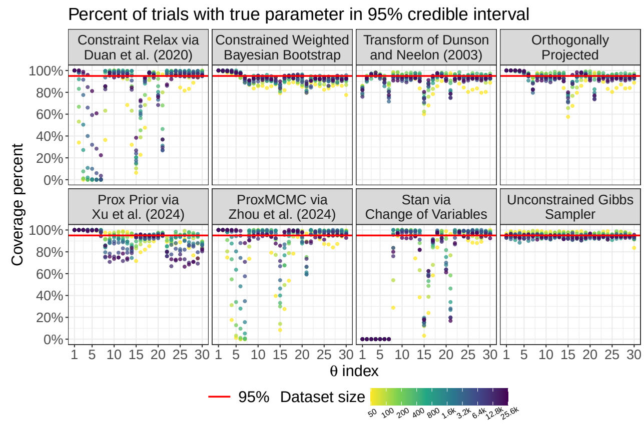
$$\begin{aligned} &y_i \mid \theta, \tau, x_i \sim N(x_i^\top \theta, \tau^{-1}), \\ &\theta \sim N(\theta_0, \Sigma_0), \quad \tau \sim \mathrm{Gamma}(a_0, b_0), \end{aligned} \quad (12)$$

where θ is restricted to the convex space

$$\mathcal{C} = \{\theta \in \mathbb{R}^p \colon 0 \leq \theta_1 \leq \theta_2 \leq \ldots \leq \theta_p\},$$



(2a) A single trial for sampling from (12) under several peer methods with $n = 100$.



(2b) Percent of trials with true θ_i in the 95% posterior credible intervals, for each θ_i, dataset size, and sampling method.

and x_i are covariates. Equivalently, there exists an invertible $D \in \mathbb{R}^{n\times n}$ such that $\theta \in \mathcal{C} \iff D\theta \geq 0$. Using arguments from Gelfand et al. [1992] surrounding constrained Gibbs sampling, Dunson and Neelon [2003] generalize a method from Hwang and Peddada [1994] to transform unconstrained posterior samples via

$$\hat{\theta}_i = \min_{i\leq t}\max_{s\leq i} \frac{\mathbf{1}_{t-s+1}^\top (\Sigma_{[s:t]})^{-1}\theta_{[s:t]}}{\mathbf{1}_{t-s+1}^\top (\Sigma_{[s:t]})^{-1}\mathbf{1}_{t-s+1}}, \qquad (13)$$

where Σ is the posterior covariance and $[s:t]$ is indexing for the sth to tth elements or submatrix. This does not guarantee transformed parameters are ordered or nonnegative, although the authors note this tends to be a nonissue in practice. The transformation requires solving a quadratic number of linear systems, accruing $O(p^4)$ operations to ensure that the parameter vector obeys the constraint.

CWBB guarantees samples satisfying the constraints much more efficiently. Each iteration samples $w_n \sim n \times \text{Dirichlet}(1,\dots,1)$ and solves

$$\begin{aligned}&\operatorname*{arg\,min}_{\theta\in\mathcal{C},\tau>0} -(n/2+a_0-1)\log\tau + b_0\tau \\ &+ \frac{\tau}{2}\sum_{i=1}^n w_{n,i}(y_i - x_i^\top\theta)^2 + \frac{1}{2}(\theta-\theta_0)^\top\Sigma_0^{-1}(\theta-\theta_0).\end{aligned}$$

Solutions to the optimization subproblem are delivered by coordinate descent on the θ and τ blocks. For fixed τ, a convex quadratic program with linear constraints emerges, giving a multitude of methods to optimize for θ; for fixed θ, the optimal τ is available in closed form.

We implement the above to sample from the posterior of (12) constrained to $\mathcal{C}$ using CWBB. Simulation parameters were in a regime with $p = 30$ and true parameters $\tau = 1/25$ and θ, a sequence of nondecreasing values with two constant contiguous subsequences (see Fig. 2a for an illustration). We sample each covariate vector, x_i, from a multivariate normal with mean zero and covariance matrix with unit diagonal, and $0.6, 0.3, 0.1$ on the main off-diagonals. For prior hyperparameters, θ_0 is the zero vector, $\Sigma_0 = 2I_p$ and $a_0, b_0 = 1$.

In addition, six other constrained sampling methods are used to sample this posterior, which are listed in Fig. 2a. Note that the Moreau-Yosida envelope of an orthogonal projection is the distance-to-set function, so sampling via ProxMCMC is equivalent to the distance-to-set prior of Presman and Xu [2023] in the case of priors with fixed constrained support. We also sample using the `positive_ordered` change of variables implemented in Stan [Stan Development Team, 2023], which leads to samples strictly in the interior of $\mathcal{C}$. This leads to difficulties in coverage as the true θ_0 is on the boundary of $\mathcal{C}$, a setting supported by CWBB according to the theory in Section 3.

With these implementations, 2500 total trials across ten different dataset sizes are simulated to determine coverage abilities of the constrained samples from a variety of methods. Figure 2a displays an example set of credible intervals for a single trial with $n = 100$. For each dataset and method, 95% credible intervals for each θ_i are constructed. The coverage percent is then the percent of trials which contain the true θ_i inside the respective credible interval. Figure 2b displays the coverage percents for all θ_i over a variety of dataset sizes for several constrained posterior sampling methods and an unconstrained Gibbs sampler as a baseline. Of the constrained sampling methods, CWBB has coverage most resembling that of the unconstrained Gibbs sampler, which does not respect the constraints and has much wider intervals, as seen in Fig. 2a. Many methods struggle with coverage where the true θ_i is flat or changes slope, with CWBB suffering the least from these difficulties. In the Supplementary Material, Table 1 has the total runtime to complete all 2500 trials, showing CWBB is orders of magnitude faster than several of the competing methods. In addition, see Fig. 5 to compare the minimum coverage percent of all θ_i for the methods, showing CWBB overall has the best worst-case performance as compared to the unconstrained Gibbs baseline.

4.2 SPARSE PRECISION MATRIX ESTIMATION

Assuming data x_i follow a multivariate normal distribution with an unknown sparse precision matrix, the maximum likelihood estimate of the precision matrix is its sample value. Sparsity is a useful characteristic of precision matrices allowing for explicit calculation of conditional dependence among the variables of x_i. To induce sparsity in the maximum likelihood estimate, the graphical lasso of Friedman et al. [2008] optimizes an ℓ_1 penalty added to the likelihood,

$$\log\det\Sigma^{-1} - \operatorname{tr}(S\Sigma^{-1}) - \rho\sum_{i\le j}|\Sigma^{-1}_{ij}|,$$

where S is the empirical covariance matrix. When a point estimate is desired, the graphical lasso is an effective tool, but uncertainty quantification around classification of zero entries can improve inference. Bayesian methods quantify uncertainty for precision matrix estimation with a variety of sparsity-inducing priors. Given samples from a posterior, a probability of edge inclusion (nonzero entry of precision matrix) is produced from the empirical distribution. The Bayesian graphical lasso of Wang [2012] uses a formulation similar to (14) where *maximum a posteriori* estimates are sparse and inclusion probabilities are estimated via partial correlations from the posterior samples. The setup of the Bayesian graphical lasso allows a prior over the possible regularization parameters, giving a smaller need to tune. Mohammadi and Wit [2015] uses the G-Wishart prior for more explicit inference of edge inclusion by directly sampling over the many possible graphical representations of variable conditional dependence. Posterior sampling is done via a birth-death process. Wang [2015] also considers a discrete approach via a spike-and-slab prior for all pairs of variables with sampling done using a block Gibbs sampler. ProxMCMC [Zhou et al., 2024] features sparse precision matrix estimation as a canonical example, where samples are constrained to an ℓ_1 ball after specifying a prior distribution over potential radii.

To sample from a posterior of sparse precision matrices, we apply CWBB by viewing the graphical lasso as the *maximum a posteriori* estimate of the following Bayesian system:

$$\begin{aligned} &x_i \mid \Sigma \sim N(0,\Sigma),\\ &\pi(\Sigma^{-1}) \propto \mathbf{1}(\Sigma^{-1}\in\mathbb{S}^{p\times p}_+)\prod_{i\le j}\exp(-\rho|\Sigma^{-1}_{ij}|), \end{aligned} \tag{14}$$

also used in Wang [2012]. The prior promotes sparsity by positing that entries of the precision matrix follow a Laplace distribution, but together they must also lie in the positive-definite cone, $\Sigma^{-1}\in\mathbb{S}^{p\times p}_+$, or equivalently $\lambda_{\min}(\Sigma^{-1}) > 0$. Hence, CWBB involves simply repeatedly sampling w_n and solving

$$\begin{aligned} \operatorname*{arg\,min}_{\Sigma^{-1}\in\mathbb{S}^{p\times p}_+} & -\frac{n}{2}\log\det\Sigma^{-1}\\ &+\frac{n}{2}\operatorname{tr}\{\Sigma^{-1}X^\top\operatorname{diag}(w_n)X\} + \sum_{i\le j} q_\rho(\Sigma^{-1}_{ij}), \end{aligned} \tag{15}$$

where $q_\rho(\Sigma^{-1}_{ij}) = \rho|\Sigma^{-1}_{ij}|$. This can be optimized efficiently by the methods in Friedman et al. [2008]. After many samples from CWBB, we calculate the posterior credible intervals for each entry.

The generality of CWBB allows uncertainty quantification for other penalties that induce sparsity. For example, the smoothly clipped absolute deviation [Fan and Li, 2001] and minimax concave penalty [Zhang, 2010] have desirable properties in problems requiring sparsity,

$$q_{\rho,SCAD}(z) = \begin{cases} \rho|z| & |z|\le\rho,\\ \frac{2a\rho|z|-z^2-\rho^2}{2(a-1)} & \rho\le|z|\le a\rho,\\ \frac{\rho^2(a+1)}{2} & |z|\ge a\rho, \end{cases} \tag{16}$$

$$q_{\rho,MCP}(z) = \begin{cases} \rho|z| - \frac{z^2}{2a} & |z|\le a\rho,\\ \frac{1}{2}a\rho^2 & |z| > a\rho. \end{cases} \tag{17}$$

By replacing the penalty q in (15) with (16), we can now utilize SCAD toward Bayesian sparse precision estimation efficiently; the point estimate was shown to be effective in Fan et al. [2009]. We also compare with the minimax concave penalty (17) from Zhang [2010], which exhibits similar performance and illustrates the flexibility of CWBB. For the analogous problem of sparse covariance estimation, CWBB also fits with principled optimization routines [Bien and Tibshirani, 2011, Xu and Lange, 2022].

The posterior credible intervals can guard against the well-documented issue of false selection with ℓ_1 penalties [Su et al., 2017, Lafit et al., 2019]. Rather than use a single graphical lasso estimate to determine non-zero entries, we employ a decision rule that classifies an entry as nonzero if the majority of the credible interval is either strictly positive or negative. To motivate casting the problem as posterior inference, if a credible interval is wide and almost symmetric about zero, it is reasonable to classify the entry as zero compared to one with a similar point estimate but with a credible interval with mostly positive support.

We perform a simulation study with $p = 100, n = 750$ and $\Sigma^{-1} = L^\top L$, where L is triangular and off-diagonal entries are independent mixtures of a standard normal and a point mass at zero. The sparsity of the true Σ^{-1} is about two-thirds. CWBB performs 250 samples from (15) for each penalty function. We use $a = 3.7$ for (16) and $a = 3$ for (17) as both were shown to be reasonable defaults in their respective works. To evaluate performance of each method, entries of the precision matrix are treated as a binary classification problem with nonzero entries considered as positives. The

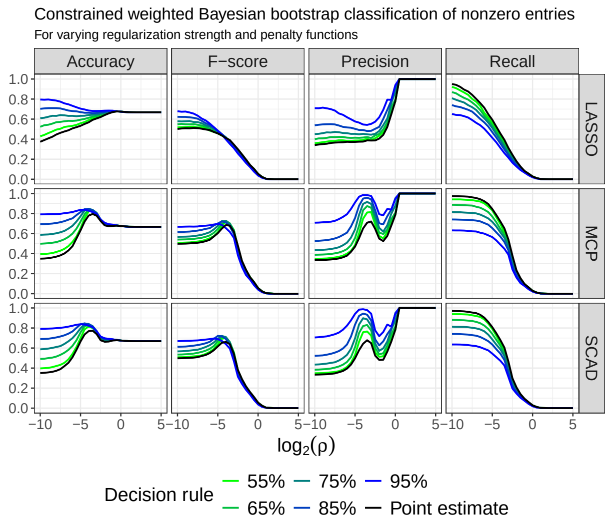


Figure 3: Graph showing binary classification performance of nonzero entries of a sparse precision matrix with CWBB. The black line represents classifying each entry according to the point estimate at ρ.

classification metrics considered for evaluation are accuracy, precision, recall and F_1-score.

