The Proximal Bootstrap for Constrained Estimators [*]

Jessie Li[†]

October 4, 2023

We demonstrate how to use the proximal bootstrap to conduct asymptotically valid inference for $\sqrt{n}$-consistent estimators defined as the solution to a constrained optimization problem with a possibly nonsmooth and nonconvex sample objective function and a constraint set defined by smooth equalities and/or inequalities which can be estimated from the data. We allow for the inequalities to drift towards equality as the sample size goes to infinity, and show how to use test-inversion to construct a uniformly asymptotically valid confidence set for the parameters. The proximal bootstrap estimator is typically much faster to compute than alternative bootstrap procedures because it can be written as the solution to a quadratic programming problem. Monte Carlo simulations illustrate the correct coverage of the proximal bootstrap in a boundary constrained maximum likelihood model, a boundary constrained nonsmooth GMM model, and a conditional logit model with estimated capacity constraints.

Keywords: bootstrap, non-standard asymptotics, constrained optimization, proximal mapping

JEL: C10; C15

1 Introduction

This paper considers using the proximal bootstrap estimator proposed in Li (2021) to conduct asymptotically valid inference for a large class of $\sqrt{n}$-consistent estimators with possibly nonstandard asymptotic distributions for which standard bootstrap procedures fail. The application

[*] I would like to thank Donald Andrews, Andres Aradillas-Lopez, Stéphane Bonhomme, Xiaohong Chen, Timothy Christensen, Jean-Jacques Forneron, Bulat Gafarov, Patrik Guggenberger, Marc Henry, Michael Jansson, Sung Jae Jun, Luofeng Liao, Joris Pinkse, Jack Porter, Demian Pouzo, Andres Santos, Xiaoxia Shi, Azeem Shaikh, Alex Torgovitsky, and Takuya Ura for helpful comments and suggestions.

[†] jeqli@ucsc.edu. Department of Economics, University of California Santa Cruz. 1156 High Street Santa Cruz, CA 95064.

which we will focus on in this paper is estimators defined by the solution to a constrained optimization problem with smooth inequality and/or equality constraints and a possibly nonsmooth and nonconvex sample objective function. A well-known example of a constrained estimator with a nonstandard distribution is the constrained MLE estimator where the true parameter lies on the boundary of the constraint set. It is well known (see e.g. Andrews (2000)) that applying a standard bootstrap procedure to estimate the distribution of the constrained estimator is inconsistent when the true parameters β_0 lie on the boundary of the constraint set C. An example of an inconsistent standard bootstrap procedure is the nonparametric bootstrap, which involves resampling the data with replacement, computing the constrained estimator on the resampled data sets, and then use the percentiles of these estimators to form confidence intervals.

Motivated by the optimization literature and recent contributions in computationally efficient bootstrap procedures (e.g. Kline and Santos (2012), Armstrong et al. (2014), Forneron and Ng (2019)), our proximal bootstrap estimator can be expressed as the solution to a convex optimization problem and efficiently computed starting from an initial consistent estimator using built-in and freely available software. The proximal bootstrap can consistently estimate the non-standard asymptotic distribution of constrained estimators when the parameters are on the boundary, but not drifting towards the boundary. When the parameters are drifting towards the boundary at an unknown rate, the proximal bootstrap typically cannot consistently replicate the estimator's distribution. However, we are still able to conduct uniformly asymptotically valid inference on the entire parameter vector using a confidence set constructed by inverting a test statistic based on the difference between two objectives. We can also conduct uniformly asymptotically valid inference on subvectors of the parameter vector using either projection or profiling of the objective functions. This idea of using test inversion to construct uniformly asymptotically valid confidence regions has similarities to the literature on partially identified models, for example, Chernozhukov et al. (2007), Romano and Shaikh (2008), Andrews and Guggenberger (2009), Andrews and Han (2009), Andrews and Guggenberger (2010), Andrews and Soares (2010), Bugni (2010), Canay (2010), and many others. However, we do not handle partial identification in this paper because our object of interest β_0 is assumed to be unique.

Another novel part of this paper is that we provide a general asymptotic distribution for estimators defined by the solution to constrained optimization problems where the Lagrangian admits a uniform local quadratic expansion in $\sqrt{n}$ neighborhoods of β_0. This local quadratic expansion rules out linear programming estimators and other estimators that have large flat regions near β_0. The asymptotic distribution is derived using ideas from the optimization literature and encompasses as special cases the results in Geyer (1994), Andrews (1999),Andrews (2000), and Andrews (2002a) for constrained estimators with non-random constraint sets and true parameters possibly lying on the boundaries of the constraint sets. Andrews (1999) derives the asymptotic distribution of constrained extremum estimators where the rescaled constraint set $\sqrt{n}\,(C - \beta_0)$ can be approximated by a convex cone. Geyer (1994) considers a more general case where the cone does not need to be convex.

Our paper was inspired by ideas in the optimization literature on sequential quadratic programming, where a local quadratic approximation is used to approximate the objective function on each iteration. The proximal bootstrap estimator is in effect applying such a local quadratic approximation centered around an initial $\sqrt{n}$-consistent estimate of the parameters. Because we want the estimation error from this initial estimate to be negligible in the proximal bootstrap approximation of our estimator's asymptotic distribution, we need to use a scaling sequence α_n that satisfies $\alpha_n \to 0$ and $\sqrt{n}\alpha_n \to \infty$. α_n will also serve as a selection device so that the active constraints are included in the asymptotic distribution while the inactive, non-drifting constraints are not. The α_n in this paper is similar to the ϵ_n in the numerical bootstrap Hong and Li (2020). However, we want to emphasize that the proximal bootstrap is a different procedure than the numerical bootstrap because it solves a different optimization problem. The proximal bootstrap works only for $\sqrt{n}$-consistent estimators but is more computationally efficient than the numerical bootstrap. Additionally, Hong and Li (2020) looked only at estimators with non-random constraints that do not depend on the data and did not consider drifting constraints. In the case of a smooth sample objective function without constraints, the proximal bootstrap is similar (but not identical) to the k-step bootstrap (for $k = 1$) proposed by Davidson and MacKinnon (1999) and investigated further by Andrews (2002b). The proximal bootstrap has an additional scaling factor of $\alpha_n\sqrt{n}$ in front of the inverse Hessian times Jacobian, which is different from the k-step bootstrap which uses a scaling of 1. There are also some similarities with the score bootstrap of Kline and Santos (2012) for unconstrained problems, but our method of inference is still different even when the constraints are not active. We have the additional scaling factor in front of the score and we can handle nonsmooth objectives.

The statistics literature contains many papers on constrained estimation such as Shapiro (1988), Shapiro (1989), Shapiro (1990), Knight (2001), Knight (2006), and Knight (2010). While several of these papers derive the non-standard asymptotic distributions of various constrained estimators, we did not see them propose a practical inference procedure as we do. Examples of econometrics papers on constrained estimation include Moon and Schorfheide (2009), Kaido and Santos (2014), Kaido (2016), Gafarov (2016), Chen et al. (2018), Hsieh et al. (2022), Kaido et al. (2019), Kaido et al. (2021), Horowitz and Lee (2019), Fang and Seo (2021), and Chernozhukov et al. (2023). While many of these papers are concerned with either conducting inference on the optimal value of the constrained optimization problem or testing whether the constraints are valid, we are interested in conducting inference on the optimal solution. Perhaps the closest paper to ours is Hsieh et al. (2022) who also consider inference for the optimal solution, but they focus on linear programming (LP) and convex quadratic programming (QP) problems with linear constraints. In contrast to Hsieh et al. (2022), we allow for nonconvex and nonlinear objective and constraint functions, but we do not handle linear programming or partially identified models. Our inference procedure is also different from theirs because we use resampling followed by inverting a test statistic while they exploit the fact that the primal-dual formulation of the Karush-Kuhn-Tucker (KKT) conditions can be written as a set of moment inequalities and then apply test inversion.

We offer simulation evidence supporting the uniform asymptotic validity of the proximal bootstrap test-inversion procedure. For a two-sided boundary constrained maximum likelihood model, we compare the empirical coverage frequencies of the proximal bootstrap test-inversion confidence interval to the intervals proposed by Hsieh et al. (2022), Fang and Santos (2019), subsampling, and the nonparametric bootstrap. The only methods that were uniformly valid across all drifting parameters were the proximal bootstrap and Hsieh et al. (2022), and we found in simulations that the proximal bootstrap is less conservative and has shorter average interval length. For a boundary constrained nonsmooth GMM model and a conditional logit model with estimated capacity constraints, we did not include a comparison with Hsieh et al. (2022) or Fang and Santos (2019) because we believe their methods do not apply. But we still compared the proximal bootstrap with subsampling and the nonparametric bootstrap and found that the proximal bootstrap achieves coverage close to the nominal level while subsampling and the nonparametric bootstrap undercover. In all simulations, we found that the coverage and average interval length of the proximal bootstrap test-inversion confidence interval were not sensitive to the choice of α_n.

The outline of our paper is as follows. Subsection 1.1 contains examples of constrained estimators and Subsection 1.2 contains the notation. Section 2 contains the main theoretical results. Subsection 2.1 shows pointwise consistency of the proximal bootstrap for constrained estimators with non-random constraint sets. Subsubsection 2.1.1 illustrates how to apply the proximal bootstrap for the Andrews (2000) example. In Subsection 2.2, by considering all rates of drift for the inequality constraints, we show how to conduct uniformly asymptotically valid inference by inverting a test statistic involving the objective function. Section 2.3 proposes a double bootstrap algorithm for choosing α_n. Section 2.4 contains an extension of our results to estimators with constraints that are estimated (meaning they depend on the data). Section 3 contains Monte Carlo simulation evidence demonstrating the uniform validity of the proximal bootstrap for a boundary constrained MLE model with a two-sided estimated constraint, a boundary constrained nonsmooth GMM model, and a conditional logit model with estimated capacity constraints. Section 4 concludes. Section 5 is the Appendix which contains proofs of the theorems.

1.1 Examples of Constrained Estimators

Example 1. An example of a constrained estimator with a non-random constraint set is the boundary constrained maximum likelihood estimator in Andrews (2000). Suppose we have a simple location model with i.i.d data:

$$y_i = \beta_0 + \epsilon_i, \quad \epsilon_i \sim N(0,1)$$

The maximum likelihood estimator subject to the constraint that $\beta \geqslant 0$ is

$$\hat{\beta}_n = \arg\min_{\beta \geqslant 0} \frac{1}{2n} \sum_{i=1}^{n} (y_i - \beta)^2$$

Example 2. Another example is a nonsmooth GMM estimator with a non-negativity constraint. Our model is

$$y_i = \beta_0 + \epsilon_i, \quad \epsilon_i \sim N(0,1)$$

For $\pi(\cdot,\beta) = [1(y_i \leqslant \beta) - \tau, y_i - \beta]'$ and $\hat{\pi}_n(\beta) = \left[\frac{1}{n}\sum_{i=1}^n 1(y_i \leqslant \beta) - 0.5, \frac{1}{n}\sum_{i=1}^n y_i - \beta\right]'$,

$$\hat{\beta}_n = \arg\min_{\beta \geqslant 0} \left\{ \hat{Q}_n(\beta) = \frac{1}{2}\hat{\pi}_n(\beta)' \hat{\pi}_n(\beta) \right\}$$

Example 3. An example involving an estimated constraint set is a conditional logit model with capacity constraints similar to the ones in de Palma et al. (2007) which state that the equilibrium demand for each housing unit should not exceed the supply of that housing unit. Let the choices be given by $y_{ij} = 1\left(y^*_{ij} > y^*_{ik} \forall k \neq j\right)$, where the utility of individual $i = 1...n$ from picking choice $j = 1...J$ is given by $y^*_{ij} = \beta_0 x_{ij} + \epsilon_{ij}$, where $\epsilon_{ij} \overset{i.i.d.}{\sim}$ Type 1 Extreme Value. The researcher observes the choices y_{ij} but not the utilities y^*_{ij} and would like to estimate the parameters β_0 using maximum likelihood. For $P_{ij}(\beta) \equiv \frac{\exp(\beta x_{ij})}{\sum_l \exp(\beta x_{il})}$,

$$\hat{\beta}_n = \arg\max_{\beta} \frac{1}{nJ}\sum_{i=1}^n \sum_{j=1}^J y_{ij} \ln P_{ij}(\beta)$$

$$\text{s.t. } \frac{1}{n}\sum_{i=1}^n P_{ij}(\beta) \leqslant \bar{b}_j \text{ for all } j = 1...J$$

1.2 Notation

Consider a random sample $\mathcal{X}_n = (X_1, X_2, ..., X_n)$ of independent draws from a probability measure P on a sample space $\mathcal{X}$. Define the empirical measure $P_n \equiv \frac{1}{n}\sum_{i=1}^n \delta_{X_i}$, where δ_x is the measure that assigns mass 1 at x and zero everywhere else. Denote the bootstrap empirical measure by $P^*_n = \frac{1}{n}\sum_{i=1}^n W_{ni}\delta_{X_i}$, which can refer to the multinomial, wild, or other exchangeable bootstraps. An exchangeable bootstrap requires that $W_n \equiv (W_{n1}, \dots, W_{nn})$ is an exchangeable vector of nonnegative weights which sum to 1. For the multinomial bootstrap, W_n is a multinomial random vector (independent of the data) with probabilities $(1/n, \dots, 1/n)$. For the wild bootstrap, $P^*_n = \frac{1}{n}\sum_{i=1}^n \left(\xi_i/\bar{\xi}_n\right)\delta_{X_i}$, where ξ_i are non-negative i.i.d. random variables (independent of the data) with finite third moments and $\bar{\xi}_n = \frac{1}{n}\sum_{i=1}^n \xi_i$. Weak convergence is defined in the sense of Kosorok (2007): $Z_n \rightsquigarrow Z$ in the metric space $(\mathbb{D}, d)$ if and only if $\sup_{f \in BL_1} |E^* f(Z_n) - Ef(Z)| \to 0$ where BL_1 is the space of functions $f : \mathbb{D} \mapsto \mathbb{R}$ with Lipschitz norm bounded by 1. $E^* f(Z_n)$ is the outer expectation of $f(Z_n)$, which is the infimum over all EU where U is measurable, $U \geqslant f(Z_n)$, and EU exists. Conditional weak convergence is also defined in the sense of Kosorok (2007): $Z_n \underset{\mathbb{W}}{\overset{\mathbb{P}}{\rightsquigarrow}} Z$ in the metric space $(\mathbb{D}, d)$ if and only if $\sup_{f\in BL_1} |E_{\mathbb{W}} f(Z_n) - Ef(Z)| \xrightarrow{p} 0$ and $E_{\mathbb{W}} f(Z_n)^* - E_{\mathbb{W}} f(Z_n)_* \xrightarrow{p} 0$ for all $f \in BL_1$, where BL_1 is the space of functions $f : \mathbb{D} \mapsto \mathbb{R}$ with Lipschitz norm bounded by 1, $E_{\mathbb{W}}$ denotes expectation with respect to the bootstrap weights $\mathbb{W}$ conditional on the data, and $f(Z_n)^*$ and $f(Z_n)_*$ denote measurable majorants and minorants with respect to the joint data (including the weights $\mathbb{W}$). Let $X_n^* = o_P^*(1)$ if $P(|X_n^*| > \epsilon|\mathcal{X}_n) = o_P(1)$ for all $\epsilon > 0$. Also define $M_n^* = O_P^*(1)$ (hence also $O_P(1)$) if $\lim_{m\to\infty} \limsup_{n\to\infty} P(P(M_n^* > m|\mathcal{X}_n) > \epsilon) \to 0\ \forall \epsilon > 0$.

2 Proximal Bootstrap

2.1 Proximal Bootstrap with non-estimated constraints

In this section, we consider constrained estimators with a finite number of non-estimated inequality and/or equality constraints $f_j(\beta)$ that are twice continuously differentiable over a compact parameter space $\mathbb{B} \subset \mathbb{R}^d$, where d is fixed. The non-random constraint set $C \subseteq \mathbb{B}$ is a closed subset of $\mathbb{B}$ and $\hat{Q}_n(\beta)$ is a possibly non-smooth, nonconvex function that converges uniformly to a function $Q(\beta)$ that is twice continuously differentiable at β_0, which is the true parameter on which we would like to conduct inference. We assume that our constraints C are correctly specified so that we can express $\beta_0 = \underset{\beta\in C}{\arg\min} Q(\beta)$, and we can estimate β_0 using

$$\hat{\beta}_n = \underset{\beta\in C}{\arg\min}\ \hat{Q}_n(\beta), \quad C = \{\beta \in \mathbb{B} : f_j(\beta) = 0 \text{ for } j \in \mathcal{E}, f_j(\beta) \leqslant 0 \text{ for } j \in \mathcal{I}\}$$

where $\mathcal{E}$ contains the indices of the equality constraints and $\mathcal{I}$ those of the inequality constraints. We will assume β_0 is the unique argmin of $Q(\beta)$ over C. We will show that the proximal bootstrap can consistently estimate the distribution of $\sqrt{n}\left(\hat{\beta}_n - \beta_0\right)$ both when β_0 lies in the interior and on the boundary of C, but not when it is drifting towards the boundary. Nevertheless, we will show in Section 2.2 by applying test-inversion, we can form a uniformly asymptotically valid confidence set.

Next, we define the proximal bootstrap estimator. For any $\bar{\beta}_n$ such that $\sqrt{n}\left(\bar{\beta}_n - \beta_0\right) = O_p(1)$, let $\bar{F}_{nj} \equiv \left.\frac{\partial f_j(\beta)}{\partial \beta}\right|_{\beta=\bar{\beta}_n}$ and $\bar{G}_{nj} \equiv \left.\frac{\partial^2 f_j(\beta)}{\partial \beta \partial \beta'}\right|_{\beta=\bar{\beta}_n}$ for all j, and let $\left\{\bar{\lambda}_{nj} \text{ for } j \in \mathcal{E} \cup \mathcal{I}\right\}$ be a set of optimal Lagrange multipliers obtained from solving for $\bar{\beta}_n$. These Lagrange multipliers can be obtained directly as outputs from the optimization algorithm used to compute $\bar{\beta}_n$. For any sequence α_n such that $\alpha_n \to 0$ and $\sqrt{n}\alpha_n \to \infty$, define $\hat{\beta}_n^* \equiv \underset{\beta \in C^*}{\arg\min} \hat{A}_n^*(\beta)$, where

$$\begin{aligned} \hat{A}_n^*(\beta) &\equiv \alpha_n\sqrt{n}\left(\hat{l}_n^*\left(\bar{\beta}_n\right) - \hat{l}_n\left(\bar{\beta}_n\right)\right)'\left(\beta - \bar{\beta}_n\right) + \frac{1}{2}\left\|\beta - \bar{\beta}_n\right\|_{\bar{H}_n}^2 + \frac{1}{2}\sum_{j \in \mathcal{E}\cup\mathcal{I}} \bar{\lambda}_{nj}\left\|\beta - \bar{\beta}_n\right\|_{\bar{G}_{nj}}^2 \\ C^* &= \left\{\beta \in \mathbb{B} : f_j\left(\bar{\beta}_n\right) + \bar{F}_{nj}'\left(\beta - \bar{\beta}_n\right) = 0 \text{ for } j \in \mathcal{E}, f_j\left(\bar{\beta}_n\right) + \bar{F}_{nj}'\left(\beta - \bar{\beta}_n\right) \leqslant 0 \text{ for } j \in \mathcal{I}\right\} \end{aligned} \quad (1)$$

Here, C^* is a linearization of C around $\bar{\beta}_n$, where $\bar{\beta}_n$ is an initial $\sqrt{n}$-consistent estimator of β_0, such as $\bar{\beta}_n = \hat{\beta}_n$. The sequence α_n ensures that $\bar{\beta}_n$'s asymptotic distribution does not enter into the proximal bootstrap estimator's asymptotic distribution. $\hat{l}_n\left(\bar{\beta}_n\right)$ is a consistent estimate of $l\left(\beta_0\right) \equiv \left.\frac{\partial Q_0(\beta)}{\partial\beta}\right|_{\beta=\beta_0}$ using $\bar{\beta}_n$, and $\hat{l}_n^*\left(\bar{\beta}_n\right)$ is a bootstrap (e.g. multinomial, wild) analog of $\hat{l}_n\left(\bar{\beta}_n\right)$. If $\hat{Q}_n(\beta)$ is differentiable, $\hat{l}_n\left(\bar{\beta}_n\right)$ can simply be the Jacobian of $\hat{Q}_n(\beta)$ evaluated at $\bar{\beta}_n$. More generally, to handle non-differentiable $\hat{Q}_n(\beta)$, $\hat{l}_n\left(\bar{\beta}_n\right)$ is a subgradient of $\hat{Q}_n(\beta)$ at $\bar{\beta}_n$, meaning that for any β, $\hat{Q}_n(\beta) - \hat{Q}_n\left(\bar{\beta}_n\right) \geqslant \hat{l}_n\left(\bar{\beta}_n\right)'\left(\beta - \bar{\beta}_n\right)$. $\bar{H}_n$ is a consistent estimate of the population Hessian $H_0 \equiv \left.\frac{\partial^2 Q_0(\beta)}{\partial\beta\partial\beta'}\right|_{\beta=\beta_0}$ constructed using $\bar{\beta}_n$.

We now discuss why we named the procedure the proximal bootstrap. Given a function $r : \mathbb{D} \mapsto \mathbb{R}$ and a symmetric positive definite matrix H, the scaled proximal mapping of r is the operator given by, for $\|\beta - z\|_H^2 = (\beta - z)' H (\beta - z)$,

$$prox_{H,r}(z) = \underset{\beta \in \mathbb{D}}{\arg\min}\left\{r(\beta) + \frac{1}{2}\|\beta - z\|_H^2\right\} \text{ for any } z \in \mathbb{D}$$

We can equivalently express the proximal bootstrap estimator using a scaled proximal map as

$$\begin{aligned} \hat{\beta}_n^* &= prox_{\bar{B}_n, \infty 1(\cdot \notin C^*)}\left(\bar{\beta}_n - \alpha_n\sqrt{n}\bar{B}_n^{-1}\left(\hat{l}_n^*\left(\bar{\beta}_n\right) - \hat{l}_n\left(\bar{\beta}_n\right)\right)\right) \\ &= \underset{\beta \in \mathbb{R}^d}{\arg\min}\left\{\infty 1\left(\beta \notin C^*\right) + \alpha_n\sqrt{n}\left(\hat{l}_n^*\left(\bar{\beta}_n\right) - \hat{l}_n\left(\bar{\beta}_n\right)\right)'\left(\beta - \bar{\beta}_n\right) + \frac{1}{2}\left\|\beta - \bar{\beta}_n\right\|_{\bar{B}_n}^2\right\} \end{aligned}$$

where $\bar{B}_n = \bar{H}_n + \sum_{j \in \mathcal{E}\cup\mathcal{I}} \bar{\lambda}_{nj}\bar{G}_{nj}$ and $\infty 1\left(\beta \notin C^*\right)$ evaluates to ∞ if β does not lie in C^* and evaluates to 0 otherwise. Because C^* is a closed, convex set, $\infty 1\left(\beta \notin C^*\right)$ will be a proper closed, convex function. The intuition for the proximal bootstrap is that $\hat{\beta}_n^*$ is the point inside C^* that is closest to, or "proximal" to, $\bar{\beta}_n - \alpha_n\sqrt{n}\bar{B}_n^{-1}\left(\hat{l}_n^*\left(\bar{\beta}_n\right) - \hat{l}_n\left(\bar{\beta}_n\right)\right)$. Note that the proximal bootstrap estimator is the solution to a quadratic programming problem, which is a convex problem if $\bar{B}_n$ is positive definite. This quadratic programming problem can be substantially faster to solve than the original constrained problem used to compute $\hat{\beta}_n$. Therefore, our proximal bootstrap estimator has a computational advantage over the standard bootstrap in cases where the standard bootstrap is consistent.

2.1.1 Proximal Bootstrap in Andrews (2000) Example

Before we go into the technical details, we illustrate how to apply the proximal bootstrap to the boundary constrained maximum likelihood estimator in Andrews (2000) (example 1):

$$\hat{\beta}_n = \arg\min_{\beta \geqslant 0} \left\{ \hat{Q}_n(\beta) = \frac{1}{2n} \sum_{i=1}^{n} (y_i - \beta)^2 \right\} = \max(\bar{y}_n, 0)$$

Andrews (2000) shows that the asymptotic distribution of $\sqrt{n}\left(\hat{\beta}_n - \beta_0\right)$ under pointwise asymptotics is given by

$$\sqrt{n}\left(\hat{\beta}_n - \beta_0\right) \rightsquigarrow \begin{cases} \arg\min\limits_{\{h:h\geqslant 0\}} \left\{h'W_0 + \frac{1}{2}h'H_0h\right\} = \max\left\{-H_0^{-1}W_0, 0\right\} & , \text{ if } \beta_0 = 0 \\ \arg\min\limits_{\{h:h\geqslant -\infty\}} \left\{h'W_0 + \frac{1}{2}h'H_0h\right\} = -H_0^{-1}W_0 & , \text{ if } \beta_0 > 0 \end{cases}$$

Andrews (2000) shows that the asymptotic distribution of $\sqrt{n}\left(\hat{\beta}_n^{\text{boot}} - \hat{\beta}_n\right)$, where $\hat{\beta}_n^{\text{boot}} = \max(\bar{y}_n^*, 0)$ is the standard nonparametric bootstrap estimator using the resampled sample mean $\bar{y}_n^*$, will not coincide with the asymptotic distribution of $\sqrt{n}\left(\hat{\beta}_n - \beta_0\right)$ when $\beta_0 = 0$. We now show that the proximal bootstrap estimator will consistently estimate the asymptotic distribution of $\sqrt{n}\left(\hat{\beta}_n - \beta_0\right)$ both when $\beta_0 = 0$ and when $\beta_0 > 0$, but is not drifting towards 0. Even though the proximal bootstrap is unable to consistently replicate the asymptotic distribution for drifting parameters, we are still able to use test-inversion to construct a uniformly asymptotically valid confidence set, as will be discussed in Section 2.2.