For many values of regularization strength, each row of Fig. 3 demonstrates an increase in performance when classifying precision matrix entries based on the credible intervals derived from CWBB as opposed to a single point estimate for all penalties considered. This performance is shown for a variety of decision rules dependent on the proportion of posterior samples that are strictly either positive or negative. This decision rule evidently results in an increase in false negatives as seen in the decrease in recall. However, the gain in precision outpaces these faults by both F_1-score and classification accuracy. Furthermore, there is a wide range of regularization hyperparameters that have competitive performance with the posterior samples, showing less sensitivity to tuning ρ for all possible penalty functions. Performance of CWBB is competitive with the more sophisticated Bayesian methods mentioned earlier, while involving fewer hyperparameters (see Table 2 of the Supplementary Material). The extra machinery of these methods requires some delicate handling of the burn-in period, starting values, and tuning, while CWBB is a simple, easy-to-use approach.

4.3 OPTION PRICING SURFACES

For a fixed stock, call options may be bought at a variety of strike prices and expiry dates. European-style call options give the buyer the right to purchase 100 shares of a stock at time t for price s. Let $C(t, s)$ be the price of this option. The pricing function C follows restrictions which prevent various types of arbitrage. Here, we focus on a set of minimal basic restrictions that are readily justifiable:

$$\begin{aligned} &C(t,\cdot) \text{ is convex and decreasing for fixed } t; \\ &C(\cdot,s) \text{ is increasing for fixed } s. \end{aligned} \tag{18}$$

This posits that the underlying price of a call option must decrease as the strike price increases, otherwise a clear arbitrage opportunity is available. Convexity is derived from principles in option pricing theory outside the scope of this work [Aït-Sahalia and Duarte, 2003]. Because call options risk expiry before an advantageous price is reached for the buyer, their value or price increases with lifespan, reflected in the second constraint.

We consider an additional constraint available from market data, as option contracts come with a bid and an ask price. The market price is generally taken to be the midpoint of the bid-ask spread, but this may not be an ideal estimate of the underlying price, and may violate the natural restrictions described above. Assuming the bid-ask is rational and the true option price falls within this spread, we add the constraint

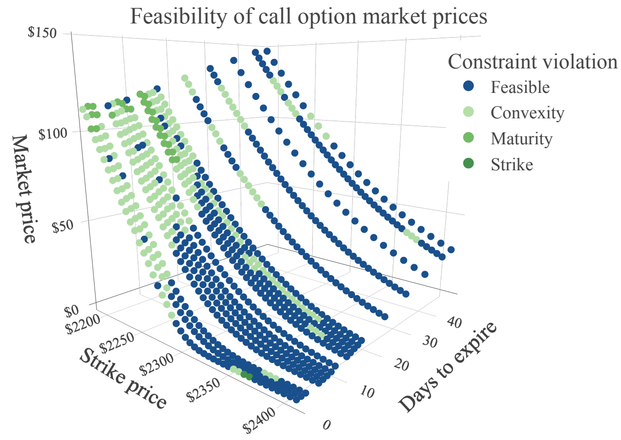
$$BID(t,s) \le C(t,s) \le ASK(t,s). \tag{19}$$

When pricing options that are part of a spread, both covariance and constraints play important roles. As an example, consider buying a vertical call spread, i.e. a trader buys a call expiring at time t for strike s_1, and sells a call expiring at the same time for strike s_2 with $s_2 > s_1$. The net price is $P = C(t, s_1) - C(t, s_2)$. Therefore, a probabilistic model for P must have support only on nonnegative reals to satisfy (18). In addition, uncertainty estimates are essential for portfolio optimization, and the uncertainty for P is changed by the covariance between $C(t, s_1)$ and $C(t, s_2)$.

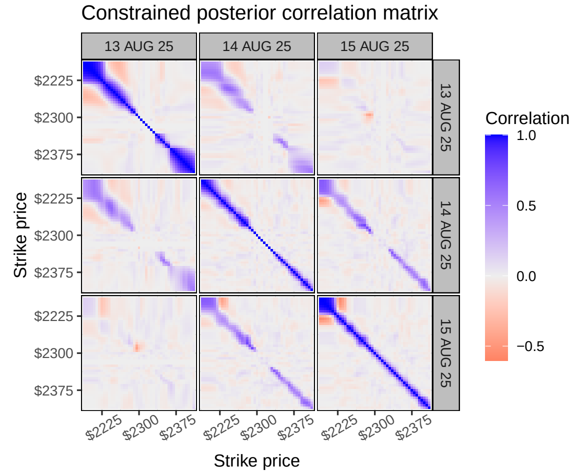
Similar to prior work [Aït-Sahalia and Duarte, 2003], we have a simple model specification, complemented by the complexity of the constraints. In particular, we choose a constrained Bayesian normal means formulation with a likelihood and unconstrained prior of

$$y \mid C \sim N_n(C, \lambda^{-1} I), \qquad C \sim N_n(0, \delta^{-1}\Sigma), \tag{20}$$

where n is the number of options available for trade, y is the observed midpoint $(BID + ASK)/2$, λ and δ are precision hyperparameters and Σ is a positive-definite matrix such that each entry is the prior covariance between a pair $C(t_1, s_1), C(t_2, s_2)$. In the notation of (20), C is flattened as a vector in $\mathbb{R}^n$, with indexing by t and s suppressed in the notation. We model the midpoint price as a noisy observation of the underlying price. The posterior distribution of (20) is multivariate normal with covariance $\Sigma' = \{\lambda I + \delta\Sigma^{-1}\}^{-1}$ and mean $\lambda\Sigma' y$. As mentioned before, covariance is essential, yet this posterior derives covariance of option prices solely from the choice of hyperparameters λ, δ and Σ, ignoring the context of the data. By factoring in the constraints (18) and (19) we induce more viable estimates of the covariance of the option prices under pertinent assumptions.



(4a) For RUT call options, 3D scatter plots showing the midpoint price as it varies in strike price and days to expiration with points colored by constraint violations.



(4b) Correlation matrix for $C \mid y$ from CWBB for the first three times to expire in the dataset.

Iterations of CWBB solve

$$\underset{DC \geq b}{\arg\min} \frac{\lambda}{2} \sum_{i=1}^{n} w_{n,i}(C_i - y_i)^2 + \frac{\delta}{2} C^\top \Sigma^{-1} C, \qquad (21)$$

where D is a matrix such that $DC \geq b$ implies feasibility. In particular, D contains signed copies of a submatrix D_1 where $D_1 x \geq 0$ implies x is nondecreasing and D_2 where $D_2 \geq 0$ implies a convex structure (see Kim et al. [2009] for details). In addition, D contains a positive and negative copy of the identity matrix to satisfy the bid-ask box constraints.

We perform experiments using market data from about 10 AM on August 13th, 2025 for the RUT European-style call options. Tradable options that expire within 60 days and have a strike price within 5% of the underlying price of \$2300 are included in the dataset ($n = 766$). Each data point consists of a bid and an ask, and the midpoint of the bid-ask is used as y_i in (21). The strike price and time to expire for each is used to calculate the prior covariance of C while defining the constraints that resulting posterior samples must satisfy. In our exploratory data analysis and resulting inference, there are patterns corresponding to a typical classification in options: (Deep) In-the-money call options have a strike price (well) below the underlying price; near-the-money call options have a strike close to the underlying; (deep) out-of-the-money have a strike (well) above the underlying.

Figure 4a shows which options have an observed midpoint price which violate some type of constraint, i.e. for option j there exists some row i in D where $\text{row}_i(D)^\top(BID + ASK) < 0$, and $D_{ij} \neq 0$. As an example, options expiring in 15 days with strike prices, \$2285, \$2290, and \$2295 have a midpoint price of \$48.00, \$45.00, and \$41.90, respectively. This violates convexity as the first derivative is not monotone increasing with respect to strike price, hence these three points are light green in Fig. 4a. Most of the near-the-money options are not part of a constraint violation. On average, as the time to expire increases, with 15 days to expiry as an exception, the range of feasible options grows. Deep in-the-money call options are likely to violate convexity, the most common violation. Sequences of call options for several expiration dates violate maturity monotonicity.

We choose hyperparameters $\lambda = 1/3$ and $\delta = 1/35$. The hyperparameter Σ is determined by an exponential kernel over the strike price–expiration date space, e.g. if $z_i = (s_i, t_i)$ is the strike and expiration of the ith option, then $\Sigma_{ij} = \exp\{-\text{dist}(z_i/\sigma_s, z_j/\sigma_t)\}$, where σ normalizes the coordinates. This makes the unconstrained prior of (20) informative, but the small value of δ compared to λ leads the posterior covariance to have many tridiagonally dominated blocks, with over 93% of entries having an absolute value less than 10^{-3}. Using the OSQP solver of Stellato et al. [2020], 1000 samples of $C \in \mathbb{R}^{766}$ under the constrained posterior are found via CWBB. With these posterior samples under the constrained model, we can estimate $\text{cov}\{C(t_1, s_1), C(t_2, s_2) \mid y\}$ for all expirations and strikes.

Figure 4b shows the empirical correlation matrix of the posterior parameters (option prices), for the first three times to expire in the snapshot, with blocks highlighting the pairs of expiration dates. Each tile corresponds to the sample correlation between the constrained posterior samples of $C(t_1, s_1)$ and $C(t_2, s_2)$. We see clear structure with blocks of high correlation for deep in-the-money call options as well as increasing amounts of correlation for deep out-of-the-money call options which expire on the same day. These blocks grow smaller as the time to expire increases, seen by the smaller correlation blocks for options expiring on the 15th of August, as opposed to the 13th. Furthermore, there is a small amount of correlation between options on adjacent days to expiry, but this correlation is weaker when examining options that expire two days apart. Near-the-money options also exhibit some weak correlation, although it does not have a block structure.

The constraints lead to uncertainty and point estimates related to the common classification of “moneyness” of the option. In the Supplementary Material, we display additional results on the constrained posterior, such as how it compares to the unconstrained problem, and how uncertainty relates to the spread of an option. In addition, we detail the limitations of several other constrained sampling methods for this case study. We find they are less viable, involving intractable runtimes, small effective sample sizes, or excessive constraint violation, highlighting both the difficulty of this problem and the generality of CWBB.

5 DISCUSSION

This article demonstrates that extending the weighted Bayesian bootstrap to constrained problems yields a simple yet effective method applicable to a diverse range of statistical tasks. Our approach bridges the advantages of uncertainty quantification via approximate posterior sampling to the computational expediency of optimization routines for constrained problems. We identify standard regularity assumptions that lead to new theoretical guarantees and asymptotic results on the samples from Algorithm 1. Just as the bootstrap enjoys widespread usage due to its simplicity—even beyond settings where all regularity conditions are checked—our method is easy to use with any plug-in optimization solution to constrained problems.

Compared to other constrained posterior sampling methods, CWBB enjoys competitive coverage, runtimes, and sampling efficiency. Our theory shows that relaxation-based methods may not entirely factor in the structure of constraint sets in their uncertainty estimates, while CWBB is centered around an efficient restricted maximum likelihood estimator. As demonstrated by the runtimes in the Supplementary Material, CWBB maintains a smaller computational footprint than several competing methods which require prohibitively expensive evaluation of proximal operators for the log-density and gradient calculations. Although our method is meant to cover cases where dedicated samplers are not available, we note CWBB performs competitively with the bespoke samplers `BDGraph` and `ssgraph` as demonstrated in Section 4.2. Furthermore, the options case study of Section 4.3 is particularly difficult, with CWBB as the only method able to produce viable samples quickly from the constrained posterior, without difficulties from mixing or infeasible samples.

There are several immediate future directions that these contributions imply. First, it is natural to strengthen the theory to encompass a broader class of settings. Indeed, while Theorem 2 resembles the standard Bernstein–von Mises result, it requires more stringent conditions than some comparable results in the literature (see Supplementary Material). As many of these technical assumptions are imposed in order to produce valid Taylor expansions around the true parameter in both the constraint and likelihood functions, weaker assumptions may apply to broaden the scope of these results with modern posterior analysis techniques such as those from Miller [2021]. Some specific tasks such as high-dimensional regression and normal means models admit attractive theoretical properties under the weighted Bayesian bootstrap with weaker regularity conditions. These include consistency [Ng and Newton, 2022] and minimax posterior contraction rates [Nie and Ročková, 2023] under penalty formulations instead of hard constraints, again suggesting that our theory could be made more applicable to specific cases by considering the geometry of the problem.