In this example, the sample objective is differentiable, so $\hat{l}_n(\beta)$ is simply the sample Jacobian.

$$l(\beta_0) = -\mathbb{E}[y_i - \beta_0]$$

$$\hat{l}_n(\beta_0) = -\frac{1}{n}\sum_{i=1}^{n}(y_i - \beta_0) \equiv -\bar{y}_n + \beta_0$$

$$\sqrt{n}\left(\hat{l}_n(\beta_0) - l(\beta_0)\right) = -\sqrt{n}(P_n - P)\, y_i \rightsquigarrow N(0, Var(y_i))$$

Our initial estimator is typically $\bar{\beta}_n = \hat{\beta}_n$, but it can also be some other $\sqrt{n}$-consistent estimator such as $\max\left(\frac{1}{n/2}\sum_{i=1}^{n/2} y_i, 0\right)$. One way to construct our proximal bootstrap estimator is by using the multinomial bootstrap for the Jacobian:

$$\begin{aligned}\hat{l}_n^*\left(\bar{\beta}_n\right) &= -\frac{1}{n}\sum_{i=1}^{n}\left(y_i^* - \bar{\beta}_n\right) \equiv -\bar{y}_n^* + \bar{\beta}_n \\ \hat{l}_n^*\left(\bar{\beta}_n\right) - \hat{l}_n\left(\bar{\beta}_n\right) &= -\left(P_n^* - P_n\right) y_i\end{aligned}$$

where $P_n^* = \frac{1}{n}\sum_{i=1}^{n} W_{ni}\delta_{X_i}$, and W_n is a multinomial random vector (independent of the data) with probabilities $(1/n, \ldots, 1/n)$.

Alternatively, we can use the wild bootstrap to estimate the Jacobian:

$$\begin{aligned}\hat{l}_n^*\left(\bar{\beta}_n\right) - \hat{l}_n\left(\bar{\beta}_n\right) &= -\frac{1}{n}\sum_{i=1}^{n}\left(\frac{\xi_i}{\bar{\xi}_n} - 1\right)\left(y_i - \bar{\beta}_n\right) \\ &= -\frac{1}{n}\sum_{i=1}^{n}\left(\frac{\xi_i}{\bar{\xi}_n} - 1\right) y_i + \underbrace{\bar{\beta}_n\left(\frac{\frac{1}{n}\sum_{i=1}^{n}\xi_i}{\bar{\xi}_n} - 1\right)}_{0} \\ &= -\left(P_n^* - P_n\right) y_i\end{aligned}$$

where $P_n^* = \frac{1}{n}\sum_{i=1}^{n}\left(\xi_i/\bar{\xi}_n\right)\delta_{X_i}$, and ξ_i are non-negative i.i.d. random variables (independent of the data) with finite third moments and $\bar{\xi}_n = \frac{1}{n}\sum_{i=1}^{n}\xi_i$.

For both the multinomial and wild bootstrap, $\sqrt{n}\left(\hat{l}_n^*\left(\beta_0\right) - \hat{l}_n\left(\beta_0\right)\right) \underset{\mathbb{W}}{\overset{\mathbb{P}}{\rightsquigarrow}} N\left(0, Var\left(y_i\right)\right) \equiv W_0$ and $\sqrt{n}\left(\hat{l}_n^*\left(\bar{\beta}_n\right) - \hat{l}_n\left(\bar{\beta}_n\right)\right) \underset{\mathbb{W}}{\overset{\mathbb{P}}{\rightsquigarrow}} W_0$, and additionally, $\sqrt{n}\left(\hat{l}_n\left(\beta_0\right) - l\left(\beta_0\right)\right) \rightsquigarrow W_0$. Thus we can use existing bootstrap procedures to estimate the distribution of the Jacobian. To estimate the distribution of the constrained estimator, we need to consider the fact that the constraint may be binding. For this example, the proximal bootstrap estimator is a scaled Newton step from an initial $\sqrt{n}$-consistent estimator, subject to a non-negativity constraint. The sequence α_n ensures that $\bar{\beta}_n$'s asymptotic distribution does not enter into the asymptotic distribution of $\hat{\beta}_n^*$, which in this example is given by

$$\hat{\beta}_n^* = \underset{\beta \geqslant 0}{\arg\min}\left\{\alpha_n\sqrt{n}\left(\hat{l}_n^*\left(\bar{\beta}_n\right) - \hat{l}_n\left(\bar{\beta}_n\right)\right)'\left(\beta - \bar{\beta}_n\right) + \frac{1}{2}\left\|\beta - \bar{\beta}_n\right\|_{\bar{H}_n}^2\right\}$$

$$
\begin{aligned}
&= \arg\min_{\beta \geqslant 0} \left\{ \alpha_n \sqrt{n} \left(\bar{y}_n - \bar{y}_n^* \right) \left(\beta - \bar{\beta}_n \right) + \frac{1}{2} \left(\beta - \bar{\beta}_n \right)^2 \right\} \\
&= \max \left(\bar{\beta}_n + \alpha_n \sqrt{n} \left(\bar{y}_n^* - \bar{y}_n \right), 0 \right)
\end{aligned}
$$

Note that since $\sqrt{n}\alpha_n \to \infty$, $\frac{\hat{\beta}_n - \beta_0}{\alpha_n} \xrightarrow{p} 0$ and the asymptotic distribution of $\frac{\hat{\beta}_n^* - \hat{\beta}_n}{\alpha_n}$ is same as the asymptotic distribution of $\frac{\hat{\beta}_n^* - \beta_0}{\alpha_n}$, which equals the asymptotic distribution of $\sqrt{n}\left(\hat{\beta}_n - \beta_0\right)$ under pointwise asymptotics:

$$
\begin{aligned}
\frac{\hat{\beta}_n^* - \beta_0}{\alpha_n} &= \arg\min_{\left\{h : \frac{\beta_0}{\alpha_n} + h \geqslant 0\right\}} \left\{ \sqrt{n} \left(\hat{l}_n^* \left(\bar{\beta}_n \right) - \hat{l}_n \left(\bar{\beta}_n \right) \right)' \left(\frac{\beta_0 - \bar{\beta}_n}{\alpha_n} + h \right) + \frac{1}{2} \left\| \frac{\beta_0 - \bar{\beta}_n}{\alpha_n} + h \right\|_{\bar{H}_n}^2 \right\} \\
&= \arg\min_{\left\{h : h \geqslant -\frac{\beta_0}{\alpha_n}\right\}} \left\{ \sqrt{n} \left(\hat{l}_n^* \left(\bar{\beta}_n \right) - \hat{l}_n \left(\bar{\beta}_n \right) \right)' h + \frac{1}{2} h' \bar{H}_n h + o_p^* (1) \right\} \\
&\overset{\mathbb{P}}{\underset{\mathbb{W}}{\rightsquigarrow}} \begin{cases} \arg\min\limits_{\{h : h \geqslant 0\}} \left\{ h' W_0 + \frac{1}{2} h' H_0 h \right\} = \max \left\{ -H_0^{-1} W_0, 0 \right\} & , \text{ if } \beta_0 = 0 \\ \arg\min\limits_{\{h : h \geqslant -\infty\}} \left\{ h' W_0 + \frac{1}{2} h' H_0 h \right\} = -H_0^{-1} W_0 & , \text{ if } \beta_0 > 0 \end{cases}
\end{aligned}
$$

where $H_0 = 1$ and $W_0 \sim N\left(0, Var\left(y_i\right)\right)$ for this example. Note that $\alpha_n \to 0$ also serves a selection device so that when the constraint $\beta_0 \geqslant 0$ is active, it enters into the asymptotic distribution, but when it is inactive (and β_0 is not drifting towards zero), it has no impact on the asymptotic distribution.

If we didn't have $\sqrt{n}\alpha_n \to \infty$, then $\frac{\beta_0 - \bar{\beta}_n}{\alpha_n} \rightsquigarrow \mathcal{Z}$ will not be $o_P(1)$ and $\mathcal{Z}$ will enter into the proximal bootstrap's asymptotic distribution:

$$
\begin{aligned}
\frac{\hat{\beta}_n^* - \beta_0}{\alpha_n} &= \arg\min_{\left\{h : \frac{\beta_0}{\alpha_n} + h \geqslant 0\right\}} \left\{ \sqrt{n} \left(\hat{l}_n^* \left(\bar{\beta}_n \right) - \hat{l}_n \left(\bar{\beta}_n \right) \right)' \left(\frac{\beta_0 - \bar{\beta}_n}{\alpha_n} + h \right) + \frac{1}{2} \left\| \frac{\beta_0 - \bar{\beta}_n}{\alpha_n} + h \right\|_{\bar{H}_n}^2 \right\} \\
&\overset{\mathbb{P}}{\underset{\mathbb{W}}{\rightsquigarrow}} \begin{cases} \arg\min\limits_{\{h : h \geqslant 0\}} \left\{ h' W_0 + \mathcal{Z}' W_0 + \frac{1}{2} \left(h + \mathcal{Z} \right)' H_0 \left(h + \mathcal{Z} \right) \right\} = \max \left\{ -H_0^{-1} W_0 - \mathcal{Z}, 0 \right\} & , \text{ if } \beta_0 = 0 \\ \arg\min\limits_{\{h : h \geqslant -\infty\}} \left\{ h' W_0 + \mathcal{Z}' W_0 + \frac{1}{2} \left(h + \mathcal{Z} \right)' H_0 \left(h + \mathcal{Z} \right) \right\} = -H_0^{-1} W_0 - \mathcal{Z} & , \text{ if } \beta_0 > 0 \end{cases}
\end{aligned}
$$

2.1.2 Assumptions

We now list some technical assumptions and discuss them afterwards.

Assumption 1. *(i)* $\mathbb{B} \subset \mathbb{R}^d$ *is compact,* $C \subseteq \mathbb{B}$ *is closed, and* d *is fixed.*

(ii) $\hat{\beta}_n$ *satisfies* $\hat{Q}_n\left(\hat{\beta}_n\right) \leqslant \inf_{\beta \in C} \hat{Q}_n(\beta) + o_p(1)$.

(iii) β_0 *is the unique value of* $\arg\min_{\beta \in C} Q(\beta)$.

(iv) $Q(\beta)$ *is twice continuously differentiable at* β_0*, and* $\sup_{\beta \in \mathbb{B}} \left|\hat{Q}_n(\beta) - Q(\beta)\right| = o_P(1)$.

Assumption 2. *(i) There exists a function* $g : \mathcal{X} \mapsto \mathbb{R}^d$ *indexed by a parameter* $\beta \in \mathbb{R}^d$ *such that for any* $\beta \in \mathbb{R}^d$, $\sqrt{n}\left(\hat{l}_n(\beta) - l(\beta)\right) = \sqrt{n}(P_n - P)g(\cdot, \beta) + o_P(1)$ *and* $\sqrt{n}\left(\hat{l}^*_n(\beta) - \hat{l}_n(\beta)\right) = \sqrt{n}(P^*_n - P_n)g(\cdot, \beta) + o^*_P(1)$*, where* $\lim_{n\to\infty} P\|g(\cdot, \beta_0)\|^2 1(\|g(\cdot, \beta_0)\| > \epsilon\sqrt{n}) = 0$ *for each* $\epsilon > 0$.

(ii) $\mathcal{G}_R \equiv \{g(\cdot, \beta) - g(\cdot, \beta_0) : \|\beta - \beta_0\| \leqslant R\}$ *is a Donsker class for some* $R > 0$ *and* $P\|g(\cdot, \beta) - g(\cdot, \beta_0)\|^2 \to 0$ *for* $\beta \to \beta_0$.

Assumption 3. $\lim_{\lambda\to\infty} \limsup_{n\to\infty} \sup_{t\geqslant\lambda} t^2 P\left\{\sup_{g(\cdot,\beta)\in\mathcal{G}_{\delta_n}} \left\|\frac{g(\cdot,\beta)-g(\cdot,\beta_0)}{1+\sqrt{n}\|\beta-\beta_0\|}\right\| > t\right\} = 0$ *for any* $\delta_n \to 0$.