Our analysis justifies the constrained weighted Bayesian bootstrap in a first-order sense. The support of the prior is considered in Theorem 2, but we do not fully make use of its functional form. Techniques such as Edgeworth expansions may produce higher-order approximations of the weighted likelihood/Bayesian bootstrap while accounting for prior structure [Newton and Raftery, 1994, Pompe, 2021]. Finally, because of the breadth statistical tasks that can be formulated as constrained optimization problems, we invite readers to explore further applications of CWBB to posterior inference where constrained spaces pose a challenge. This includes sampling over valid correlation matrices, intersections of convex sets or split problems [Xu et al., 2018], and combinatorial problems [Xu and Duan, 2023].

References

Yacine Aït-Sahalia and Jefferson Duarte. Nonparametric option pricing under shape restrictions. *Journal of Econometrics*, 116(1):9–47, September 2003. ISSN 0304-4076. doi: 10.1016/S0304-4076(03)00102-7.

J. Aitchison and S. D. Silvey. Maximum-Likelihood Estimation of Parameters Subject to Restraints. *The Annals of Mathematical Statistics*, 29(3):813–828, 1958. ISSN 0003-4851. doi: 10.1214/aoms/1177706538.

Lachlan Astfalck, Deborshee Sen, Sayan Patra, Edward Cripps, and David Dunson. Posterior Projection for Inference in Constrained Spaces. *arXiv:* 1812.05741v5, December 2024. doi: 10.48550/arXiv.1812.05741.

Michael Betancourt. A Conceptual Introduction to Hamiltonian Monte Carlo. *arXiv:* 1701.02434, July 2018. doi: 10.48550/arXiv.1701.02434.

Jacob Bien and Robert J. Tibshirani. Sparse estimation of a covariance matrix. *Biometrika*, 98(4):807–820, December 2011. ISSN 0006-3444. doi: 10.1093/biomet/asr054.

Leo L. Duan, Alexander L. Young, Akihiko Nishimura, and David B. Dunson. Bayesian constraint relaxation.

Biometrika, 107(1):191–204, March 2020. ISSN 0006-3444. doi: 10.1093/biomet/asz069.

David B. Dunson and Brian Neelon. Bayesian Inference on Order-Constrained Parameters in Generalized Linear Models. *Biometrics*, 59(2):286–295, 2003. ISSN 0006-341X. doi: 10.1111/1541-0420.00035.

Jasper M. Everink, Yiqiu Dong, and Martin S. Andersen. Bayesian Inference with Projected Densities. *SIAM/ASA Journal on Uncertainty Quantification*, 11(3):1025–1043, September 2023. ISSN 2166-2525. doi: 10.1137/22M150695X.

Jianqing Fan and Runze Li. Variable Selection via Nonconcave Penalized Likelihood and its Oracle Properties. *Journal of the American Statistical Association*, 96(456):1348–1360, December 2001. ISSN 0162-1459. doi: 10.1198/016214501753382273.

Jianqing Fan, Yang Feng, and Yichao Wu. Network Exploration via the Adaptive LASSO and SCAD Penalties. *The Annals of Applied Statistics*, 3(2):521–541, 2009. ISSN 1932-6157. doi: 10.1214/08-AOAS215SUPP.

Edwin Fong, Simon Lyddon, and Chris Holmes. Scalable Nonparametric Sampling from Multimodal Posteriors with the Posterior Bootstrap. In *Proceedings of the 36th International Conference on Machine Learning*, pages 1952–1962. PMLR, May 2019.

Jerome Friedman, Trevor Hastie, and Robert Tibshirani. Sparse inverse covariance estimation with the graphical lasso. *Biostatistics*, 9(3):432–441, July 2008. ISSN 1465-4644. doi: 10.1093/biostatistics/kxm045.

Alan E. Gelfand, Adrian F. M. Smith, and Tai-Ming Lee. Bayesian Analysis of Constrained Parameter and Truncated Data Problems Using Gibbs Sampling. *Journal of the American Statistical Association*, 87(418):523–532, 1992. ISSN 0162-1459. doi: 10.2307/2290286.

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. ISSN 1369-7412. doi: 10.1111/j.1467-9868.2010.00765.x.

Wei Han and Yun Yang. Statistical Inference in Mean-Field Variational Bayes. *arXiv:* 1911.01525, November 2019. doi: 10.48550/arXiv.1911.01525.

J. T. Gene Hwang and Shyamal Das Peddada. Confidence Interval Estimation Subject to Order Restrictions. *The Annals of Statistics*, 22(1):67–93, 1994. ISSN 0090-5364. doi: 10.1214/aos/1176325358.

Seung-Jean Kim, Kwangmoo Koh, Stephen Boyd, and Dimitry Gorinevsky. ℓ_1 Trend Filtering. *SIAM Review*, 51(2):339–360, May 2009. ISSN 0036-1445. doi: 10.1137/070690274.

Ginette Lafit, Francis Tuerlinckx, Inez Myin-Germeys, and Eva Ceulemans. A Partial Correlation Screening Approach for Controlling the False Positive Rate in Sparse Gaussian Graphical Models. *Scientific Reports*, 9(1):17759, November 2019. ISSN 2045-2322. doi: 10.1038/s41598-019-53795-x.

Jason D. Lee, Dennis L. Sun, Yuekai Sun, and Jonathan E. Taylor. Exact post-selection inference, with application to the lasso. *The Annals of Statistics*, 44(3):907–927, June 2016. ISSN 0090-5364, 2168-8966. doi: 10.1214/15-AOS1371.

S. P. Lyddon, C. C. Holmes, and S. G. Walker. General Bayesian updating and the loss-likelihood bootstrap. *Biometrika*, 106(2):465–478, June 2019. ISSN 0006-3444. doi: 10.1093/biomet/asz006.

Anna Menacher, Thomas E. Nichols, Chris Holmes, and Habib Ganjgahi. Bayesian Lesion Estimation with a Structured Spike-and-Slab Prior. *Journal of the American Statistical Association*, 119(545):66–80, January 2024. ISSN 0162-1459. doi: 10.1080/01621459.2023.2278201.

Jeffrey W. Miller. Asymptotic normality, concentration, and coverage of generalized posteriors. *Journal of Machine Learning Research*, 22(1), January 2021. ISSN 1532-4435.

A. Mohammadi and E. C. Wit. Bayesian Structure Learning in Sparse Gaussian Graphical Models. *Bayesian Analysis*, 10(1):109–138, March 2015. ISSN 1936-0975, 1931-6690. doi: 10.1214/14-BA889.

Reza Mohammadi. *ssgraph: Bayesian Graphical Estimation using Spike-and-Slab Priors*, December 2022. Version 1.15. doi: 10.32614/CRAN.package.ssgraph.

Reza Mohammadi and Ernst C. Wit. BDgraph: An R Package for Bayesian Structure Learning in Graphical Models. *Journal of Statistical Software*, 89:1–30, May 2019. ISSN 1548-7660. doi: 10.18637/jss.v089.i03.

Michael A. Newton. *The Weighted Likelihood Bootstrap and an Algorithm for Prepivoting*. PhD thesis, University of Washington, 1991. URL `http://hdl.handle.net/1773/8962`.

Michael A. Newton and Adrian E. Raftery. Approximate Bayesian Inference with the Weighted Likelihood Bootstrap. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 56(1):3–48, 1994. ISSN 0035-9246. doi: 10.1111/j.2517-6161.1994.tb01956.x.

Michael A. Newton, Nicholas G. Polson, and Jianeng Xu. Weighted Bayesian bootstrap for scalable posterior distributions. *Canadian Journal of Statistics*, 49(2):421–437, 2021. ISSN 1708-945X. doi: 10.1002/cjs.11570.

Tun Lee Ng and Michael A. Newton. Random weighting in LASSO regression. *Electronic Journal of Statistics*, 16(1): 3430–3481, January 2022. ISSN 1935-7524, 1935-7524. doi: 10.1214/22-EJS2020.

Lizhen Nie and Veronika Ročková. Bayesian Bootstrap Spike-and-Slab LASSO. *Journal of the American Statistical Association*, 118(543):2013–2028, July 2023. ISSN 0162-1459. doi: 10.1080/01621459.2022.2025815.

Jorge Nocedal and Stephen J. Wright. Theory of Constrained Optimization. In *Numerical Optimization*, pages 304–354. Springer, New York, NY, 2006. ISBN 978-0-387-40065-5. doi: 10.1007/978-0-387-40065-5_12.

John Ormerod, Mohammad Javad Davoudabadi, Garth Tarr, Samuel Mueller, and Jonathon Tidswell. *Bayesian Lasso Regression and Tools for the Lasso Distribution*, 2025. Version 0.3.0. doi: 10.32614/CRAN.package.BayesianLasso.

Snigdha Panigrahi and Jonathan Taylor. Scalable methods for Bayesian selective inference. *Electronic Journal of Statistics*, 12(2):2355–2400, January 2018. ISSN 1935-7524, 1935-7524. doi: 10.1214/18-EJS1452.

K. B. Petersen and M. S. Pedersen. The matrix cookbook, Nov 2012. URL `http://www2.compute.dtu.dk/pubdb/pubs/3274-full.html`. Version 20121115.

Emilia Pompe. Introducing prior information in Weighted Likelihood Bootstrap with applications to model misspecification. *arXiv:* 2103.14445v2, April 2021. doi: 10.48550/arXiv.2103.14445.

Rick Presman and Jason Xu. Distance-to-Set Priors and Constrained Bayesian Inference. In *Proceedings of The 26th International Conference on Artificial Intelligence and Statistics*, volume 206, pages 2310–2326. PMLR, April 2023.

Veronika Ročková and Edward I. George. The Spike-and-Slab LASSO. *Journal of the American Statistical Association*, 113(521):431–444, January 2018. ISSN 0162-1459. doi: 10.1080/01621459.2016.1260469.

Stan Development Team. Stan modeling language users guide and reference manual, version 2.32, 2023. URL `https://mc-stan.org/docs/2_32/reference-manual/`.

Bartolomeo Stellato, Goran Banjac, Paul Goulart, Alberto Bemporad, and Stephen Boyd. OSQP: An operator splitting solver for quadratic programs. *Mathematical Programming Computation*, 12(4):637–672, December 2020. ISSN 1867-2957. doi: 10.1007/s12532-020-00179-2.

P. Stoica and Boon Chong Ng. On the Cramer-Rao bound under parametric constraints. *IEEE Signal Processing Letters*, 5(7):177–179, July 1998. ISSN 1558-2361. doi: 10.1109/97.700921.

Weijie Su, Małgorzata Bogdan, and Emmanuel Candès. False discoveries occur early on the Lasso path. *The Annals of Statistics*, 45(5):2133–2150, October 2017. ISSN 0090-5364, 2168-8966. doi: 10.1214/16-AOS1521.

A. W. van der Vaart. *Asymptotic Statistics*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press, Cambridge, 1998. ISBN 978-0-521-78450-4. doi: 10.1017/CBO9780511802256.

Hao Wang. Bayesian Graphical Lasso Models and Efficient Posterior Computation. *Bayesian Analysis*, 7(4):867–886, December 2012. ISSN 1936-0975, 1931-6690. doi: 10.1214/12-BA729.

Hao Wang. Scaling It Up: Stochastic Search Structure Learning in Graphical Models. *Bayesian Analysis*, 10(2): 351–377, June 2015. ISSN 1936-0975, 1931-6690. doi: 10.1214/14-BA916.

Jason Xu and Kenneth Lange. A proximal distance algorithm for likelihood-based sparse covariance estimation. *Biometrika*, 109(4):1047–1066, December 2022. ISSN 1464-3510. doi: 10.1093/biomet/asac011.

Jason Xu, Eric C. Chi, Meng Yang, and Kenneth Lange. A majorization–minimization algorithm for split feasibility problems. *Computational Optimization and Applications*, 71(3):795–828, December 2018. ISSN 1573-2894. doi: 10.1007/s10589-018-0025-z.

Maoran Xu and Leo L. Duan. Bayesian inference with the l 1-ball prior: solving combinatorial problems with exact zeros. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 85(5):1538–1560, 2023. ISSN 1369-7412. doi: 10.1093/jrsssb/qkad076.

Maoran Xu, Hua Zhou, Yujie Hu, and Leo L. Duan. Bayesian Inference Using the Proximal Mapping: Uncertainty Quantification Under Varying Dimensionality. *Journal of the American Statistical Association*, 119 (547):1847–1858, July 2024. ISSN 0162-1459. doi: 10.1080/01621459.2023.2220170.

Cun-Hui Zhang. Nearly unbiased variable selection under minimax concave penalty. *The Annals of Statistics*, 38(2): 894–942, April 2010. ISSN 0090-5364, 2168-8966. doi: 10.1214/09-AOS729.

Xinkai Zhou, Qiang Heng, Eric C. Chi, and Hua Zhou. Proximal MCMC for Bayesian Inference of Constrained and Regularized Estimation. *The American Statistician*, 78(4):379–390, October 2024. ISSN 0003-1305. doi: 10.1080/00031305.2024.2308821.

Constrained Weighted Bayesian Bootstrap (Supplementary Material)

Sam Rosen[1] **Jason Xu**[2]

[1]Department of Statistical Science, Duke University
[2]Department of Biostatistics, University of California, Los Angeles

A ADDITIONAL EXPERIMENT RESULTS

A.1 NONNEGATIVE ORDER-CONSTRAINED PARAMETERS IN REGRESSION

This experiment is a comprehensive comparison of the constrained posterior sampling methods mentioned in Section 1. In Table 1, we show that CWBB has runtime orders of magnitude faster than several other constrained sampling methods. Most methods performed 250 posterior samples per trial. However, methods requiring HMC were afforded a warmup and 1000 samples (except Proximal Prior with 250 samples) to account for potential correlation across samples. In addition, Fig. 5 shows the minimum coverage (the lowest point for each dataset size in Fig. 2b) over all $p = 30$ parameters in θ. CWBB is most similar to the baseline unconstrained Gibbs sampler, as compared to the competing methods. Finally, we use the experiments from Section 4.1 to empirically demonstrate Theorems 1 and 2. Figure 6 shows convergence of CWBB samples to the true underlying parameter and the covariance matrix of the samples to the projected inverse Fisher information.

Table 1: Time to complete 2500 trials for coverage simulation study.

Method	Samples per Trial	Total Runtime (seconds)
Constraint Relaxation [Duan et al., 2020]	1000	37958
Constrained Weighted Bayesian Bootstrap	250	1242
Transformation of Dunson and Neelon [2003]	250	2538
Orthogonally Projected Gibbs Samples	250	255
Proximal Prior [Xu et al., 2024]	250	83830
ProxMCMC/Distance-to-set [Presman and Xu, 2023, Zhou et al., 2024]	1000	3200
`positive_ordered` Change of Variables in Stan	1000	3304
Unconstrained Gibbs Sampler	250	110

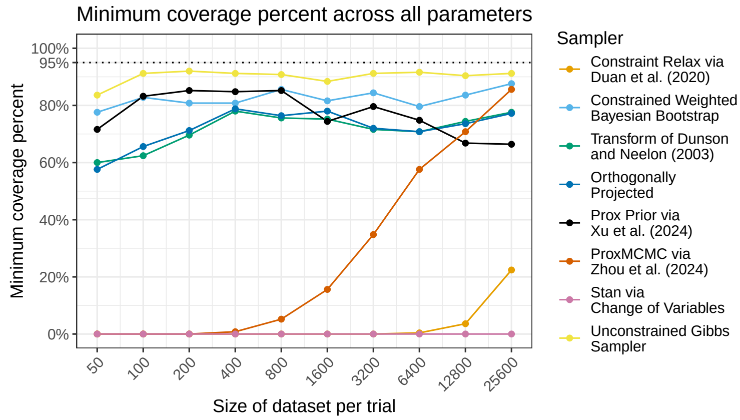


Figure 5: For each sampling method, the minimum coverage percent over all θ_i for each dataset size.

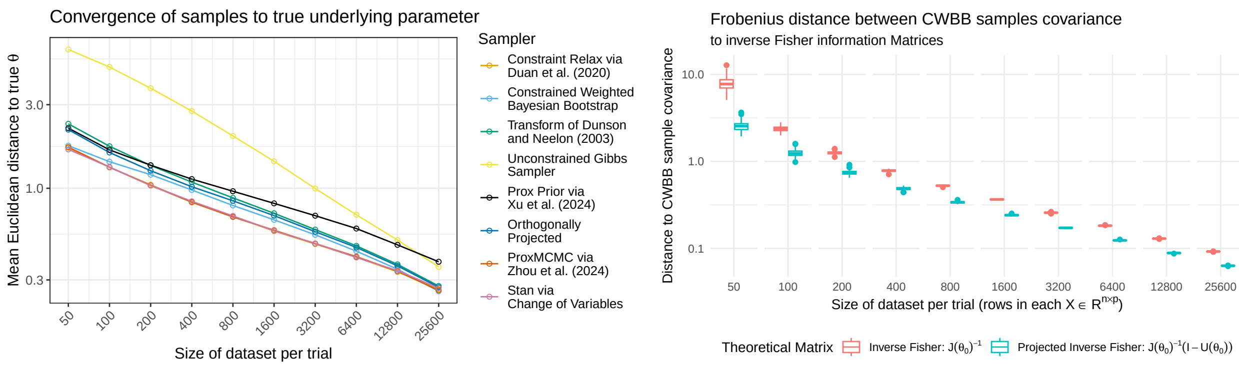


Figure 6: (Left) Empirical verification of Theorem 1, showing CWBB samples converge to the true θ, at a rate similar to that of competing methods. (Right) Empirical verification of Theorem 2, showing the empirical covariance matrix of CWBB samples is closer in Frobenius norm to the theoretical covariance described in Theorem 2, as opposed to the theoretical covariance from unconstrained settings.

A.2 SPARSE PRECISION MATRIX ESTIMATION

The first three rows of Table 2 show the peak performance from Fig. 3 while the remaining rows show peak performance over a hyperparameter grid for other Bayesian methods. Bayesian methods `BDgraph` and `ssgraph` perform 5000 iterations of their respective samplers with the first half discarded as burn-in. ProxMCMC performs 1000 iterations of a No-U-turn sampler [Betancourt, 2018], along with a dynamic warmup, as implemented in the Julia package `DynamicHMC.jl`. These methods are afforded more iterations due to potential correlation among successive samples. ProxMCMC is flexible enough to implement the nonconvex penalties SCAD and MCP (see code for implementation details), although performance was not improved with them. The methods of Mohammadi and Wit [2015] and Wang [2015] have a starting graph where all values of the precision matrix are nonzero; performance was poor otherwise. All F_1-scores shown are for the best performing threshold for posterior probability of a precision matrix entry being nonzero. For the first seven rows, this corresponds to the maximum posterior probability of the entry being strictly positive or negative, while the remaining two rows are posterior probabilities of the edge indicator being nonzero. `R` packages `BayesianGLasso` [Ormerod et al., 2025], `BDGraph` [Mohammadi and Wit, 2019], and `ssgraph` [Mohammadi, 2022] implement the methods of Wang [2012], Mohammadi and Wit [2015] and Wang [2015], respectively. The code to implement ProxMCMC was provided with Zhou et al. [2024].

A.3 OPTION PRICING SURFACES

Examining the left side of Fig. 7, the violations explored in Fig. 4a are correlated with a larger spread. Although one might expect options with a wider range of possible prices to have more uncertainty in their underlying price, by factoring in the constraints we are able to incur smaller uncertainty in options with large spread. This is shown in the right side of Fig. 7

Table 2: Best performances for Bayesian sparse precision estimates. Runtime is the time to complete samples for the set of given hyperparameters.

Method	F_1-score	Decision Rule	Hyperparameters	Runtime (seconds)
CWBB - LASSO	0.68	95%	$\rho = 2^{-10}$	23
CWBB - SCAD	0.72	75%	$\rho = 2^{-4.5}, a = 3.7$	250
CWBB - MCP	0.73	65%	$\rho = 2^{-4}, a = 3$	215
ProxMCMC - LASSO	0.67	95%	$\lambda = 0.1, \alpha_{\text{scale}} = 400$	68
ProxMCMC - SCAD	0.66	95%	$\lambda = 10^{-4}, \alpha_{\text{scale}} = 50$	294
ProxMCMC - MCP	0.67	95%	$\lambda = 10^{-3}, \alpha_{\text{scale}} = 200$	145
`BayesianGLasso`	0.61	85%	$a = 10^{-4}, b = 2^{14}$	343
`BDGraph`	0.76	75%	`g.prior` $= 0.35$, `df.prior` $= 3 \times 2^8$, `g.start` = "full"	78
`ssgraph`	0.78	55%	$\lambda = 2^{-4}, \pi = 0.75, v_0 = 4\times10^{-4}, v_1 = 1$, `g.start` = "full"	217

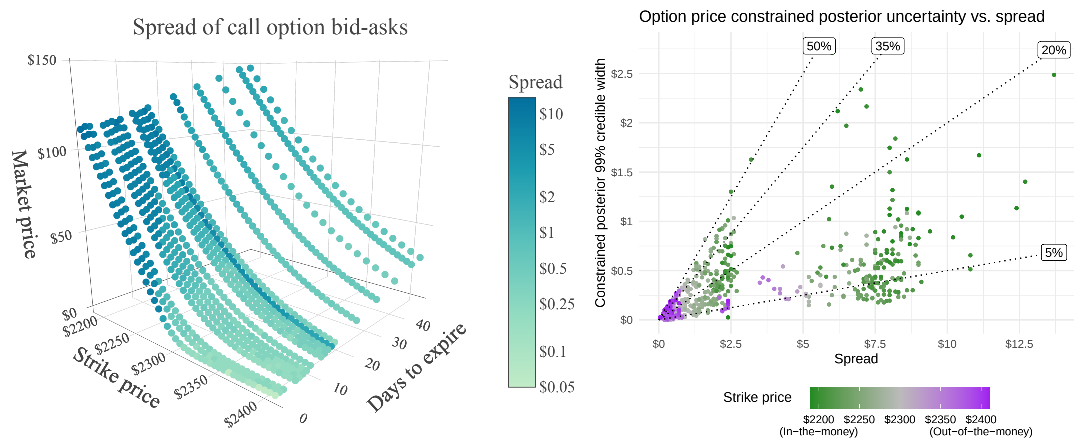


Figure 7: (Left) For the dataset of `RUT` call options, scatter plots in 3D showing the midpoint price (z-axis) as it varies in both strike price (x-axis) and days to expiration (y-axis). Points are colored according to the log of the spread. (Right) Scatter plot showing the spread of options on the x-axis, the posterior credible interval widths on the y-axis, with points colored by underlying strike price. Several references lines are included to show what percent of the spread the credible interval contains.

showing many options deep in-the-money (a strike price well below the underlying price of $2300) have a credible interval width that is a small proportion of the spread, as opposed to those at-the-money which have a width that contains up to half of the spread.

The left side of Fig. 8 shows the unconstrained posterior correlation matrix, which contains less rich structure than the constrained one showed in Fig. 4b. There are some similarities, such as weak correlation when examining options that expire two days apart. The right side of Fig. 8 shows the difference in point estimates between the constrained and unconstrained posteriors. Deep in-the-money options tend to have a larger difference, although most options have very similar price estimates. However, for a subset of days to expire, there is a noticeable difference in price for some options at-the-money.

Comparison with other samplers: Due to the geometry of the constraint set, modern constrained posterior methods struggle heavily on the options case study, while CWBB is able to produce 1000 independent samples in less than 10 minutes (see Table 3). Both constraint relaxation [Duan et al., 2020] and ProxMCMC [Zhou et al., 2024] (equivalent to distance-to-set priors [Presman and Xu, 2023]) take a significant amount of time longer than CWBB to produce 1000 samples. The samples from ProxMCMC/distance-to-set struggle to mix, giving a mean effective sample size below 5, despite a proper warmup with a No-U-turn-sampler. Both constraint relaxation and ProxMCMC/distance-to-set are unable to produce samples in the constraint set with many samples violating the bid-ask or shape constraints. Attempts to increase the regularization only lead to prohibitive sampling times and difficulty mixing. The proximal prior of Xu et al. [2024] has extreme computational costs

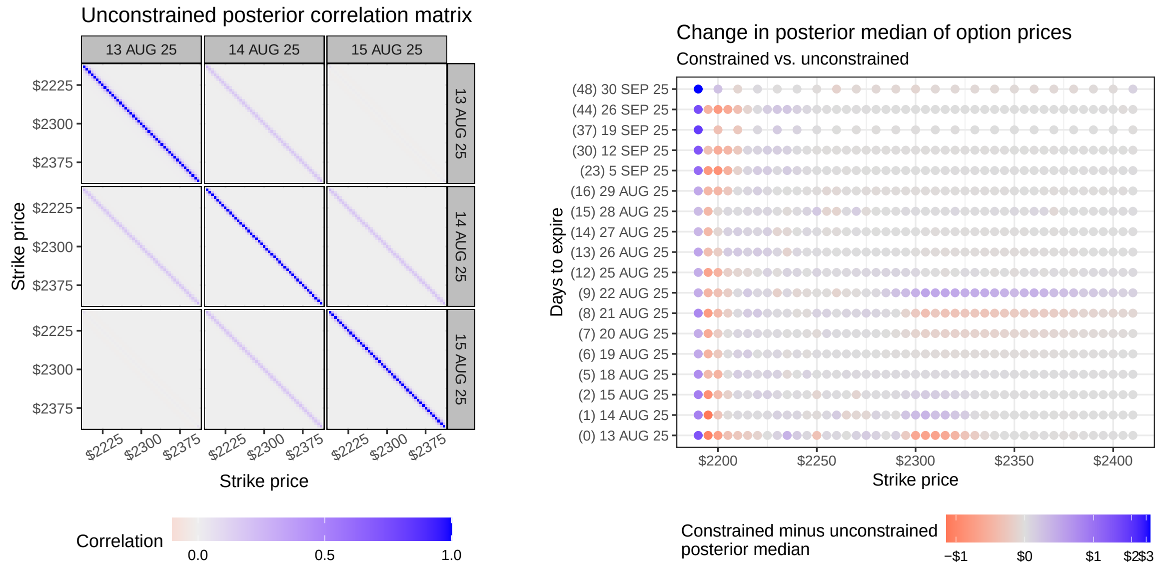


Figure 8: (Left) Correlation matrix for unconstrained posterior samples of normal means for the first three times to expire in the dataset. Options are ordered in the matrix by time to expire and then by strike price. (Right) Scatter plot with strike price on the x-axis, days to expire on the y-axis, and points colored by how much larger the constrained posterior median is than the unconstrained.

as it requires a numerical approximation of the gradient of the projection onto the constraint polytope. Despite optimized code, it is unable to produce samples in a reasonable amount of time.