Assumption 4. *Suppose Linear Independence Constraint Qualification (LICQ) holds at* β_0 *: the gradients of the active constraints* $F_{0j} \equiv \left.\frac{\partial f_j(\beta)}{\partial \beta}\right|_{\beta=\beta_0}$ *for* $j \in \mathcal{E} \cup \mathcal{I}^*$*, where* $\mathcal{I}^* \equiv \{j \in \mathcal{I} : f_j(\beta_0) = 0\}$*, are linearly independent.*

Assumption 5. *Suppose* $f_j : \mathbb{B} \mapsto \mathbb{R}$ *for all* $j \in \mathcal{E} \cup \mathcal{I}$ *are twice continuously differentiable functions. Let* λ_{0j} *be the unique Lagrange multipliers that satisfy* $\lambda_{0j} f_j(\beta_0) = 0$ *for all* $j \in \mathcal{E} \cup \mathcal{I}$, $\lambda_{0j} \geqslant 0$ *for all* $j \in \mathcal{E} \cup \mathcal{I}$*, and* $\nabla\mathcal{L}(\beta_0, \lambda_0) \equiv l(\beta_0) + \sum_{j\in\mathcal{E}\cup\mathcal{I}} \lambda_{0j} F_{0j} = 0$*, where* $\mathcal{L}(\beta_0, \lambda_0) \equiv Q(\beta) + \sum_{j\in\mathcal{E}\cup\mathcal{I}} \lambda_{0j} f_j(\beta)$ *and* $F_{0j} \equiv \left.\frac{\partial f_j(\beta)}{\partial \beta}\right|_{\beta=\beta_0}$*. Define* $\tilde{\mathcal{L}}_n(\beta) \equiv \hat{Q}_n(\beta) + \sum_{j\in\mathcal{E}\cup\mathcal{I}} \lambda_{0j} f_j(\beta)$*, and* $G_{0j} \equiv \left.\frac{\partial^2 f_j(\beta)}{\partial\beta\partial\beta'}\right|_{\beta=\beta_0}$*. Then for any* $\delta_n \to 0$*, and* $\mathcal{B}_{\delta_n} = \left\{h \in \mathbb{R}^d : \frac{\|h\|}{\sqrt{n}} \leqslant \delta_n\right\}$,

$$\sup_{h\in\mathcal{B}_{\delta_n}} \left|\frac{n\tilde{\mathcal{L}}_n\left(\beta_0 + \frac{h}{\sqrt{n}}\right) - n\tilde{\mathcal{L}}_n(\beta_0) - h'\sqrt{n}\hat{l}_n(\beta_0) - \frac{1}{2}h'H_0h - \sum_{j\in\mathcal{E}\cup\mathcal{I}} \lambda_{0j}\left(\sqrt{n}F'_{0j}h + \frac{1}{2}h'G_{0j}h\right)}{1 + \|h\|^2}\right|$$

$$= o_P(1)$$

Assumption 1 is a standard assumption for showing consistency of $\hat{\beta}_n$ for β_0. Assumption 2 allows us to apply Theorem 2.6 of Kosorok (2007) and show that $\sqrt{n}\left(\hat{l}_n(\beta_0) - l(\beta_0)\right)$ and

$\sqrt{n}\left(\hat{l}_n^*(\beta_0)-\hat{l}_n(\beta_0)\right)$ have the same asymptotic distribution. In the case of the wild bootstrap $P_n^*=\frac{1}{n}\sum_{i=1}^n\left(\xi_i/\bar{\xi}_n\right)\delta_{X_i}$, we need to ensure that the weights ξ_i have mean equal to variance.

We use Assumption 3 to show bootstrap equicontinuity which will imply $\sqrt{n}\left(\hat{l}_n^*(\bar{\beta}_n)-\hat{l}_n(\bar{\beta}_n)\right)$ and $\sqrt{n}\left(\hat{l}_n^*(\beta_0)-\hat{l}_n(\beta_0)\right)$ have the same asymptotic distribution. Assumption 2(ii) will imply stochastic equicontinuity, which in combination with the envelope function integrability condition in Assumption 3 will imply bootstrap equicontinuity (see Lemma 4.2 of Wellner and Zhan (1996)). A sufficient condition for Assumption 3 is that $\sup_{g(\cdot,\beta)\in\mathcal{G}_{\delta_n}}\left\|\frac{g(\cdot,\beta)-g(\cdot,\beta_0)}{1+\sqrt{n}\|\beta-\beta_0\|}\right\|\leqslant\kappa$ for some constant $\kappa>0$ and any $\delta_n\to 0$.

For the Andrews (2000) example (example 1), Assumptions 2 and 3 are satisfied. Note that we showed earlier that $\sqrt{n}\left(\hat{l}_n(\beta)-l(\beta)\right)=\sqrt{n}\left(P_n-P\right)g(\cdot,\beta)+o_P(1)$ and $\sqrt{n}\left(\hat{l}_n^*(\beta)-\hat{l}_n(\beta)\right)=\sqrt{n}\left(P_n^*-P_n\right)g(\cdot,\beta)+o_P^*(1)$, where $g(\cdot,\beta)=-(y_i-\beta)$. Since $g(\cdot,\beta)-g(\cdot,\beta_0)=-(y_i-\beta)+(y_i-\beta_0)=\beta-\beta_0$, $\mathcal{G}_R\equiv\{g(\cdot,\beta)-g(\cdot,\beta_0):|\beta-\beta_0|\leqslant R\}$ for any $R>0$ is a fixed function class and therefore also a Donsker class, and $P\left|g(\cdot,\beta)-g(\cdot,\beta_0)\right|^2\to 0$ as $\beta\to\beta_0$. Additionally, $\sup_{g(\cdot,\beta)\in\mathcal{G}_{\delta_n}}\left|\frac{g(\cdot,\beta)-g(\cdot,\beta_0)}{1+\sqrt{n}|\beta-\beta_0|}\right|\leqslant 1$ so Assumption 3 is satisfied. In the Appendix, we verify that Assumptions 2 and 3 are satisfied for examples 2 and 3.

Assumption 4 imposes that the constraints satisfy Linear Independence Constraint Qualification (LICQ), which says that the gradients of the active constraints are linearly independent. LICQ is the weakest possible constraint qualification that ensures the set of optimal Lagrange multipliers that satisfy the first order Karush-Kuhn-Tucker (KKT) conditions is a singleton (Wachsmuth (2013)). We note that LICQ will be violated when some active constraint gradients are linear combinations of other active constraint gradients. In particular, LICQ will be violated when some of the active constraint gradients are zero. Examples of when LICQ is violated appear in e.g. Kaido et al. (2021) and Nocedal and Wright (2006). It is fine to relax LICQ to Mangasarian-Fromovitz constraint qualification (MFCQ) as long as we impose the additional condition that there are unique optimal Lagrange multipliers. MFCQ is weaker than LICQ because it does not require the gradients of the equality constraints to be linearly independent. Both MFCQ and LICQ are clearly satisfied for our examples 1-2 because there can be at most one constraint active at β_0. For example 3, MFCQ and LICQ will be satisfied if the constraint gradients corresponding to the active constraints at β_0 (those j for which $\frac{1}{n}\sum_{i=1}^n P_{ij}(\beta_0)=\bar{b}_j$) are linearly independent.

Assumption 5 requires that the sample Lagrangian evaluated at the population Lagrange multipliers has a uniform local quadratic approximation in $\sqrt{n}$ neighborhoods of β_0. This assumption is similar to the stochastic differentiability assumption in Pollard (1985) and is needed to derive the asymptotic distribution of $\sqrt{n}\left(\hat{\beta}_n - \beta_0\right)$. The importance of using the Lagrangian instead of the objective function is that it allows for β_0 to not be a solution of the unconstrained population optimization problem; in other words, we allow for the possibility that $l\left(\beta_0\right) \neq 0$. Note that since the derivative of the population Lagrangian satisfies $\nabla \mathcal{L}\left(\beta_0, \lambda_0\right) \equiv l\left(\beta_0\right) + \sum_{j \in \mathcal{E} \cup \mathcal{I}} \lambda_{0j} F_{0j} = 0$ by the KKT conditions, Assumption 5 can also be written as follows: for any $\delta_n \to 0$, and $\mathcal{B}_{\delta_n} = \left\{h \in \mathbb{R}^d : \frac{\|h\|}{\sqrt{n}} \leqslant \delta_n\right\}$,

$$\sup_{h \in \mathcal{B}_{\delta_n}} \left| \frac{n\tilde{\mathcal{L}}_n\left(\beta_0 + \frac{h}{\sqrt{n}}\right) - n\tilde{\mathcal{L}}_n\left(\beta_0\right) - h'\sqrt{n}\left(\hat{l}_n\left(\beta_0\right) - l\left(\beta_0\right)\right) - \frac{1}{2}h'H_0h - \frac{1}{2}\sum_{j \in \mathcal{E} \cup \mathcal{I}} \lambda_{0j} h' G_{0j} h}{1 + \|h\|^2} \right| = o_P(1)$$

When $l\left(\beta_0\right) = 0$ and LICQ is satisfied, $\lambda_{0j} = 0$ for all $j \in \mathcal{E} \cup \mathcal{I}$. A more in-depth discussion of why $\lambda_{0j} = 0$ appears in Remark 1. Assumption 5 can then be rewritten as follows: for any $\delta_n \to 0$,

$$\sup_{h \in \mathcal{B}_{\delta_n}} \left| \frac{n\hat{Q}_n\left(\beta_0 + \frac{h}{\sqrt{n}}\right) - n\hat{Q}_n\left(\beta_0\right) - h'\sqrt{n}\hat{l}_n\left(\beta_0\right) - \frac{1}{2}h'H_0h}{1 + \|h\|^2} \right| = o_P(1)$$

Assumption 5 is satisfied in Example 1 because the objective $\hat{Q}_n\left(\beta\right) = \frac{1}{2n}\sum_{i=1}^n \left(y_i - \beta\right)^2$ is quadratic and the constraint $\beta \geqslant 0$ is linear. In particular, $n\tilde{\mathcal{L}}_n\left(\beta_0 + \frac{h}{\sqrt{n}}\right) - n\tilde{\mathcal{L}}_n\left(\beta_0\right) = n\hat{Q}_n\left(\beta_0 + \frac{h}{\sqrt{n}}\right) - n\hat{Q}_n\left(\beta_0\right) - \lambda_0\sqrt{n}h = h'\sqrt{n}\hat{l}_n\left(\beta_0\right) + \frac{1}{2}h'H_0h + \lambda_0\sqrt{n}h = h'\sqrt{n}\left(\hat{l}_n\left(\beta_0\right) - l\left(\beta_0\right)\right) + \frac{1}{2}h'H_0h$, where we have used the KKT condition $l\left(\beta_0\right) - \lambda_0 = 0$. To check that Assumption 5 is satisfied in Example 2, we can use Proposition 1 of Chernozhukov and Hong (2003) who show that the uniform local quadratic approximation of the objective in a neighborhood of β_0 follows from compactness of the parameter space, continuity of the population Jacobian and Hessian, and the moments $\{\pi\left(\cdot, \beta\right) : \|\beta - \beta_0\| \leqslant R\}$ being a Donsker class for any $R > 0$. The constraint $\beta \geqslant 0$ is linear and can be dealt with using the same KKT condition $l\left(\beta_0\right) - \lambda_0 = 0$ as in Example 1. For Example 3, we can use Lemma 2 of Chernozhukov and Hong (2003) which says that the uniform local quadratic approximation of the objective in a neighborhood of β_0 will hold when $\hat{Q}_n\left(\beta\right)$ is twice continuously differentiable with a second derivative matrix that is uniformly consistent for the population hessian

H_0 in a neighborhood of β_0. The constraints in this example are estimated, and we will require the second derivatives of the estimated sample constraints to be uniformly consistent for the second derivatives of the population constraints.

In the following theorem, we show that when the inequality constraints $f_j(\beta_0)$ for $j \in \mathcal{I}$ are not drifting towards zero and when there are no strongly active constraints, the proximal bootstrap is able to consistently replicate the non-standard asymptotic distribution of constrained estimators for which the standard bootstrap is inconsistent. We denote $\bar{\beta}_n$ as the initial $\sqrt{n}$-consistent estimator for β_0, $\bar{G}_{nj} \equiv \left.\frac{\partial^2 f_j(\beta)}{\partial\beta\partial\beta'}\right|_{\beta=\bar{\beta}_n}$ for all j, and $\left\{\bar{\lambda}_{nj} \text{ for } j \in \mathcal{E} \cup \mathcal{I}\right\}$ are a set of optimal Lagrange multipliers obtained from the optimization problem used to compute $\bar{\beta}_n$.

Theorem 1. *Suppose Assumptions 1 - 5 are satisfied in addition to the following:*

(i) $\nabla^2\mathcal{L}(\beta_0, \lambda_0) \equiv H_0 + \sum_{j\in\mathcal{E}\cup\mathcal{I}} \lambda_{0j} G_{0j}$ *is positive definite on* $M(\lambda_0) = \left\{h : F_{0j}' h = 0, j \in \mathcal{E}\right\}$.

(ii) $\bar{H}_n \xrightarrow{p} H_0$, $\max_{j\in\mathcal{E}\cup\mathcal{I}} \left|\bar{G}_{nj} - G_{0j}\right| \xrightarrow{p} 0$, *and* $\max_{j\in\mathcal{E}\cup\mathcal{I}} \left|\bar{\lambda}_{nj} - \lambda_{0j}\right| \xrightarrow{p} 0$.

(iii) $\mathcal{I}_+^*(\lambda_0) \equiv \{j \in \mathcal{I}^* : \lambda_{0j} > 0\} = \varnothing$, *where* $\mathcal{I}^* \equiv \{j \in \mathcal{I} : f_j(\beta_0) = 0\}$.

Suppose $f_j(\beta_0)$ for $j \in \mathcal{I}$ is fixed (not changing with the sample size n). For any sequence α_n such that $\alpha_n \to 0$ and $\sqrt{n}\alpha_n \to \infty$, let $\hat{\beta}_n^* \equiv \underset{\beta\in C^*}{\arg\min} \hat{A}_n^*(\beta)$, where $\hat{A}_n^*(\beta)$ and C^* are defined in equation 1. Then, $\sqrt{n}\left(\hat{\beta}_n - \beta_0\right) \rightsquigarrow \mathcal{J}$ and $\frac{\hat{\beta}_n^* - \hat{\beta}_n}{\alpha_n} \underset{\mathbb{W}}{\overset{\mathbb{P}}{\rightsquigarrow}} \mathcal{J}$, where

$$\mathcal{J} = \underset{h\in\Sigma}{\arg\min}\left\{h'W_0 + \frac{1}{2}h'H_0h + \frac{1}{2}\sum_{j\in\mathcal{E}} \lambda_{0j} h' G_{0j} h\right\}$$

$$\Sigma = \left\{h : F_{0j}'h = 0 \text{ for } j \in \mathcal{E}, F_{0j}'h \leqslant 0 \text{ for } j \in \mathcal{I}^*\right\}$$

$W_0 \sim N\left(0, P\left(g(\cdot,\beta_0) - Pg(\cdot,\beta_0)\right)\left(g(\cdot,\beta_0) - Pg(\cdot,\beta_0)\right)'\right)$.

In condition (i), we do not require the Lagrangian's hessian $\nabla^2\mathcal{L}(\beta_0, \lambda_0)$ to be positive definite on $\mathbb{R}^d$ because β_0 is typically a saddle-point of $\mathcal{L}(\beta_0, \lambda_0)$. Condition (ii) says that the sample analogs of the hessians and Lagrange multipliers are consistent for their population limits. Condition (iii) rules out the strongly active inequality constraints at β_0 because the proximal bootstrap cannot distinguish between strongly versus weakly active inequality constraints (weakly active inequality constraints are those $j \in \mathcal{I}^*$ such that $\lambda_{0j} = 0$). If both strong and weakly active inequality constraints are present, then the constraint set in $\mathcal{J}$ should be $\Sigma = \left\{h : F_{0j}'h = 0 \text{ for } j \in \mathcal{E} \cup \mathcal{I}_+^*(\lambda_0), F_{0j}'h \leqslant 0 \text{ for } j \in \mathcal{I}_0^*(\lambda_0)\right\}$. The rate conditions on α_n will ensure that the nonactive inequality constraints will not be included in Σ; however, among the active inequality constraints, the proximal bootstrap is not able to determine which of them have positive Lagrange multipliers and turn them into equality constraints inside Σ. We think that ruling out strongly active inequality constraints at β_0 is a plausible assumption because we are effectively ruling out misspecified inequality constraints. In Example 1, the constraint $\beta \geqslant 0$ will be misspecified at β_0 if $E[y_i] < 0$ and we are interested in conducting inference on $\beta_0 = \underset{\beta \geqslant 0}{\arg\min} E\left[(y_i - \beta)^2\right]$. The proximal bootstrap cannot handle this misspecified inequality constraint because λ_0 is positive. Notice that the proximal bootstrap can allow for all types of equality constraints, including misspecified ones, because all of them remain as equality constraints inside Σ.

Remark 1. If $l(\beta_0) = 0$, meaning that the population unconstrained minimum is the same as the constrained minimum, then $\mathcal{J}$ reduces down to

$$\mathcal{J} = \underset{h \in \Sigma}{\arg\min} \left\{h'W_0 + \frac{1}{2}h'H_0h\right\}$$
$$\Sigma = \left\{h : F_{0j}'h = 0 \text{ for } j \in \mathcal{E}, F_{0j}'h \leqslant 0 \text{ for } j \in \mathcal{I}^*\right\}$$

This is because by the KKT conditions, λ_{0j} satisfies $l(\beta_0) + \sum_{j \in \mathcal{E} \cup \mathcal{I}} \lambda_{0j} F_{0j} = 0$, so if $l(\beta_0) = 0$, then $\sum_{j \in \mathcal{E} \cup \mathcal{I}} \lambda_{0j} F_{0j} = 0$. By LICQ, the active constraint gradients F_{0j} for $j \in \mathcal{E} \cup \mathcal{I}^*$ are all nonzero, and furthermore, the optimal Lagrange multipliers for the nonactive inequality constraints $j \in \mathcal{I} \backslash \mathcal{I}^*$ are zero by the complementary slackness conditions $\lambda_{0j} f_j(\beta_0) = 0$ for all $j \in \mathcal{E} \cup \mathcal{I}$. Therefore, $\lambda_{0j} = 0$ for all $j \in \mathcal{E} \cup \mathcal{I}$ is a solution to $\sum_{j \in \mathcal{E} \cup \mathcal{I}} \lambda_{0j} F_{0j} = 0$. Since the set of Lagrange multipliers that satisfy the KKT conditions is a singleton under LICQ, $\lambda_{0j} = 0$ for all $j \in \mathcal{E} \cup \mathcal{I}$ are the unique optimal Lagrange multipliers, which implies $\sum_{j \in \mathcal{E} \cup \mathcal{I}} \lambda_{0j} h' G_{0j} h = 0$.

In this case, we can redefine the proximal bootstrap estimator as $\hat{\beta}_n^* \equiv \underset{\beta \in C^*}{\arg\min} \hat{Z}_n^*(\beta)$, where

$$\hat{Z}_n^*(\beta) \equiv \alpha_n \sqrt{n} \left(\hat{l}_n^*(\bar{\beta}_n) - \hat{l}_n(\bar{\beta}_n)\right)' (\beta - \bar{\beta}_n) + \frac{1}{2} \|\beta - \bar{\beta}_n\|_{\bar{H}_n}^2 \qquad (2)$$

From results in Shapiro (1988) and Shapiro (1989), when $l(\beta_0) = 0$ and LICQ is satisfied, the Tangent cone $T_C(\beta_0) \equiv \limsup_{\tau \downarrow 0} \frac{C-\beta_0}{\tau}$ is equal to $\Sigma = \left\{h : F_{0j}'h = 0 \text{ for } j \in \mathcal{E}, F_{0j}'h \leqslant 0 \text{ for } j \in \mathcal{I}^*\right\}$, so $\mathcal{J}$ can be written as $\arg\min_{h \in T_C(\beta_0)} \left\{h'W_0 + \frac{1}{2}h'H_0h\right\}$, which coincides with the asymptotic distribution given in Geyer (1994). Additionally, LICQ implies C is Chernoff Regular at β_0, and this cone K will be the Tangent cone $T_C(\beta_0)$. The constraint set C is Chernoff Regular at β_0 if C is well-approximated by a cone K at β_0, meaning that $\inf_{w \in K} \|(\beta - \beta_0) - w\| = o(\|\beta - \beta_0\|)$ for all $\beta \in C$, and $\inf_{\beta \in C} \|(\beta - \beta_0) - w\| = o(\|w\|)$ for all $w \in K$ (see Theorem 2.1 of Geyer (1994) for more details).

Remark 2. If there are only equality constraints, then the asymptotic distribution becomes $\mathcal{J} = \arg\min_{h \in \Sigma} \left\{h'W_0 + \frac{1}{2}h'\left(H_0 + \sum_{j \in \mathcal{E}} \lambda_{0j}G_{0j}\right)h\right\}$ for $\Sigma = \left\{h : F_{0j}'h = 0 \text{ for } j \in \mathcal{E}\right\}$. Using standard arguments in Amemiya (1985) Section 1.4.1 or Newey and McFadden (1994) Section 9.1, $\mathcal{J} = -B_0^{-1}\left(I - F_0\left(F_0'B_0^{-1}F_0\right)^{-1}F_0'B_0^{-1}\right)W_0$, where $B_0 = H_0 + \sum_{j \in \mathcal{E}} \lambda_{0j}G_{0j}$. If W_0 is multivariate normal, then the asymptotic distribution will be multivariate normal.

If $l(\beta_0) = 0$ or if the constraints are linear, then $\sum_{j \in \mathcal{E}} \lambda_{0j}G_{0j} = 0$ and $B_0 = H_0$, so $\mathcal{J} = -H_0^{-1}\left(I - F_0\left(F_0'H_0^{-1}F_0\right)^{-1}F_0'H_0^{-1}\right)W_0$.

2.2 Uniformity

In the case of drifting inequality constraints $f_j(\beta_0) = c/n^\rho$ for some $\rho > 0$ and $c < 0$, the proximal bootstrap will typically not consistently replicate the estimator's asymptotic distribution; however we can still obtain a uniformly asymptotically valid confidence set for β_0 using test-inversion. Throughout this section, we will assume the constraints are not necessary for identification of β_0, meaning $l(\beta_0) = 0$. We will benchmark the distribution of the test statistic $n\left(\hat{Q}_n(\beta_0) - \inf_{h \in \mathcal{B}_{\delta_n}} \hat{Q}_n\left(\beta_0 + \frac{h}{\sqrt{n}}\right)\right)$ against the empirical distribution of $-\frac{\inf_{\beta \in \mathbb{B}} \hat{Z}_n^*(\beta)}{\alpha_n^2}$, where $\hat{Z}_n^*(\beta) = \alpha_n\sqrt{n}\left(\hat{l}_n^*(\bar{\beta}_n) - \hat{l}_n(\bar{\beta}_n)\right)'(\beta - \bar{\beta}_n) + \frac{1}{2}\|\beta - \bar{\beta}_n\|_{\bar{H}_n}^2$, $\mathcal{B}_{\delta_n} = \left\{h \in \mathbb{R}^d : \frac{\|h\|}{\sqrt{n}} \leqslant \delta_n\right\}$ is a shrinking neighborhood, and $\delta_n \to 0$ satisfies $\sqrt{n}\delta_n \to \kappa$ for $\kappa \in (0, \infty]$. Let $\hat{c}_{1-\alpha}^*$ be the $1 - \alpha$ quantile of $-\frac{\inf_{\beta \in \mathbb{B}} \hat{Z}_n^*(\beta)}{\alpha_n^2}$. We will show that $\mathcal{C}_{1-\alpha}^* = \left\{\beta : n\left(\hat{Q}_n(\beta) - \inf_{h \in \mathcal{B}_{\delta_n}} \hat{Q}_n\left(\beta + \frac{h}{\sqrt{n}}\right)\right) \leqslant \hat{c}_{1-\alpha}^*\right\}$ is a uniformly asymptotically valid nominal $1 - \alpha$ confidence set for $\beta_0 \equiv \beta(P)$.

In the theorem below, $J_n(\cdot, P)$ denotes the CDF of $n\left(\hat{Q}_n(\beta_0) - \inf_{h \in \mathcal{B}_{\delta_n}} \hat{Q}_n\left(\beta_0 + \frac{h}{\sqrt{n}}\right)\right)$ under P, and $J(\cdot, P)$ denotes the CDF of its limiting distribution under P. Similarly, $J_{\alpha_n}^*(\cdot, P)$ denotes the conditional CDF of $-\frac{\inf_{\beta\in\mathbb{B}}\hat{Z}_n^*(\beta)}{\alpha_n^2}$ under P, and $J^*(\cdot,P)$ denotes the CDF of its limiting distribution under P.

Theorem 2. *Let $\mathcal{P}$ be a class of distributions for which $l(\beta_0)=0$, Assumptions 1 - 5 hold uniformly in $P\in\mathcal{P}$ and condition (ii) of Theorem 1 is satisfied uniformly in $P\in\mathcal{P}$, and $\{J(\cdot,P):P\in\mathcal{P}\}$ and $\{J^*(\cdot,P):P\in\mathcal{P}\}$ are equicontinuous at $J_n^{-1}(1-\alpha,P)$. Then $\liminf_{n\to\infty}\inf_{P\in\mathcal{P}}P\left(\beta_0\in\mathcal{C}^*_{1-\alpha}\right)\geqslant 1-\alpha$, where $\mathcal{C}^*_{1-\alpha}=\left\{\beta:n\left(\hat{Q}_n(\beta)-\inf_{h\in\mathcal{B}_{\delta_n}}\hat{Q}_n\left(\beta+\frac{h}{\sqrt{n}}\right)\right)\leqslant\hat{c}^*_{1-\alpha}\right\}$, $\mathcal{B}_{\delta_n}=\left\{h\in\mathbb{R}^d:\frac{\|h\|}{\sqrt{n}}\leqslant\delta_n\right\}$, $\delta_n\to 0$ satisfies $\sqrt{n}\delta_n\to\kappa$ for $\kappa\in(0,\infty]$, and $\hat{c}^*_{1-\alpha}$ is the $1-\alpha$ quantile of $-\frac{\inf_{\beta\in\mathbb{B}}\hat{Z}_n^*(\beta)}{\alpha_n^2}$ for any α_n satisfying $\alpha_n\to 0$ and $\sqrt{n}\alpha_n\to\infty$.*

Remark 3. If we would like to construct a nominal $1-\alpha$ confidence set for a subvector $\gamma_0=a'\beta_0$, where a is a known vector, we could use projection: $CI^{Proj}_{1-\alpha}=\left[\inf_{\beta\in\mathcal{C}^*_{1-\alpha}}a'\beta,\ \sup_{\beta\in\mathcal{C}^*_{1-\alpha}}a'\beta\right]$. The uniform asymptotic validity of these projection intervals follows directly from the uniform asymptotic validity of $\mathcal{C}^*_{1-\alpha}$.