Table 3: Properties of 1000 samples for the case study in Section 4.3.

Method	Feasible Samples	Mean Effective Sample Size	Total Runtime (seconds)
Constraint Relaxation	0	603.10	60865
CWBB	1000	Gives independent samples	494
Proximal Prior	-	-	Failed
ProxMCMC/Distance-to-set	0	4.78	6775

B PROOF OF MAIN RESULTS

B.1 PROOF PRELIMINARIES

The techniques used in the resulting proofs are found in Michael Newton's 1991 PhD dissertation at the University of Washington [Newton, 1991], and work on restricted maximum likelihood estimation by Aitchison and Silvey [1958]. We combine aspects of notation and regularity conditions from each work and introduce it here. Recall the weights are generated such that

$$(w_{n,1}, \dots, w_{n,n}) = \frac{n}{\sum_{j=1}^{n} Y_j}(Y_1, \dots, Y_n),$$

where n is the sample size, and each $Y_i \sim \text{Exp}(1)$ independently, admitting a Dirichlet$(1, \dots, 1)$ distribution for w_n/n. The typical empirical score and information functions, denoting the weighted versions with a tilde, are defined as

$$S_n(\theta) = \frac{1}{n}\sum_{i=1}^{n} \nabla_\theta \log f_\theta(X_i), \qquad \tilde{S}_n(\theta) = \frac{1}{n}\sum_{i=1}^{n} w_{n,i} \nabla_\theta \log f_\theta(X_i),$$

$$J_n(\theta) = -\frac{1}{n}\sum_{i=1}^{n} \nabla_\theta^2 \log f_\theta(X_i), \qquad \tilde{J}_n(\theta) = -\frac{1}{n}\sum_{i=1}^{n} w_{n,i} \nabla_\theta^2 \log f_\theta(X_i).$$

Elements of the Jacobian of the constraint functions and a variant (with expectation always with respect to θ_0) of the Fisher information are defined as

$$[D_h(\theta)]_{ij} = \frac{\partial h_i(\theta)}{\partial \theta_j}, \qquad [J(\theta)]_{ij} = -\mathbb{E}_{\theta_0}\left\{\frac{\partial^2 \log f_\theta(X)}{\partial \theta_i \partial \theta_j}\right\}.$$

Finally, the third-order partial derivatives of the log-likelihood (see Condition 5) are pertinent for Taylor expansions. Weighted sums of elements of the tensor are denoted

$$[\tilde{\Psi}^j_{1:n}(\theta)]_{k\ell} = \sum_{i=1}^{n} w_{n,i} \frac{\partial^3 \log f_\theta(X_i)}{\partial \theta_j \partial \theta_k \partial \theta_\ell}.$$

Conditional convergence is the end result of the theory. The definition is repeated for clarity.

Definition 1 (Convergence in Conditional Probability)**.** *Let U and $V_1, V_2, \dots$ be defined on the same probability space. Then V_n converges in conditional probability $a.s.[X_{1:\infty}]$ to U if for all $\epsilon > 0$*

$$\mathrm{pr}(\|V_n - U\|_2 > \epsilon \mid X_{1:n}) \to 0 \quad a.s.[X_{1:\infty}],$$

as $n \to \infty$. In shorthand, this property is denoted $V_n \to_{c.p.} U$ $a.s.[X_{1:\infty}]$ and applies for almost every infinite sample sequence.

There are several shared motifs in the regularity conditions of Newton [1991] and Aitchison and Silvey [1958] as many are used in order to have proper Taylor expansions around the score function. In the following conditions θ_0 refers to the true underlying value of θ.

Condition 1 (Identifiability)**.** *For any $\theta_1 \neq \theta_0$, there exists a set $\mathcal{D}$ where $P_{\theta_1}(\mathcal{D}) \neq P_{\theta_0}(\mathcal{D})$ and P_θ is the probability measure induced from probability density function f_θ. In addition,*

$$\mathbb{E}_{\theta_0}\left\{\frac{\log f_{\theta_0}(x)}{\log f_{\theta_1}(x)}\right\} < \infty.$$

Condition 2 (Feasibility)**.** *The support set, Θ, has a nonempty subset $\tilde{\Theta} = \{\theta\colon h(\theta) = 0\}$ and $\theta_0 \in \tilde{\Theta}$.*

Condition 3 (Neighborhood of θ_0)**.** *There is a closed ball $U_\alpha = \{\theta\colon \|\theta - \theta_0\|_2 \leq \alpha\} \subset \Theta$ around θ_0 for some α.*

Condition 4 (Smoothness of unconstrained prior)**.** *For every $\theta \in U_\alpha$, the function $\log \pi(\theta)$ is bounded and continuous. Furthermore, for every $\theta \in U_\alpha$, the following derivatives exist, are bounded, and are continuous,*

$$\frac{\partial \log \pi(\theta)}{\partial \theta_j}, \quad (j = 1, \dots, p).$$

Condition 5 (Smoothness of log-likelihood)**.** *For every $\theta \in U_\alpha$, the following derivatives exist for almost all x and are continuous with respect to θ,*

$$\frac{\partial \log f_\theta(x)}{\partial \theta_j}, \quad \frac{\partial^2 \log f_\theta(x)}{\partial \theta_j \partial \theta_k}, \quad \frac{\partial^3 \log f_\theta(x)}{\partial \theta_j \partial \theta_k \partial \theta_\ell}, \quad (j, k, \ell = 1, \dots, p).$$

Condition 6 (Boundedness of likelihood)**.** *For every $\theta \in U_\alpha$, for all $(j, k, \ell = 1, \dots, p)$, there exist functions $m^{(\cdot)}$ such that*

$$\begin{aligned}
\left|\frac{\partial f_\theta(x)}{\partial \theta_j}\right| &< m^{(1)}_j(x), \quad \int m^{(1)}_j(x)\mu(dx) < \infty, \\
\left|\frac{\partial^2 f_\theta(x)}{\partial \theta_j \partial \theta_k}\right| &< m^{(2)}_{jk}(x), \quad \int m^{(2)}_{jk}(x)\mu(dx) < \infty, \\
\left|\frac{\partial^2 \log f_\theta(x)}{\partial \theta_j \partial \theta_k}\right| &< m^{(3)}_{jk}(x), \quad \mathbb{E}_{\theta_0}\{m^{(3)}_{jk}(X)\} < \infty, \\
\left|\frac{\partial^3 \log f_\theta(x)}{\partial \theta_j \partial \theta_k \partial \theta_\ell}\right| &< m^{(4)}_{jk\ell}(x), \quad \mathbb{E}_{\theta_0}\{m^{(4)}_{jk\ell}(X)\} < \infty, \\
\left|\frac{\partial \log f_\theta(x)}{\partial \theta_j} \cdot \frac{\partial^2 \log f_\theta(x)}{\partial \theta_k \partial \theta_\ell}\right| &< m^{(5)}_{jk\ell}(x), \quad \mathbb{E}_{\theta_0}\{m^{(5)}_{jk\ell}(X)\} < \infty.
\end{aligned} \tag{22}$$

Condition 7 (Positive-definite information matrix)**.** *The Fisher information $J(\theta_0)$ is positive-definite with finite elements. Furthermore, $I(\theta) = \mathbb{E}_\theta[\{\nabla_\theta \log f_\theta(X)\}\{\nabla_\theta \log f_\theta(X)\}^\top]$ is positive-definite with finite elements for all for all $\theta \in U_\alpha$. Note that $I(\theta_0) = J(\theta_0)$ under Conditions 5 and 6 [Newton, 1991].*

Condition 8 (Smoothness of equality constraint function)**.** *For every $\theta \in U_\alpha$, the following derivatives exist and are continuous functions of θ,*

$$\frac{\partial h_k(\theta)}{\partial \theta_j}, \quad (j = 1, \dots, p;\ k = 1, \dots, r).$$

In addition, for every $\theta \in U_\alpha$, the following derivatives exist, are continuous functions of θ and are bounded above by some constant c_h,

$$\left| \frac{\partial^2 h_k(\theta)}{\partial \theta_j \partial \theta_\ell} \right| \leq c_h, \quad (j, \ell = 1, \dots, p;\ k = 1, \dots, r).$$

Condition 9 (Full-rank Jacobian at θ_0)**.** *The $r \times p$ matrix $D_h(\theta_0)$ has full row-rank.*

We quickly prove under the conditions that the covariance matrix from Theorem 2 is invariant to h.

Lemma 1. *Suppose $g(\theta)$ and $h(\theta)$ satisfy Conditions 2, 8 and 9. If*

$$U_h(\theta) = D_h^\top(\theta)\{D_h(\theta_0)J(\theta_0)^{-1}D_h^\top(\theta)\}^{-1}D_h(\theta_0)J(\theta_0)^{-1},$$

then $U_h(\theta_0) = U_g(\theta_0)$.

Proof. Let $\mathcal{F}_h = \{D_h(\theta_0)d = 0\}$, be the set of feasible directions at θ_0, under the definition of h (Definition 12.3 of Nocedal and Wright [2006]). Under Condition 9 and Lemma 12.2 of Nocedal and Wright [2006], $\mathcal{F}_h$ is equal to the tangent space of $\tilde{\Theta}$ at θ_0. The same is true of $\mathcal{F}_g$. Hence, $D_h(\theta_0)$ and $D_g(\theta_0)$ have same kernel and row space. Hence, there exists some invertible matrix $P \in \mathbb{R}^{r \times r}$ such that $D_h(\theta_0) = PD_g(\theta_0)$. Then

$$\begin{aligned} U_h(\theta_0) &= D_h^\top(\theta_0)\{D_h(\theta_0)J(\theta_0)^{-1}D_h^\top(\theta_0)\}^{-1}D_h(\theta_0)J(\theta_0)^{-1}, \\ &= D_g^\top(\theta_0)P^\top[P\{D_g(\theta_0)J(\theta_0)^{-1}D_g^\top(\theta_0)\}P^\top]^{-1}PD_g(\theta_0)J(\theta_0)^{-1}, \\ &= D_g^\top(\theta_0)P^\top(P^\top)^{-1}\{D_g(\theta_0)J(\theta_0)^{-1}D_g^\top(\theta_0)\}^{-1}P^{-1}PD_g(\theta_0)J(\theta_0)^{-1}, \\ &= U_g(\theta_0). \end{aligned}$$

□

Without proof we repeat the following useful lemmas from Newton [1991].

Lemma 2 (Lemma 3 of Newton [1991])**.** *Let the data X_i be drawn from density f_{θ_0} and let the random variables Y_i be independent, identically distributed from a unit exponential, and independent of X. If m is a real-valued, measurable function such that $\mathbb{E}_{\theta_0}\{m(X_i)\} < \infty$, then*

$$\frac{1}{n}\sum_{i=1}^{n} Y_i m(X_i) \to_{c.p.} \mathbb{E}_{\theta_0}\{m(X_i)\} \quad a.s.[X_{1:\infty}].$$

Lemma 3 (Lemmas 4 and 5 of Newton [1991])**.** *Under Conditions 1, 3, 5, and 6, as $n \to \infty$,*

$$\begin{aligned} \|\tilde{S}_n(\theta_0)\|_2 &\to_{c.p.} 0 \quad a.s.[X_{1:\infty}], \\ \|\tilde{J}_n(\theta) - J(\theta)\|_2 &\to_{c.p.} 0 \quad a.s.[X_{1:\infty}]. \end{aligned}$$

Lemma 4 (Lemma 8 of Newton [1991])**.** *Let $z \in \mathbb{R}^p$ be a unit vector. Let $a_{in} = z^\top \nabla_\theta\{\log f_{\hat{\theta}_n}(X_i)\}$, for a strongly consistent estimator, $\hat{\theta}_n$. Under Conditions 1, 3, 5, 6, and 7, as $n \to \infty$,*

$$\begin{aligned} \frac{1}{n}\sum_{i=1}^{n} a_{in}^2 &\to z^\top J(\theta_0) z \quad a.s.[X_{1:\infty}], \\ \frac{1}{n}\max_{i=1,\dots,n} a_{in}^2 &\to 0 \quad a.s.[X_{1:\infty}]. \end{aligned}$$

Lemma 5 (Theorem 15 of Newton [1991]). *Let $Z_1, Z_2, \dots$ be independent and identically distributed random variables with mean μ and variance σ^2. Let (a_{in}) be a non-vanishing sequence of constants satisfying*

$$\frac{\sum_{i=1}^{n} a_{in}^2}{\max_{i=1,\dots,n} a_{in}^2} \to \infty, \quad n \to \infty.$$

Then for $T_n = \sum_{i=1}^{n} a_{in} Z_i,\ \mu_n = \mu \sum_{i=1}^{n} a_{in}$, and $\sigma_n^2 = \sigma^2 \sum_{i=1}^{n} a_{in}^2$,

$$\frac{T_n - \mu_n}{\sigma_n}$$

converges to a standard normal distribution as $n \to \infty$.

B.2 PROOF OF PROPOSITION 1

Proof. First, the conditions in the proof statement are made rigorous by Condition 11 from Miller [2021] which we assume is satisfied. Under Theorem 12 of Miller [2021], with probability 1, $h_n(\theta) = -n^{-1} \sum_{i=1}^{n} \log f_\theta(x_i)$ satisfies the conditions of Theorem 5 of Miller [2021]. We now show that $g_n(\theta) = h_n(\theta) + \rho q(\theta)$ also satisfies these conditions with probability 1, which implies the proposition statement.

Continuity of g_n at θ_0 is followed by the assumed continuity of q at θ_0. In addition, there is some E, with $\theta_0 \in E$, which is open and bounded where h_n has continuous third derivatives which are uniformly bounded. Let B be the interior of the intersection of E and the closed ball around θ_0 where $q(\theta)$ is three-times continuously differentiable. Then, B is an open set containing θ_0 where g_n has continuous third derivatives. Furthermore, g_n has uniformly bounded third derivatives on the closure of B, as it is compact. This implies uniformly bounded third derivatives on B itself. Trivially, as h_n converges pointwise to some function h (established by Theorem 5 of Miller [2021]), then g_n converges pointwise to g where $g(\theta) = h(\theta) + \rho q(\theta)$. We also have $\nabla_\theta^2 h(\theta_0) = -\mathbb{E}_{\theta_0}\{\nabla_\theta^2 \log f_{\theta_0}(X)\}$ which is positive-definite via properties of natural exponential families. Hence, under assumption, $\nabla_\theta^2 g(\theta_0)$ is also positive-definite.

Finally we show Condition (1) of Theorem 5 of Miller [2021]. Apply Theorem 12 of Miller [2021] which says with probability one, for any open ball G with $\theta_0 \in G$ and $\mathrm{closure}(G) \subset \Theta$ that there exists some compact set K with θ_0 in the interior where $\liminf_n \inf_{\theta \in \Theta \setminus K} h_n(\theta) > h(\theta_0)$. Hence,

$$g(\theta_0) = h(\theta_0) < \liminf_n \inf_{\theta \in \Theta \setminus K} h_n(\theta) \le \liminf_n \inf_{\theta \in \Theta \setminus K} h_n(\theta) + \rho q(\theta) = \liminf_n \inf_{\theta \in \Theta \setminus K} g_n(\theta), \tag{23}$$

via $q(\theta) \ge 0$ and $q(\theta_0) = 0$. Also on K we have $h(\theta) > h(\theta_0)$ for all $\theta \in K \setminus \{\theta_0\}$. Because $q(\theta_0) = 0$ and $q(\theta) \ge 0$, this further implies $g(\theta) > g(\theta_0)$ for all $\theta \in K \setminus \{\theta_0\}$. $\square$

B.3 PROOF OF THEOREM 1

The following two lemmas of Aitchison and Silvey [1958] help establish the existence of a constrained weighted Bayesian bootstrap solution. The first is equivalent to Brouwer's fixed-point theorem, and the second is reproved for conditional probabilities.

Lemma 6 (Lemma 2 of Aitchison and Silvey [1958]). *If m is a continuous function mapping a closed ball to itself, such that for every $\|z\|_2 = 1$, $z^\top m(z) < 0$, then there exists a point $\hat{z}$ such that $\|\hat{z}\|_2 < 1$ and $m(z) = 0$.*

Lemma 7 (Analogue to Lemma 1 of Aitchison and Silvey [1958]). *Consider the Lagrangian multiplier system of the constrained weighted Bayesian bootstrap with no inequality constraints,*

$$0 = n\tilde{S}_n(\theta) + \nabla_\theta \log \pi(\theta) + D_h^\top(\theta)\lambda, \tag{24}$$

$$0 = h(\theta). \tag{25}$$

Subject to Conditions 1–9, for $\delta < \alpha$ and almost all w_n we have that $\check{\theta}_n \in U_\delta$ and $\check{\lambda}_n$ satisfy (24) *and* (25) *if and only if $\check{\theta}_n$ satisfies an equation of the form*

$$-J(\theta_0)(\check{\theta}_n - \theta_0) + \tilde{v}(\check{\theta}_n) = 0. \tag{26}$$

Furthermore, $\tilde{v}$ is continuous with respect to θ on U_δ for almost all $X_{1:n}$ and w_n. In addition, there exists a constant c_v (that does not depend on w_n, X, δ, or n) such that

$$\mathrm{pr}\left\{\sup_{\theta \in U_\delta} \|\tilde{v}(\theta)\|_2 \le \delta^2 c_v \mid X_{1:n}\right\} \to 1 \quad a.s.[X_{1:\infty}].$$

Proof. Let $\theta \in U_\delta$ with U_δ well defined from Conditions 1–3. Under Conditions 2, 5, and 8, perform a Taylor expansion of the weighted score function and the constraint function around θ_0 and define $\tilde{v}^{(1)}$ and $v^{(2)}$ such that

$$n\tilde{S}_n(\theta) = n\tilde{S}_n(\theta_0) - n\tilde{J}_n(\theta_0)(\theta - \theta_0) + \tilde{v}^{(1)}(\theta), \tag{27}$$

$$h(\theta) = D_h(\theta_0)(\theta - \theta_0) + v^{(2)}(\theta). \tag{28}$$

Components of $\tilde{v}^{(1)}(\theta) \in \mathbb{R}^p$ and $v^{(2)}(\theta) \in \mathbb{R}^r$ can be written as

$$\tilde{v}_j^{(1)}(\theta) = \frac{1}{2}(\theta - \theta_0)^\top \tilde{\Psi}_{1:n}^j(\theta^{(1,*)})(\theta - \theta_0), \tag{29}$$

$$v_j^{(2)}(\theta) = \frac{1}{2}(\theta - \theta_0)^\top \nabla_\theta^2 h_j(\theta^{(2,*)})(\theta - \theta_0), \tag{30}$$

where $\theta^{(\cdot,*)}$ are some vectors on the line segment between θ and θ_0. Substitute the expansions (27) and (28) into the Lagrangian system (24) and (25)

$$n\tilde{S}_n(\theta_0) - n\tilde{J}_n(\theta_0)(\theta - \theta_0) + D_h^\top(\theta)\lambda + \nabla_\theta \log \pi(\theta) + \tilde{v}^{(1)}(\theta) = 0, \tag{31}$$

$$D_h(\theta_0)(\theta - \theta_0) + v^{(2)}(\theta) = 0. \tag{32}$$

Now assume $\check{\theta}_n \in U_\delta$ and $\check{\lambda}_n$ satisfy (24) and (25), substitute $n\tilde{S}_n(\check{\theta}_n) = -D_h^\top(\check{\theta}_n)\lambda - \nabla_\theta \log \pi(\check{\theta}_n)$ into (31), $h(\check{\theta}_n)$ into (32) and then simplify:

$$n\tilde{S}_n(\theta_0) - n\tilde{J}_n(\theta_0)(\check{\theta}_n - \theta_0) + \tilde{v}^{(1)}(\check{\theta}_n) = n\tilde{S}_n(\check{\theta}_n), \tag{33}$$

$$D_h(\theta_0)(\check{\theta}_n - \theta_0) + v^{(2)}(\check{\theta}_n) = h(\check{\theta}_n) = 0. \tag{34}$$

Divide (33) by n and define $\tilde{v}^{(3)}(\theta)$ as

$$-J(\theta_0)(\check{\theta}_n - \theta_0) + \tilde{v}^{(3)}(\check{\theta}_n) = \tilde{S}_n(\check{\theta}_n) = -\frac{D_h^\top(\check{\theta}_n)\check{\lambda}_n}{n} - \frac{\nabla_\theta \log \pi(\check{\theta}_n)}{n}, \tag{35}$$

$$\tilde{v}^{(3)}(\theta) = \tilde{S}_n(\theta_0) - \{\tilde{J}_n(\theta_0) - J(\theta_0)\}(\theta - \theta_0) + \frac{\tilde{v}^{(1)}(\theta)}{n}. \tag{36}$$

Via the definition of $\tilde{v}^{(1)}(\theta)$ we also have

$$\tilde{v}^{(3)}(\theta) = \tilde{S}_n(\theta) + J(\theta_0)(\theta - \theta_0). \tag{37}$$

Multiply both sides of (35) by $D_h(\theta_0)J(\theta_0)^{-1}$ and substitute $D_h(\theta_0)(\check{\theta}_n - \theta_0)$ from (34) to get

$$v^{(2)}(\check{\theta}_n) + D_h(\theta_0)J(\theta_0)^{-1}\left\{\tilde{v}^{(3)}(\check{\theta}_n) + \frac{\nabla_\theta \log \pi(\check{\theta}_n)}{n}\right\} = -D_h(\theta_0)J(\theta_0)^{-1}D_h^\top(\check{\theta}_n)\frac{\check{\lambda}_n}{n}.$$

This gives an expression for $\check{\lambda}_n$

$$\check{\lambda}_n = -n\{D_h(\theta_0)J(\theta_0)^{-1}D_h^\top(\check{\theta}_n)\}^{-1}\left[D_h(\theta_0)J(\theta_0)^{-1}\left\{\tilde{v}^{(3)}(\check{\theta}_n) + \frac{\nabla_\theta \log \pi(\check{\theta}_n)}{n}\right\} + v^{(2)}(\check{\theta}_n)\right].$$

Substitute $\check{\lambda}_n$ into (35) to get

$$-J(\theta_0)(\check{\theta}_n - \theta_0) + \tilde{v}(\check{\theta}_n) = 0,$$

where

$$\begin{aligned}\tilde{v}(\theta) = &-D_h^\top(\theta)\{D_h(\theta_0)J(\theta_0)^{-1}D_h^\top(\theta)\}^{-1}\left[D_h(\theta_0)J(\theta_0)^{-1}\left\{\tilde{v}^{(3)}(\theta) + \frac{\nabla_\theta \log \pi(\theta)}{n}\right\} + v^{(2)}(\theta)\right] \\ &+ \tilde{v}^{(3)}(\theta) + \frac{\nabla_\theta \log \pi(\theta)}{n}.\end{aligned} \tag{38}$$

Using (24) for $\tilde{S}_n(\check{\theta}_n) = -D_h^\top(\check{\theta}_n)\check{\lambda}_n/n - \nabla_\theta \log \pi(\check{\theta}_n)/n$, (25) for $h(\check{\theta}_n) = 0$, and $\tilde{v}^{(3)}(\check{\theta}_n)$ from (37), it is clear $-J(\theta_0)(\check{\theta}_n - \theta_0) + \tilde{v}(\check{\theta}_n) = 0$.