Andrews (2000) Example Revisited Suppose in the Andrews (2000) example the parameter is drifting at some τ_n rate: $\beta_0=c/\tau_n$ for some constant $c>0$. When $c>0$, $l(\beta_0)=0$ and the inequality constraint $\beta>0$ is weakly active in the limit as $\tau_n\to\infty$. To conduct uniformly valid inference, we can use $\mathcal{C}^*_{1-\alpha}=\left\{\beta:n\left(\hat{Q}_n(\beta)-\inf_{h\in\mathcal{B}_{\delta_n}}\hat{Q}_n\left(\beta+\frac{h}{\sqrt{n}}\right)\right)\leqslant\hat{c}^*_{1-\alpha}\right\}$, where $\hat{c}^*_{1-\alpha}$ is the $1-\alpha$ quantile of $-\frac{\inf_{\beta\in\mathbb{B}}\hat{Z}_n^*(\beta)}{\alpha_n^2}$, and $\hat{Z}_n^*(\beta)=\alpha_n\sqrt{n}\left(\bar{y}_n-\bar{y}_n^*\right)\left(\beta-\bar{\beta}_n\right)+\frac{1}{2}\left(\beta-\bar{\beta}_n\right)^2$. We can show that $n\left(\hat{Q}_n(\beta_0)-\inf_{h\in\mathcal{B}_{\delta_n}}\hat{Q}_n\left(\beta_0+\frac{h}{\sqrt{n}}\right)\right)\rightsquigarrow-\inf_{h\in\mathcal{B}_\kappa}q(h)$, where $\mathcal{B}_\kappa=\left\{h\in\mathbb{R}^d:\|h\|\leqslant\kappa\right\}$ for $\sqrt{n}\delta_n\to\kappa\in(0,\infty]$, and $-\frac{\inf_{\beta\in\mathbb{B}}\hat{Z}_n^*(\beta)}{\alpha_n^2}\overset{\mathbb{P}}{\underset{\mathbb{W}}{\rightsquigarrow}}-\min_{h\in\mathbb{R}^d}q(h)$, where $q(h)=h'W_0+\frac{1}{2}h'H_0h$, $H_0=1$ and $W_0\sim N(0,Var(y_i))$.

2.3 Choice of α_n

In order to determine the optimal value of α_n, one possibility is to use a double bootstrap algorithm similar to the one in Chakraborty et al. (2013). Starting from the smallest value in a grid of α_n, draw B_1 bootstrap samples and compute initial $\sqrt{n}$-consistent estimates $\bar{\beta}_n^{(b_1)}$ for $b_1=1,\ldots,B_1$. To obtain these initial $\sqrt{n}$-consistent estimates, we could use the proximal bootstrap or other consistent procedures such as subsampling, but we cannot use the standard bootstrap which can be inconsistent when parameters are not in the interior. We can use these $\bar{\beta}^{(b_1)}_n$ to estimate the Jacobians $\hat{l}^{(b_1)}_n\left(\bar{\beta}^{(b_1)}_n\right)$ and Hessians $\bar{H}^{(b_1)}_n$ and $\bar{G}^{(b_1)}_{nj}$ and all $j \in \mathcal{E} \cup \mathcal{I}$. Conditional on each of these bootstrap samples $b_1 = 1, \ldots, B_1$, draw B_2 bootstrap samples and compute $\hat{l}^{(b_2)}_n\left(\bar{\beta}^{(b_1)}_n\right) - \hat{l}^{(b_1)}_n\left(\bar{\beta}^{(b_1)}_n\right)$ for $b_2 = 1, \ldots, B_2$. Pick some nominal frequency $1-\tau$. Compute the empirical frequency with which $\hat{\mathcal{C}}^{(b_1)}_{1-\alpha} = \left\{\beta : n\left(\hat{Q}^{(b_1)}_n(\beta) - \inf_{h\in\mathcal{B}_{\delta_n}} \hat{Q}^{(b_1)}_n\left(\beta + \frac{h}{\sqrt{n}}\right)\right) \leqslant \hat{c}^*_{1-\tau}\right\}$ covers $\hat{\beta}_n$, where $\hat{c}^*_{1-\tau}$ is the $1-\tau$ quantile of $-\frac{\inf_{\beta\in\mathbb{B}} \hat{A}^{(b_1,b_2)}_n(\beta)}{\alpha_n^2}$ and

$$\hat{Z}^{(b_1,b_2)}_n(\beta) \equiv \alpha_n\sqrt{n}\left(\hat{l}^{(b_2)}_n\left(\bar{\beta}^{(b_1)}_n\right) - \hat{l}^{(b_1)}_n\left(\bar{\beta}^{(b_1)}_n\right)\right)'\left(\beta - \bar{\beta}^{(b_1)}_n\right) + \frac{1}{2}\left\|\beta - \bar{\beta}^{(b_1)}_n\right\|^2_{\bar{H}^{(b_1)}_n} \tag{3}$$

If the current value of α_n achieves coverage at or above $1-\tau$, then it picks that value as the optimal α_n. Otherwise, increment α_n to the next highest value in the grid and repeat the steps above.

The justification for why this procedure works is similar to the arguments in Hall and Martin (1988) for using bootstrap iteration to reduce coverage error for confidence intervals. We are trying to estimate the coverage frequency of $\mathcal{C}^*_{1-\alpha}$ by constructing confidence sets $\hat{\mathcal{C}}^{(b_1)}_{1-\alpha}$ using the resampled data. We need B_1 and B_2 to be large enough so that we can estimate the coverage frequency well enough.

2.4 Estimated Constraints

We can also apply the proximal bootstrap to constrained estimators with a finite number of $\sqrt{n}$-consistently estimated inequality and/or equality constraints that are twice continuously differentiable over a compact parameter space $\mathbb{B} \subset \mathbb{R}^d$.

$$\hat{\beta}_n = \arg\min_{\beta\in C} \hat{Q}_n(\beta), \quad C = \{\beta \in \mathbb{B} : f_{nj}(\beta) = 0 \text{ for } j \in \mathcal{E}, f_{nj}(\beta) \leqslant 0 \text{ for } j \in \mathcal{I}\}$$

We will define the population analog of $C \subseteq \mathbb{B}$ to be $C_0 \equiv \{\beta \in \mathbb{B} : f_{0j}(\beta) = 0 \text{ for } j \in \mathcal{E}, f_{0j}(\beta) \leqslant 0 \text{ for } j \in \mathcal{I}\}$, where $\sup_{\beta\in\mathbb{B}} |f_{nj}(\beta) - f_{0j}(\beta)| = o_P(1)$ for all $j \in \mathcal{E} \cup \mathcal{I}$. We are interested in conducting inference on $\beta_0 \equiv \arg\min_{\beta\in C_0} Q(\beta)$, which is assumed to be unique. $Q(\beta)$ is twice continuously differentiable at β_0 and $\sup_{\beta\in\mathbb{B}} \left|\hat{Q}_n(\beta) - Q(\beta)\right| = o_P(1)$.

Let $f_{nj}^*(\beta)$ be the bootstrap analog of $f_{nj}(\beta)$ and let $F_{nj}^*(\beta) \equiv \frac{\partial f_{nj}^*(\beta)}{\partial \beta}$. For any $\bar{\beta}_n$ such that $\sqrt{n}\left(\bar{\beta}_n - \beta_0\right) = O_p(1)$, let $\bar{F}_{nj} \equiv F_{nj}\left(\bar{\beta}_n\right)$, $\bar{F}_{nj}^* \equiv F_{nj}^*\left(\bar{\beta}_n\right)$, $\bar{G}_{nj} \equiv \left.\frac{\partial^2 f_{nj}(\beta)}{\partial \beta \partial \beta'}\right|_{\beta=\bar{\beta}_n}$ for all j, and let $\bar{\lambda}_{nj}$ be a set of optimal Lagrange multipliers for $\bar{\beta}_n$. These Lagrange multipliers can be obtained directly as outputs from the optimization algorithm's function call for computing $\bar{\beta}_n$. We modify our proximal bootstrap to account for the sampling variation in the constraint Jacobians:

$$\begin{aligned}
\hat{A}_n^*(\beta) &\equiv \alpha_n \sqrt{n}\left(\hat{l}_n^*\left(\bar{\beta}_n\right) - \hat{l}_n\left(\bar{\beta}_n\right)\right)'\left(\beta - \bar{\beta}_n\right) + \frac{1}{2}\left\|\beta - \bar{\beta}_n\right\|_{\bar{H}_n}^2 \\
&\quad + \sum_{j \in \mathcal{E} \cup \mathcal{I}} \bar{\lambda}_{nj}\left(\alpha_n \sqrt{n}\left(\bar{F}_{nj}^* - \bar{F}_{nj}\right)'\left(\beta - \bar{\beta}_n\right) + \frac{1}{2}\left\|\beta - \bar{\beta}_n\right\|_{\bar{G}_{nj}}^2\right) \\
C^* &\equiv \left\{\beta \in \mathbb{B} : f_{nj}\left(\bar{\beta}_n\right) + \bar{F}_{nj}'\left(\beta - \bar{\beta}_n\right) + \alpha_n \sqrt{n}\left(f_{nj}^*\left(\bar{\beta}_n\right) - f_{nj}\left(\bar{\beta}_n\right)\right) = 0 \text{ for } j \in \mathcal{E},\right. \\
&\qquad \left. f_{nj}\left(\bar{\beta}_n\right) + \bar{F}_{nj}'\left(\beta - \bar{\beta}_n\right) + \alpha_n \sqrt{n}\left(f_{nj}^*\left(\bar{\beta}_n\right) - f_{nj}\left(\bar{\beta}_n\right)\right) \leqslant 0 \text{ for } j \in \mathcal{I}\right\}
\end{aligned} \tag{4}$$

We will modify Assumption 1 to account for the difference between the sample versus the population constraints.

Assumption 1′. (i) $\mathbb{B} \subset \mathbb{R}^d$ is compact, $C \subseteq \mathbb{B}$ is closed, and d is fixed.

(ii) $\hat{\beta}_n$ satisfies $\hat{Q}_n\left(\hat{\beta}_n\right) \leqslant \inf_{\beta \in C} \hat{Q}_n(\beta) + o_p(1)$.

(iii) β_0 is the unique value of $\arg\min_{\beta \in C_0} Q(\beta)$.

(iv) $Q(\beta)$ is twice continuously differentiable at β_0, and $\sup_{\beta \in \mathbb{B}}\left|\hat{Q}_n(\beta) - Q(\beta)\right| = o_P(1)$.

(v) $f_{nj} : \mathbb{B} \mapsto \mathbb{R}$ and $f_{0j} : \mathbb{B} \mapsto \mathbb{R}$ are twice continuously differentiable functions that satisfy $\sup_{\beta \in \mathbb{B}}\left|f_{nj}(\beta) - f_{0j}(\beta)\right| = o_P(1)$ for all $j \in \mathcal{E} \cup \mathcal{I}$.

We also modify Assumption 5 to account for estimated constraints.