Suppose now that $-J(\theta_0)(\theta-\theta_0)+\tilde{v}(\theta)=0$ and $\theta \in U_\delta$. Multiply this expression by $D_h(\theta_0)J(\theta_0)^{-1}$ to get

$$\begin{aligned}
0 &= D_h(\theta_0)J(\theta_0)^{-1}\{\tilde{v}(\theta)-J(\theta_0)(\theta-\theta_0)\}\\
&= -D_h^\top(\theta_0)J(\theta_0)^{-1}\left\{\tilde{v}^{(3)}(\theta)+\frac{\nabla_\theta \log\pi(\theta)}{n}\right\}-v^{(2)}(\theta)\\
&\quad + D_h(\theta_0)J(\theta_0)^{-1}\left\{\tilde{v}^{(3)}(\theta)+\frac{\nabla_\theta \log\pi(\theta)}{n}\right\}-D_h(\theta_0)(\theta-\theta_0)
\end{aligned}$$

Substituting (28) implies $h(\theta)=0$. Using $h(\theta)=0$ and (37) we can do similar substitution as above for

$$\tilde{S}_n(\theta)+\frac{\nabla_\theta \log\pi(\theta)}{n} = D_h^\top(\theta)\underbrace{\{D_h(\theta_0)J(\theta_0)^{-1}D_h^\top(\theta)\}^{-1}D_h(\theta_0)J(\theta_0)^{-1}\left\{\tilde{S}_n(\theta)+\frac{\nabla_\theta \log\pi(\theta)}{n}\right\}}_{-\lambda/n}.$$

This then implies $n\tilde{S}_n(\theta)+\nabla_\theta\log\pi(\theta)+D_h^\top(\theta)\lambda=0$, for some λ due to (26) being satisfied.

We now show for small δ that $\tilde{v}(\theta)$ is bounded and continuous for $\theta\in U_\delta$ with high conditional probability. We reiterate definitions for $\tilde{v}^{(1)}$, $v^{(2)}$, $\tilde{v}^{(3)}$ and show each is bounded. We can easily bound the components of $v^{(2)}$ defined in (30) due to Condition 8:

$$\begin{aligned}
v_j^{(2)}(\theta) &= \frac{1}{2}(\theta-\theta_0)^\top\nabla_\theta^2 h_j(\theta^{(2,*)})(\theta-\theta_0),\\
&\le \sup_{\theta'\in U_\delta}\frac{1}{2}|\lambda_{\max}\{\nabla_\theta^2 h_j(\theta')\}|\|\theta'-\theta\|_2^2,\\
&\le \frac{\delta^2 p c_h}{2}. \qquad (39)
\end{aligned}$$

Line (39) is due to the maximum eigenvalue being bounded by the magnitude of the entries times the number of rows from the Gershgorin circle theorem. This holds almost surely with respect to the random weights. Similarly for $\tilde{v}^{(1)}(\theta)$ defined in (29):

$$\begin{aligned}
|\tilde{v}_j^{(1)}(\theta)| &= \frac{1}{2}\left|\sum_{k=1}^p\sum_{\ell=1}^p[\theta-\theta_0]_k[\theta-\theta_0]_\ell\left\{\sum_{i=1}^n w_{n,i}\frac{\partial^3\log f_{\theta^*}(X_i)}{\partial\theta_j\partial\theta_k\partial\theta_\ell}\right\}\right|,\\
&\le \frac{1}{2}\|\theta-\theta_0\|_\infty^2\sum_{k=1}^p\sum_{\ell=1}^p\sum_{i=1}^n\left|w_{n,i}\frac{\partial^3\log f_{\theta^*}(X_i)}{\partial\theta_j\partial\theta_k\partial\theta_\ell}\right|,\\
&\le \frac{\delta^2}{2}\sum_{k=1}^p\sum_{\ell=1}^p\sum_{i=1}^n w_{n,i}m_{jk\ell}^{(4)}(X_i),\\
&= \frac{\delta^2}{2\bar{Y}}\sum_{k=1}^p\sum_{\ell=1}^p\frac{1}{n}\left\{\sum_{i=1}^n Y_i m_{jk\ell}^{(4)}(X_i)\right\}\to_{c.p.}\frac{\delta^2p^2c_4}{2}\quad a.s.[X_{1:\infty}].
\end{aligned}$$

Convergence in probability is due to (22) in Condition 6 and Lemma 2. The above implies

$$\begin{aligned}
\mathrm{pr}\left\{\|\tilde{v}^{(1)}(\theta)\|_\infty>\delta^2p^2c_4\mid X_{1:n}\right\} &= \mathrm{pr}\left\{\bigcup_{j=1}^p|\tilde{v}_j^{(1)}(\theta)|>\delta^2p^2c_4\mid X_{1:n}\right\},\\
&\le \sum_{j=1}^p\mathrm{pr}\left\{|\tilde{v}_j^{(1)}(\theta)|>\delta^2p^2c_4\mid X_{1:n}\right\},\\
&\le \sum_{j=1}^p\mathrm{pr}\left\{\frac{\delta^2}{2\bar{Y}}\sum_{k=1}^p\sum_{\ell=1}^p\frac{1}{n}\left\{\sum_{i=1}^n Y_i m_{jk\ell}^{(4)}(X_i)\right\}>\delta^2p^2c_4\mid X_{1:n}\right\},\\
&\to 0\quad a.s.[X_{1:\infty}].
\end{aligned}$$

Using the explicit definition of $\tilde{v}^{(3)}(\theta)$ from (36), Cauchy-Schwarz, triangle inequality, and norm equivalence imply

$$\|\tilde{v}^{(3)}(\theta)\|_2\le\|\tilde{S}_n(\theta_0)\|_2+\|\tilde{J}_n(\theta_0)-J(\theta_0)\|_2\|\theta-\theta_0\|_2+\frac{p^{1/2}}{n}\|\tilde{v}^{(1)}(\theta)\|_\infty.$$

Via Lemma 3 we can choose n large enough so $\mathrm{pr}\{\|\tilde{S}_n(\theta_0)\|_2 \geq \delta^2 \mid X_{1:n}\} < \eta/3$, $\mathrm{pr}\{\|\tilde{J}_n(\theta_0) - J(\theta_0)\|_2 \geq \delta \mid X_{1:n}\} < \eta/3$, and $\mathrm{pr}\{\|\tilde{v}^{(1)}(\theta)\|_\infty > \delta^2 p^2 c_4 \mid X_{1:n}\} < \eta/3$, for $\eta > 0$. This then implies via a standard union bound

$$\mathrm{pr}\left\{\|\tilde{v}^{(3)}(\theta)\|_2 > \delta^2\left(2 + \frac{p^{5/2}c_4}{2n}\right) \mid X_{1:n}\right\} \leq \eta.$$

We examine the terms of (38) to continue. We see $\nabla_\theta \log \pi$ is bounded and defined in U_δ via Conditions 3 and 4. Choosing n large enough will ensure $\nabla \log \pi(\theta)/n$ will have entries below δ^2. Bounding the operator norms of the matrices in (38) will further bound $\tilde{v}$. The operator norms of $D_h(\theta_0)$ and $J(\theta_0)$ are constants and hence bounded. Because D_h is continuous it will have bounded entries within the ball U_α. Since $\delta < \alpha$, the operator norm of $D_h(\theta)$ is uniformly bounded in U_δ. By choosing α small enough, $D_h(\theta)$ will also be full rank for all $\theta \in U_\alpha$ ensuring $\{D_h(\theta_0)J(\theta_0)^{-1}D_h^\top(\theta)\}^{-1}$ is nonsingular, has bounded entries, and bounded operator norm. All of these operator norms will factor into c_v. Given all components not containing w_n in (38) can be bounded below a factor of δ^2 for all $\theta \in U_\delta$, combined with (B.3) the ℓ_2-norm of the $\tilde{v}$ function for $\theta \in U_\delta$ is bounded by a factor of δ^2 with arbitrarily high conditional probability as n grows.

Continuity follows from similar arguments. Via (28), $v^{(2)}$ is continuous in U_δ, since h is also continuous in this neighborhood (Condition 8). Each component of $\tilde{v}^{(1)}(\theta)$ is continuous due to the continuity of the third partial derivatives of $\log f_\theta$ for all $\theta \in U_\delta$ (Condition 5). Continuity of $\tilde{v}^{(3)}$ follows from continuity of $\tilde{v}^{(1)}$. Both D_h and $\nabla_\theta \log \pi$ are continuous by the same conditions that have boundedness. This leaves showing $\{D_h(\theta_0)J(\theta_0)^{-1}D_h^\top(\theta)\}^{-1}$ is continuous on U_δ. We have that there is some δ where $D_h(\theta)$ is full row-rank for all $\theta \in U_\delta$, since $D_h(\theta_0)$ is full row-rank and D_h is continuous (Conditions 8, 9), all perturbations below a certain magnitude do not change the rank. As a result, $D_h(\theta_0)J(\theta_0)^{-1}D_h^\top(\theta)$ is full-rank and bounded in U_δ, for δ small enough, so the inverse is also continuous. Finally, all terms in the definition of $\tilde{v}$ on U_δ are continuous so it is as well. This holds for almost all w_n. □

We now prove Theorem 1.

Proof. We emulate the proof of Theorem 1 from Aitchison and Silvey [1958], while taking additional care for the conditional probabilities. Let c_v be defined in Lemma 7. Fix $\delta < \min[\lambda_{\min}\{J(\theta_0)\}/c_v, c_v^{-1/2}\lambda_{\max}^{1/2}\{J(\theta_0)\}, \alpha, 1]$. Let θ be such that $\|\theta - \theta_0\|_2 = \delta$. Define on the closed unit ball in $\mathbb{R}^p$,

$$m\left(\frac{\theta - \theta_0}{\delta}\right) = \frac{1}{2\lambda_{\max}\{J(\theta_0)\}}\{-J(\theta_0)(\theta - \theta_0) + \tilde{v}(\theta)\}. \tag{40}$$

Via Lemma 7, m is continuous for all $\theta \in U_\delta$ and almost all $X_{1:n}, w_n$. In addition, for all $\theta \in U_\delta$,

$$\left\|m\left(\frac{\theta - \theta_0}{\delta}\right)\right\|_2 \leq \frac{\|\theta - \theta_0\|_2}{2} + \frac{1}{2\lambda_{\max}\{J(\theta_0)\}}\|\tilde{v}(\theta)\|_2.$$

If $\|\tilde{v}(\theta)\|_2 < \lambda_{\max}\{J(\theta_0)\}$, then $\|m\|_2 \leq 1$ for all $\theta \in U_\delta$. Let $\epsilon < 1$, so via Lemma 7 there exists $n(\delta, \epsilon)$ large enough such that

$$\mathrm{pr}\left\{\sup_{\theta \in U_\delta} \|\tilde{v}(\theta)\|_2 \leq \delta^2 c_v \mid X_{1:n}\right\} > 1 - \epsilon/2, \quad (n = n(\delta, \epsilon), \ldots, \infty).$$

Then for $\delta < c_v^{-1/2}\lambda_{\max}^{1/2}\{J(\theta_0)\}$, we have $\|\tilde{v}(\theta)\|_2 < \lambda_{\max}\{J(\theta_0)\}$ with conditional probability at least $1 - \epsilon/2$. Hence, all m maps its domain to itself with conditional probability at least $1 - \epsilon/2$ for $n > n(\delta, \epsilon)$.

We now show that Lemma 6 can be applied with high conditional probability. With probability at least $1 - \epsilon/2$, for θ with $\|\theta - \theta_0\|_2 = \delta$,

$$\begin{aligned}\frac{1}{\delta}(\theta - \theta_0)^\top m\left(\frac{\theta - \theta_0}{\delta}\right) &= \frac{1}{2\delta\lambda_{\max}\{J(\theta_0)\}}\{-(\theta - \theta_0)^\top J(\theta_0)(\theta - \theta_0) + (\theta - \theta_0)^\top \tilde{v}(\theta)\},\\
&\leq \frac{1}{2\delta\lambda_{\max}\{J(\theta_0)\}}[-\lambda_{\min}\{J(\theta_0)\}\|\theta - \theta_0\|_2^2 + \|\theta - \theta_0\|_2\|\tilde{v}(\theta)\|_2],\\
&\leq \frac{1}{2\delta\lambda_{\max}\{J(\theta_0)\}}[-\lambda_{\min}\{J(\theta_0)\}\|\theta - \theta_0\|_2^2 + \delta^2 c_v\|\theta - \theta_0\|_2],\\
&= \frac{1}{2\lambda_{\max}\{J(\theta_0)\}}[-\delta\lambda_{\min}\{J(\theta_0)\} + \delta^2 c_v].\end{aligned} \tag{41}$$

Line (41) is negative when $\delta < \lambda_{\min}\{J(\theta_0)\}/c_v$. As a result, we can apply Lemma 6 and see there must exist a point $\check{\theta}_n$ where $\check{\theta}_n \in U_\delta$, but not on the boundary, and $m\{(\check{\theta}_n - \theta_0)/\delta\} = 0$. (40) is satisfied by $\check{\theta}_n$ so via Lemma 7, $\check{\theta}_n$ also satisfies (24) and (25) and this holds with arbitrarily high conditional probability as n grows. Furthermore, the Lagrange multipliers exist from Lemma 7. We immediately see $\check{\theta}_n$ is conditionally consistent because there exists $n(\delta, \epsilon)$ for arbitrarily small δ and ϵ such that both qualifiers for Lemma 6 are satisfied and

$$\mathrm{pr}(\|\check{\theta}_n - \theta_0\|_2 < \delta \mid X_{1:n}) > 1 - \epsilon, \tag{42}$$

for all $n > n(\delta, \epsilon)$. □

B.4 PROOF OF THEOREM 2

We start with an additional lemma to assist in the proof.