Assumption 5′. Let λ_{0j} be the unique Lagrange multipliers that satisfy $\lambda_{0j} f_{0j}(\beta_0) = 0$ for all $j \in \mathcal{E} \cup \mathcal{I}$, $0 \leqslant \lambda_{0j} < \infty$ for all $j \in \mathcal{E} \cup \mathcal{I}$, and $\nabla \mathcal{L}(\beta_0, \lambda_0) \equiv l(\beta_0) + \sum_{j \in \mathcal{E} \cup \mathcal{I}} \lambda_{0j} F_{0j} = 0$. Define $\tilde{\mathcal{L}}_n(\beta) \equiv \hat{Q}_n(\beta) + \sum_{j \in \mathcal{E} \cup \mathcal{I}} \lambda_{0j} f_{nj}(\beta)$, $F_{nj}(\beta_0) \equiv \left.\frac{\partial f_{nj}(\beta)}{\partial \beta}\right|_{\beta=\beta_0}$, and $G_{0j} \equiv \left.\frac{\partial^2 f_{0j}(\beta)}{\partial \beta \partial \beta'}\right|_{\beta=\beta_0}$. For any

$\delta_n \to 0$, and $\mathcal{B}_{\delta_n} = \left\{h \in \mathbb{R}^d : \frac{\|h\|}{\sqrt{n}} \leqslant \delta_n\right\}$,

$$\sup_{h \in \mathcal{B}_{\delta_n}} \left| \frac{n\tilde{\mathcal{L}}_n\left(\beta_0 + \frac{h}{\sqrt{n}}\right) - n\tilde{\mathcal{L}}_n\left(\beta_0\right) - h'\sqrt{n}\hat{l}_n\left(\beta_0\right) - \frac{1}{2}h'H_0h - \sum_{j\in\mathcal{E}\cup\mathcal{I}} \lambda_{0j}\left(\sqrt{n}F_{nj}\left(\beta_0\right)' h + \frac{1}{2}h'G_{0j}h\right)}{1 + \|h\|^2} \right|$$
$$= o_P(1)$$

Note that since $\nabla\mathcal{L}\left(\beta_0, \lambda_0\right) \equiv l\left(\beta_0\right) + \sum_{j\in\mathcal{E}\cup\mathcal{I}} \lambda_{0j}F_{0j} = 0$, Assumption $5'$ can also be written as follows: for any $\delta_n \to 0$, and $\mathcal{B}_{\delta_n} = \left\{h \in \mathbb{R}^d : \frac{\|h\|}{\sqrt{n}} \leqslant \delta_n\right\}$,

$$\sup_{h \in \mathcal{B}_{\delta_n}} \left| \frac{n\tilde{\mathcal{L}}_n\left(\beta_0 + \frac{h}{\sqrt{n}}\right) - n\tilde{\mathcal{L}}_n\left(\beta_0\right) - h'\sqrt{n}\left(\hat{l}_n\left(\beta_0\right) - l\left(\beta_0\right)\right) - \frac{1}{2}h'H_0h - \sum_{j\in\mathcal{E}\cup\mathcal{I}} \lambda_{0j}\left(\sqrt{n}\left(F_{nj}\left(\beta_0\right) - F_{0j}\right)' h + \frac{1}{2}h'G_{0j}h\right)}{1 + \|h\|^2} \right| = o_P(1)$$

Assumption $5'$ will hold in Example 3 if uniform local quadratic expansions exist for $\hat{Q}_n\left(\beta\right)$ and the constraints $f_{nj}\left(\beta\right) = \frac{1}{n}\sum_{i=1}^n P_{ij}\left(\beta\right) - \bar{b}_j$. Since both $\hat{Q}_n\left(\beta\right)$ and $f_{nj}\left(\beta\right)$ are twice continuously differentiable, the uniform local quadratic expansions will hold if the second derivative matrices of $\hat{Q}_n\left(\beta\right)$ and $f_{nj}\left(\beta\right)$ are uniformly consistent for the population hessians H_0 and G_{0j} when β lies in a neighborhood of β_0.

We next impose that the bootstrapped constraint Jacobians converge weakly in probability to the same limiting distribution as the unbootstrapped constraint Jacobians.

Assumption 6. *(i)* $\sqrt{n}\left(\hat{l}_n\left(\beta_0\right) - l\left(\beta_0\right)\right) + \sum_{j\in\mathcal{E}\cup\mathcal{I}} \lambda_{0j}\sqrt{n}\left(F_{nj}\left(\beta_0\right) - F_{0j}\right) \rightsquigarrow W_0 + \sum_{j\in\mathcal{E}\cup\mathcal{I}} \lambda_{0j}V_{0j}$, *a tight random vector.*

(ii) $\sqrt{n}\left(\hat{l}^*_n\left(\beta_0\right) - \hat{l}_n\left(\beta_0\right)\right) + \sum_{j\in\mathcal{E}\cup\mathcal{I}} \lambda_{0j}\sqrt{n}\left(F^*_{nj}\left(\beta_0\right) - F_{nj}\left(\beta_0\right)\right) \underset{\mathbb{W}}{\overset{\mathbb{P}}{\rightsquigarrow}} W_0 + \sum_{j\in\mathcal{E}\cup\mathcal{I}} \lambda_{0j}V_{0j}$.

(iii) $\sup_{\|\beta-\beta_0\| \leqslant o(1)} \sqrt{n}\left(F^*_n\left(\beta\right) - F_n\left(\beta\right) - F^*_n\left(\beta_0\right) + F_n\left(\beta_0\right)\right) = o^*_P(1)$.

A sufficient condition for $\sqrt{n}\left(\hat{l}_n\left(\beta_0\right) - l\left(\beta_0\right)\right) + \sum_{j\in\mathcal{E}\cup\mathcal{I}} \lambda_{0j}\sqrt{n}\left(F_{nj}\left(\beta_0\right) - F_{0j}\right) \rightsquigarrow W_0 + \sum_{j\in\mathcal{E}\cup\mathcal{I}} \lambda_{0j}V_{0j}$ is $\begin{pmatrix} \sqrt{n}\left(\hat{l}_n\left(\beta_0\right) - l\left(\beta_0\right)\right) \\ \sqrt{n}\left(F_n\left(\beta_0\right) - F_0\right) \end{pmatrix} \rightsquigarrow \begin{pmatrix} W_0 \\ V_0 \end{pmatrix}$, where $F_0 = \left(F_{0j} \text{ for } j \in \mathcal{E} \cup \mathcal{I}\right)$ and $V_0 = \left(V_{0j} \text{ for } j \in \mathcal{E} \cup \mathcal{I}\right)$. Similarly, a sufficient condition for $\sqrt{n}\left(\hat{l}^*_n\left(\beta_0\right) - \hat{l}_n\left(\beta_0\right)\right) + \sum_{j\in\mathcal{E}\cup\mathcal{I}} \lambda_{0j}\sqrt{n}\left(F^*_{nj}\left(\beta_0\right) - F_{nj}\left(\beta_0\right)\right) \underset{\mathbb{W}}{\overset{\mathbb{P}}{\rightsquigarrow}} W_0 + \sum_{j\in\mathcal{E}\cup\mathcal{I}} \lambda_{0j}V_{0j}$ is $\begin{pmatrix} \sqrt{n}\left(\hat{l}^*_n\left(\beta_0\right) - \hat{l}_n\left(\beta_0\right)\right) \\ \sqrt{n}\left(F^*_n\left(\beta_0\right) - F_n\left(\beta_0\right)\right) \end{pmatrix} \underset{\mathbb{W}}{\overset{\mathbb{P}}{\rightsquigarrow}} \begin{pmatrix} W_0 \\ V_0 \end{pmatrix}$. When $F_n\left(\beta\right) = P_n\pi(\cdot, \beta)$ and $F^*_n\left(\beta\right) =$

$P_n^* \pi(\cdot, \beta)$ are sample averages, these joint weak convergence statements can be verified under a joint Lindeberg condition.

In the next theorem, we show that when the population inequality constraints $f_{0j}(\beta_0)$ for $j \in \mathcal{I}$ are not drifting towards zero, the proximal bootstrap is able to consistently replicate the non-standard asymptotic distribution of constrained estimators for which the standard bootstrap is inconsistent.

Theorem 3. *Suppose Assumptions 1′, 2 - 4, 5′, and 6 are satisfied in addition to the following:*

(i) $\nabla^2 \mathcal{L}(\beta_0, \lambda_0) \equiv H_0 + \sum_{j \in \mathcal{E} \cup \mathcal{I}} \lambda_{0j} G_{0j}$ *is positive definite on* $M(\lambda_0) = \left\{ h : F_{0j}' h = 0, j \in \mathcal{E} \right\}$.

(ii) $\bar{H}_n \xrightarrow{p} H_0$, $\max\limits_{j \in \mathcal{E} \cup \mathcal{I}} \left| \bar{G}_{nj} - G_{0j} \right| \xrightarrow{p} 0$, *and* $\max\limits_{j \in \mathcal{E} \cup \mathcal{I}} \left| \bar{\lambda}_{nj} - \lambda_{0j} \right| \xrightarrow{p} 0$.

(iii) $\mathcal{I}_{n,+}^*(\lambda_0) \equiv \{ j \in \mathcal{I}_n^* : \lambda_{0j} > 0 \} = \varnothing$, *where* $\mathcal{I}_n^* \equiv \{ j \in \mathcal{I} : f_{nj}(\beta_0) = 0 \}$, *and*

$\mathcal{I}_+^*(\lambda_0) \equiv \{ j \in \mathcal{I}^* : \lambda_{0j} > 0 \} = \varnothing$, *where* $\mathcal{I}^* \equiv \{ j \in \mathcal{I} : f_{0j}(\beta_0) = 0 \}$.

Suppose $f_{0j}(\beta_0)$ *for* $j \in \mathcal{I}$ *is fixed (not changing with* n *). Then, for any sequence* α_n *such that* $\alpha_n \to 0$ *and* $\sqrt{n}\alpha_n \to \infty$, $\sqrt{n}\left(\hat{\beta}_n - \beta_0\right) \rightsquigarrow \mathcal{J}$ *and* $\frac{\hat{\beta}_n^* - \hat{\beta}_n}{\alpha_n} \underset{\mathbb{W}}{\overset{\mathbb{P}}{\rightsquigarrow}} \mathcal{J}$,

$$\mathcal{J} = \underset{h \in \Sigma}{\arg\min} \left\{ h' W_0 + \frac{1}{2} h' H_0 h + \sum_{j \in \mathcal{E}} \lambda_{0j} \left(h' V_{0j} + \frac{1}{2} h' G_{0j} h \right) \right\}$$

$$\Sigma = \left\{ h : U_{0j} + F_{0j}' h = 0 \text{ for } j \in \mathcal{E}, U_{0j} + F_{0j}' h \leqslant 0 \text{ for } j \in \mathcal{I}^* \right\}$$

Note that uniformity results in Theorem 2 still hold in the case of estimated constraints, as long as $l(\beta_0) = 0$.

3 Monte Carlo Simulations

3.1 Two-sided Boundary Constraint

We consider a simple location model with i.i.d data:

$$y_i = \beta_0 + \epsilon_i \ , \epsilon_i \overset{i.i.d.}{\sim} N(0, 1)$$

We would like to compute the maximum likelihood estimator subject to the constraint that the parameter lies between 0 and $\bar{x}_n = \frac{1}{n}\sum_{i=1}^{n} x_i$, where $x_i \overset{i.i.d.}{\sim} N(5,1)$ and $x_i \perp y_i$.

$$\hat{\beta}_n = \underset{0 \leqslant \beta \leqslant \bar{x}_n}{\arg\min} \frac{1}{2n}\sum_{i=1}^{n}(y_i - \beta)^2$$

Note that we can express our estimator as a function of $\bar{y}_n$, treating $\bar{x}_n$ as given.

$$\hat{\beta}_n = \max\left(\min\left(\bar{y}_n, \bar{x}_n\right), 0\right) \equiv \phi\left(\bar{y}_n\right)$$

We will examine the empirical coverage and average interval length of the proximal bootstrap confidence set $\mathcal{C}^*_{1-\alpha} = \left\{\beta : n\left(\hat{Q}_n(\beta) - \inf_{h \in \mathcal{B}_{\delta_n}} \hat{Q}_n\left(\beta + \frac{h}{\sqrt{n}}\right)\right) \leqslant \hat{c}^*_{1-\alpha}\right\}$, where $\hat{c}^*_{1-\alpha}$ is the $1-\alpha$ quantile of $-\frac{\inf_{\beta \in \mathbb{B}} \hat{Z}^*_n(\beta)}{\alpha_n^2}$ and $\hat{Z}^*_n(\beta) = \alpha_n\sqrt{n}\left(\hat{l}^*_n\left(\bar{\beta}_n\right) - \hat{l}_n\left(\bar{\beta}_n\right)\right)'\left(\beta - \bar{\beta}_n\right) + \frac{1}{2}\left\|\beta - \bar{\beta}_n\right\|^2_{\bar{H}_n}$, for $\bar{\beta}_n = \hat{\beta}_n$, $\hat{l}^*_n\left(\bar{\beta}_n\right) - \hat{l}_n\left(\bar{\beta}_n\right) = \bar{y}_n - \bar{y}^*_n$ and $\bar{H}_n = 1$. The true parameter β_0 takes on 7 different values: $\beta_0 \in \left\{1, n^{-1/6}, n^{-1/4}, n^{-1/3}, n^{-1/2}, n^{-1}, 0\right\}$. We consider four different sample sizes $n \in \{100, 500, 1000, 5000\}$ and use 1000 bootstrap iterations and 2000 Monte Carlo simulations. We chose $\alpha_n = n^{-1/4}$ after performing the double bootstrap procedure described in Section 2.3 using $n = 5000$, $B_1 = B_2 = 5000$, and $\beta_0 = 0$. The empirical coverage frequencies over a grid of $\alpha_n \in \left\{n^{-1/3}, n^{-1/4}, n^{-1/6}, n^{-1/7}, n^{-1/8}, n^{-1/9}, n^{-1/10}\right\}$ were $\{0.9496, 0.9538, 0.9460, 0.9536, 0.9508, 0.9526, 0.9496\}$. $\alpha_n = n^{-1/4}$ was the smallest value which achieved coverage at or above the nominal level of 0.95. We also tried using all the other values of α_n and found that the coverage was the same up to three decimal places across the different values of α_n. We did not constrain h when computing $\inf_{h \in \mathcal{B}_{\delta_n}} \hat{Q}_n\left(\beta + \frac{h}{\sqrt{n}}\right)$, which effectively sets δ_n to $\sqrt{n}\delta_n \to \infty$.