Lemma 8. *Let $U(\theta) = D_h^\top(\theta)\{D_h(\theta_0)J(\theta_0)^{-1}D_h^\top(\theta)\}^{-1}D_h(\theta_0)J(\theta_0)^{-1}$ be defined under Conditions 7 and 9. Let $\hat{\theta}_n$ be a strongly consistent estimator for θ_0 such that $h(\hat{\theta}_n) = 0$ almost surely and*

$$\|n^{1/2}\{I - U(\hat{\theta}_n)\}S_n(\hat{\theta}_n)\| \to 0 \quad a.s.[X_{1:\infty}]. \tag{43}$$

Furthermore, let $\tilde{M}_n$ be such that

$$\tilde{M}_n \to_{c.p.} J(\theta_0)^{-1}\{I - U(\theta_0)\} \quad a.s.[X_{1:\infty}]. \tag{44}$$

Then

$$n^{1/2}\tilde{M}_n\tilde{S}_n(\hat{\theta}_n) \mid X_{1:n} \Rightarrow N[0, J(\theta_0)^{-1}\{I - U(\theta_0)\}] \quad a.s.[X_{1:\infty}].$$

Proof. Let $t_n(z) = n^{1/2}z^\top\tilde{S}_n(\hat{\theta}_n)$ and $a_{in}(z) = z^\top\nabla_\theta\{\log f_{\hat{\theta}_n}(X_i)\}$. Rearranging terms we have

$$t_n(z) = n^{1/2}\sum_{j=1}^{p} z_j\left(\frac{\sum_{i=1}^{n} Y_i \frac{\partial}{\partial\theta_j}\log f_{\hat{\theta}_n}(X_i)}{\sum_{i=1}^{n} Y_i}\right) = \frac{1}{n^{1/2}\bar{Y}}\sum_{i=1}^{n} Y_i a_{in}(z).$$

The strong law of large numbers implies $\bar{Y}$ converges almost surely (with respect to the weights probability measure) to one, so $t_n(z)$ will converge in distribution similarly to $n^{-1/2}\sum_{i=1}^{n} Y_i a_{in}(z)$. Under Lemma 4 we have

$$\frac{1}{n}\sum_{i=1}^{n} a_{in}^2(z) \to z^\top J(\theta_0)z, \quad \frac{1}{n}\max_{i=1,\ldots,n} a_{in}^2(z) \to 0 \quad a.s.[X_{1:\infty}],$$

so the conditions to apply Lemma 5 hold $a.s.[X_{1:\infty}]$ on this quantity and as a result,

$$t_n(z) - n^{1/2}z^\top S_n(\hat{\theta}_n) \mid X_{1:n} \Rightarrow N\{0, z^\top J(\theta_0)z\} \quad a.s.[X_{1:\infty}].$$

Applying the Cramér-Wold theorem,

$$n^{1/2}\{\tilde{S}_n(\hat{\theta}_n) - S_n(\hat{\theta}_n)\} \mid X_{1:n} \Rightarrow N\{0, J(\theta_0)\} \quad a.s.[X_{1:\infty}].$$

Multiplying both sides by $\tilde{M}_n$ and applying a conditional version of Slutsky's lemma (along almost every sample path) with (44) gives

$$n^{1/2}\tilde{M}_n\{\tilde{S}_n(\hat{\theta}_n) - S_n(\hat{\theta}_n)\} \mid X_{1:n} \Rightarrow N[0, J(\theta_0)^{-1}\{I - U(\theta_0)\}] \quad a.s.[X_{1:\infty}].$$

To complete the proof we show the asymptotic distribution of $n^{1/2}\tilde{M}_n S_n(\hat{\theta}_n) \mid X_{1:n}$ is a point mass at zero. First, by (44), using conditional Slutsky's, for almost every sample path,

$$n^{1/2}\tilde{M}_n S_n(\hat{\theta}_n) \mid X_{1:n} \to_{c.p.} J(\theta_0)^{-1}\{I - U(\theta_0)\}S_n(\hat{\theta}_n) \quad a.s.[X_{1:\infty}].$$

As a consequence of (43), $n^{1/2}\{I - U(\theta_0)\}S_n(\hat{\theta}_n)$ is trivially conditionally consistent for 0. As a result, for almost every sample path, $n^{1/2}\tilde{M}_n S_n(\hat{\theta}_n)$ is conditionally consistent to 0 and the proof statement holds. □

We now prove the full theorem.

Proof. Using Condition 8, assume without loss of generality, that α is small enough such that $D_h(\theta)$ is full-rank with linearly independent rows for all $\theta \in U_\alpha$. Then Conditions 2 and 9 apply and a necessary condition for $\check{\theta}_n$ to be a local optima (as a sample from Algorithm 1) is satisfying the system

$$\begin{aligned} 0 &= \tilde{S}_n(\check{\theta}_n) + \frac{\nabla_\theta \log \pi(\check{\theta}_n)}{n} + \frac{D_h^\top(\check{\theta}_n)\check{\lambda}_n}{n}, \\ 0 &= h(\check{\theta}_n). \end{aligned} \tag{45}$$

Via Theorem 1 we have shown that $\check{\theta}_n$ is conditionally consistent for θ_0, so we can let n be large enough that $\check{\theta}_n \in U_\alpha$ with conditional probability at least $1-\epsilon$.

Using the strong consistency of $\hat{\theta}_n$ we can let n be large enough so Condition 5 applies almost surely, giving a Taylor expansion around $\hat{\theta}_n$ of $\tilde{S}_n(\theta)$ to get

$$\tilde{S}_n(\theta) = \tilde{S}_n(\hat{\theta}_n) - \tilde{J}_n(\hat{\theta}_n)(\theta - \hat{\theta}_n) + r(\theta), \tag{46}$$

where $r(\theta)$ has jth entry $\frac{1}{2}(\theta - \hat{\theta}_n)^\top \tilde{\Psi}^j_{1:n}(\theta^*_S)(\theta - \hat{\theta}_n)$ such that θ^*_S is some point on the line segment between θ and $\hat{\theta}_n$. Equation (46) can then be rearranged to

$$\tilde{S}_n(\hat{\theta}_n) = \tilde{S}_n(\theta) + \{\tilde{J}_n(\hat{\theta}_n) - R_n(\theta)\}(\theta - \hat{\theta}_n), \tag{47}$$

where $R_n \in \mathbb{R}^{p\times p}$ such that

$$\mathrm{row}_j\{R_n(\theta)\} = \frac{1}{2}(\theta - \hat{\theta}_n)^\top \tilde{\Psi}^j_{1:n}(\theta^*_S).$$

Substitute the stationary condition of (45) into (47) evaluated at $\check{\theta}_n$ for

$$\tilde{S}_n(\hat{\theta}_n) = -\frac{\nabla_\theta \log \pi(\check{\theta}_n)}{n} - \frac{D_h^\top(\check{\theta}_n)\check{\lambda}_n}{n} + \{\tilde{J}_n(\hat{\theta}_n) - R_n(\check{\theta}_n)\}(\check{\theta}_n - \hat{\theta}_n). \tag{48}$$

By definition $h(\hat{\theta}_n) = 0$. From Condition 8 we can expand h around $\hat{\theta}_n$ and putting $\check{\theta}_n$ into the expansion gives

$$\begin{aligned} 0 &= \{D_h^\top(\hat{\theta}_n) + Q_h(\check{\theta}_n)\}^\top(\check{\theta}_n - \hat{\theta}_n), \\ \mathrm{row}_j\{Q_h(\check{\theta}_n)\} &= (\check{\theta}_n - \hat{\theta}_n)^\top \nabla^2_\theta h_j(\theta^*_h), \end{aligned} \tag{49}$$

where θ^*_h is some point on the line segment between $\check{\theta}_n$ and $\hat{\theta}_n$.

Let

$$M_n(\hat{\theta}_n, \check{\theta}_n) = \begin{bmatrix} \tilde{J}_n(\hat{\theta}_n) - R_n(\check{\theta}_n) & -D_h^\top(\check{\theta}_n) \\ -\{D_h^\top(\hat{\theta}_n) + Q_h(\check{\theta}_n)\}^\top & 0 \end{bmatrix}.$$

In the proof of Theorem 7 of Newton [1991], from Conditions 1, 3, 5, 6, 7, and the strong consistency of $\hat{\theta}_n$, they show $\tilde{J}_n(\hat{\theta}_n) - R_n(\check{\theta}_n) \to_{c.p.} J(\theta_0)\ a.s.[X_{1:\infty}]$. Due to the continuity of D_h, Q_h (Condition 8) and the conditional consistency of $\check{\theta}_n$, the continuous mapping theorem implies $D_h(\check{\theta}_n) \to_{c.p.} D_h(\theta_0)\ a.s.[X_{1:\infty}]$ and $D_h(\hat{\theta}_n) + Q_h(\check{\theta}_n) \to_{c.p.} D_h(\theta_0)\ a.s.[X_{1:\infty}]$. From these conditional consistency results, we have

$$M_n(\hat{\theta}_n, \check{\theta}_n) \to_{c.p.} \begin{bmatrix} J(\theta_0) & -D_\star^\top \\ -D_\star & 0 \end{bmatrix} \quad a.s.[X_{1:\infty}],$$

where $D_\star^\top = D_h^\top(\theta_0)$. The matrix M_n converges to is nonsingular as $J(\theta_0)$ and $D_h(\theta_0)$ are full-rank.

We combine (48) and (49) and multiply by $n^{1/2}$ to get the following linear system

$$\begin{bmatrix} n^{1/2}\tilde{S}_n(\hat{\theta}_n) + n^{-1/2}\nabla_\theta \log \pi(\check{\theta}_n) \\ 0 \end{bmatrix} = M_n(\hat{\theta}_n, \check{\theta}_n) \begin{bmatrix} n^{1/2}(\check{\theta}_n - \hat{\theta}_n) \\ n^{-1/2}\check{\lambda}_n \end{bmatrix}.$$

On a set with arbitrarily high conditional probability we may write

$$M_n^{-1}(\hat{\theta}_n, \check{\theta}_n) \begin{bmatrix} n^{1/2}\tilde{S}_n(\hat{\theta}_n) + n^{-1/2}\nabla_\theta \log \pi(\check{\theta}_n) \\ 0 \end{bmatrix} = \begin{bmatrix} n^{1/2}(\check{\theta}_n - \hat{\theta}_n) \\ n^{-1/2}\check{\lambda}_n \end{bmatrix}.$$

The matrix inverse is continuous in neighborhoods of full-rank, and all functions in M_n are continuous, so we can apply the continuous mapping theorem for

$$M_n^{-1}(\hat{\theta}_n, \check{\theta}_n) \rightarrow_{c.p.} \begin{bmatrix} J(\theta_0) & -D_\star^\top \\ -D_\star & 0 \end{bmatrix}^{-1} = \begin{bmatrix} P & Q \\ Q^\top & R \end{bmatrix}, \tag{50}$$

where, using formulas for the inverse of block matrices [Petersen and Pedersen, 2012],

$$\begin{aligned} P &= J(\theta_0)^{-1}(I - U), \\ Q &= -J(\theta_0)^{-1} D_\star^\top \{D_\star J(\theta_0)^{-1} D_\star^\top\}^{-1}, \\ R &= -\{D_\star J(\theta_0)^{-1} D_\star^\top\}^{-1}, \\ U &= D_\star^\top \{D_\star J(\theta_0)^{-1} D_\star^\top\}^{-1} D_\star J(\theta_0)^{-1}. \end{aligned}$$

Let $\tilde{M}_n$ be the top-left block of $M_n^{-1}(\hat{\theta}_n, \check{\theta}_n)$, so $n^{1/2}(\check{\theta}_n - \hat{\theta}_n) = n^{1/2}\tilde{M}_n\{\tilde{S}_n(\hat{\theta}_n) + n^{-1/2}\nabla_\theta \log \pi(\check{\theta}_n)\}$. Applying Lemma 8 with (50) and noting $n^{-1/2}\nabla_\theta \log \pi(\check{\theta}_n)$ converges in conditional distribution to zero (from Condition 4), we have the proof statement. □