We will compare the empirical coverage frequency of the proximal bootstrap to alternative methods. Fang and Santos (2019)'s equal-tailed two-sided interval is $\left[\phi\left(\bar{y}\right) - \frac{1}{\sqrt{n}}\hat{c}_{1-\alpha/2}, \phi\left(\bar{y}_n\right) - \frac{1}{\sqrt{n}}\hat{c}_{\alpha/2}\right]$, where $\hat{c}_\alpha$ is the αth quantile of $\hat{\phi}'\left(\sqrt{n}\left(\bar{y}^*_n - \bar{y}_n\right)\right)$, $\bar{y}^*_n$ is the nonparametric bootstrap analog of $\bar{y}_n$, and

$$\hat{\phi}'(h) = \begin{cases} h & \text{if } \kappa_n < \sqrt{n}\bar{y}_n/\hat{\sigma} \text{ and } \sqrt{n}\left(\bar{y}_n - \bar{x}_n\right)/\hat{\sigma} < -\kappa_n \\ \max\left(h, 0\right) & \text{if } \left|\sqrt{n}\bar{y}_n/\hat{\sigma}\right| \leqslant \kappa_n \\ \min\left(h, \bar{x}_n\right) & \text{if } \left|\sqrt{n}\left(\bar{y}_n - \bar{x}_n\right)/\hat{\sigma}\right| \leqslant \kappa_n \\ 0 & \text{if } \sqrt{n}\bar{y}_n/\hat{\sigma} < -\kappa_n \text{ or } \sqrt{n}\left(\bar{y}_n - \bar{x}_n\right)/\hat{\sigma} > \kappa_n \end{cases}$$

We use Fang and Santos (2019)'s recommended choice of $\kappa_n = \Phi^{-1}\left(1 - \delta_n\right)$ for some $\delta_n \downarrow 0$ (see their Example 2.1 on page 391-392). We tried three different types of confidence intervals (two-sided equal-tailed, one-sided lower, and one-sided upper) and four different values of $\delta_n \in \left\{n^{-1}, n^{-1/2}, n^{-1/3}, n^{-1/6}\right\}$, and none of them produced uniformly valid coverage for all drifting parameter sequences.

Hsieh et al. (2022) propose using

$$CS_n^{PD}\left(1-\alpha\right) = \left\{\beta : \min_{\lambda, s \geqslant 0, \lambda_1 \beta = 0, \lambda_2 s = 0} n g\left(y, \beta, \lambda_1, \lambda_2, s\right)' \left(G\hat{V}G'\right)^{-1} g\left(y, \beta, \lambda_1, \lambda_2, s\right) \leqslant \chi_2^2\left(1-\alpha\right)\right\}$$

$$g\left(y, \beta, \lambda_1, \lambda_2\right) = \begin{bmatrix} -\frac{1}{n}\sum_{i=1}^n \left(y_i - \beta\right) - \lambda_1 + \lambda_2 \\ \bar{x}_n - \beta - s \end{bmatrix} = \begin{bmatrix} -1 & 0 \\ 0 & 1 \end{bmatrix} \begin{bmatrix} \bar{y}_n \\ \bar{x}_n \end{bmatrix} + \begin{bmatrix} \beta - \lambda_1 + \lambda_2 \\ -\beta - s \end{bmatrix}$$

$$G = \begin{bmatrix} -1 & 0 \\ 0 & 1 \end{bmatrix}$$

$$\hat{V} = \begin{bmatrix} \widehat{Var}\left(y\right) & 0 \\ 0 & \widehat{Var}\left(x\right) \end{bmatrix} = \begin{bmatrix} \frac{1}{n-1}\sum_{i=1}^n \left(y_i - \bar{y}_n\right)^2 & 0 \\ 0 & \frac{1}{n-1}\sum_{i=1}^n \left(x_i - \bar{x}_n\right)^2 \end{bmatrix}$$

Table 1 compares the empirical coverage frequencies and average interval lengths (in parentheses) of the proximal bootstrap simultaneous confidence set to Hsieh et al. (2022)'s confidence set, Fang and Santos (2019)'s equal-tailed two-sided intervals, subsampling (using $\lfloor\sqrt{n}\rfloor$ as the subsample size) and standard nonparametric bootstrap two-sided equal-tailed confidence intervals. For n large enough, the proximal bootstrap coverage frequencies are close to 95% for all drifting parameters. Hsieh et al. (2022)'s coverage is more conservative than the proximal bootstrap for all parameters, and the average interval lengths for Hsieh et al. (2022) are longer. The coverage of Fang and Santos

(2019), subsampling, and the standard nonparametric bootstrap can be far below 95%, especially for $\beta_0 = 1/n$ where the coverage drops to around 50%.

3.2 Boundary Constrained Nonsmooth GMM

We consider a simple location model with i.i.d data:

$$y_i = \beta_0 + \epsilon_i, \quad \epsilon_i \sim N(0,1)$$

For $\pi(\cdot,\beta) = [1(y_i \leqslant \beta) - \tau, y_i - \beta]'$, let the population and sample moments be

$$\pi(\beta) = [P(y_i \leqslant \beta) - 0.5, Ey_i - \beta]', \qquad \hat{\pi}_n(\beta) = \left[\frac{1}{n}\sum_{i=1}^{n} 1(y_i \leqslant \beta) - 0.5, \frac{1}{n}\sum_{i=1}^{n} y_i - \beta\right]'$$

Our GMM estimator has a non-negativity constraint:

$$\hat{\beta}_n = \arg\min_{\beta \geqslant 0} \left\{\hat{Q}_n(\beta) = \frac{1}{2}\hat{\pi}_n(\beta)'\hat{\pi}_n(\beta)\right\}$$

We will examine the empirical coverage and average interval length of the proximal bootstrap confidence set $\mathcal{C}^*_{1-\alpha} = \left\{\beta : n\left(\hat{Q}_n(\beta) - \inf_{h \in \mathcal{B}_{\delta_n}} \hat{Q}_n\left(\beta + \frac{h}{\sqrt{n}}\right)\right) \leqslant \hat{c}^*_{1-\alpha}\right\}$, where $\hat{c}^*_{1-\alpha}$ is the $1-\alpha$ quantile of $-\frac{\inf_{\beta \in \mathbb{B}} \hat{Z}^*_n(\beta)}{\alpha_n^2}$ and $\hat{Z}^*_n(\beta) = \alpha_n\sqrt{n}\left(\hat{l}^*_n(\bar{\beta}_n) - \hat{l}_n(\bar{\beta}_n)\right)'(\beta - \bar{\beta}_n) + \frac{1}{2}\|\beta - \bar{\beta}_n\|^2_{\bar{H}_n}$, for $\bar{H}_n = \hat{G}'_n\hat{G}_n + \hat{L}'_n\hat{\pi}_n(\bar{\beta}_n)$, $\hat{l}_n(\bar{\beta}_n) = \hat{G}'_n\hat{\pi}_n(\bar{\beta}_n)$, $\hat{l}^*_n(\bar{\beta}_n) = \hat{G}^{*\prime}_n\hat{\pi}^*_n(\bar{\beta}_n)$, and

$$\hat{G}_n = \begin{bmatrix} \frac{1}{nh}\sum_{i=1}^n K_h\left(y_i - \hat{\beta}_n\right) \\ -1 \end{bmatrix}, \hat{G}^*_n = \begin{bmatrix} \frac{1}{nh}\sum_{i=1}^n K_h\left(y^*_i - \hat{\beta}_n\right) \\ -1 \end{bmatrix}, \hat{L}_n = \begin{bmatrix} \frac{1}{nh^2}\sum_{i=1}^n K'_h\left(y_i - \hat{\beta}_n\right) \\ 0 \end{bmatrix},$$

$K_h(x) = K(x/h)$, $K(x) = (2\pi)^{-1/2}\exp(-x^2/2)$, $K'_h(x) = K'(x/h)$ and $K'(x) = -(2\pi)^{-1/2}x\exp(-x^2/2)$. We use the Silverman's rule of thumb bandwidth $h = 1.06n^{-1/5}$.

We consider possibly drifting sequences of parameters $\beta_0 \in \{0, n^{-1}, n^{-1/2}, n^{-1/3}, n^{-1/4}, n^{-1/6}, 2\}$. We consider four different sample sizes $n \in \{100, 500, 1000, 5000\}$ and we use 1000 bootstrap iterations and 2000 Monte Carlo simulations. Table 2 shows the empirical coverage frequencies and average interval lengths (in parentheses) of nominal 95% confidence intervals constructed using the proximal bootstrap, subsampling, and the standard nonparametric bootstrap. To the best of

Table 1: Empirical Coverage Frequencies and Average Interval Lengths

β_0	0	n^{-1}	$n^{-1/2}$	$n^{-1/3}$	$n^{-1/4}$	$n^{-1/6}$	1
$n = 100$							
Proximal Bootstrap	0.944	0.944	0.947	0.940	0.950	0.951	0.943
	(0.379)	(0.380)	(0.379)	(0.379)	(0.377)	(0.380)	(0.379)
Hsieh et al. (2022)	0.993	0.984	0.983	0.985	0.989	0.980	0.981
	(0.478)	(0.478)	(0.478)	(0.479)	(0.477)	(0.478)	(0.479)
Fang and Santos (2019)	0.975	0.470	0.586	0.953	0.954	0.939	0.952
	(0.028)	(0.028)	(0.043)	(0.055)	(0.055)	(0.055)	(0.055)
Subsampling	0.980	0.620	0.534	0.663	0.777	0.944	0.947
	(0.028)	(0.028)	(0.029)	(0.035)	(0.042)	(0.055)	(0.055)
Nonparametric Bootstrap	0.972	0.609	0.679	0.948	0.946	0.955	0.953
	(0.028)	(0.027)	(0.041)	(0.055)	(0.055)	(0.055)	(0.055)
$n = 500$							
Proximal Bootstrap	0.952	0.941	0.949	0.957	0.945	0.960	0.952
	(0.165)	(0.165)	(0.165)	(0.165)	(0.165)	(0.165)	(0.165)
Hsieh et al. (2022)	0.990	0.987	0.987	0.984	0.990	0.985	0.987
	(0.209)	(0.209)	(0.209)	(0.209)	(0.209)	(0.209)	(0.209)
Fang and Santos (2019)	0.975	0.470	0.586	0.953	0.954	0.939	0.952
	(0.028)	(0.028)	(0.043)	(0.055)	(0.055)	(0.055)	(0.055)
Subsampling	0.980	0.620	0.534	0.663	0.777	0.944	0.947
	(0.028)	(0.028)	(0.029)	(0.035)	(0.042)	(0.055)	(0.055)
Nonparametric Bootstrap	0.972	0.609	0.679	0.948	0.946	0.955	0.953
	(0.028)	(0.027)	(0.041)	(0.055)	(0.055)	(0.055)	(0.055)
$n = 1000$							
Proximal Bootstrap	0.954	0.946	0.947	0.953	0.953	0.945	0.950
	(0.114)	(0.114)	(0.114)	(0.114)	(0.114)	(0.114)	(0.114)
Hsieh et al. (2022)	0.994	0.988	0.983	0.984	0.987	0.989	0.989
	(0.145)	(0.145)	(0.145)	(0.145)	(0.145)	(0.145)	(0.145)
Fang and Santos (2019)	0.975	0.470	0.586	0.953	0.954	0.939	0.952
	(0.028)	(0.028)	(0.043)	(0.055)	(0.055)	(0.055)	(0.055)
Subsampling	0.980	0.620	0.534	0.663	0.777	0.944	0.947
	(0.028)	(0.028)	(0.029)	(0.035)	(0.042)	(0.055)	(0.055)
Nonparametric Bootstrap	0.972	0.609	0.679	0.948	0.946	0.955	0.953
	(0.028)	(0.027)	(0.041)	(0.055)	(0.055)	(0.055)	(0.055)
$n = 5000$							
Proximal Bootstrap	0.948	0.951	0.947	0.950	0.945	0.951	0.958
	(0.045)	(0.045)	(0.045)	(0.045)	(0.046)	(0.045)	(0.046)
Hsieh et al. (2022)	0.990	0.985	0.991	0.985	0.982	0.990	0.987
	(0.059)	(0.059)	(0.059)	(0.059)	(0.059)	(0.059)	(0.059)
Fang and Santos (2019)	0.975	0.470	0.586	0.953	0.954	0.939	0.952
	(0.028)	(0.028)	(0.043)	(0.055)	(0.055)	(0.055)	(0.055)
Subsampling	0.980	0.620	0.534	0.663	0.777	0.944	0.947
	(0.028)	(0.028)	(0.029)	(0.035)	(0.042)	(0.055)	(0.055)
Nonparametric Bootstrap	0.972	0.609	0.679	0.948	0.946	0.955	0.953
	(0.028)	(0.027)	(0.041)	(0.055)	(0.055)	(0.055)	(0.055)

our knowledge, Hsieh et al. (2022)'s method does not apply for this example because it is not a quadratic programming problem. We are also unable to use Fang and Santos (2019) because there is no closed form solution to the optimization problem. For choosing α_n, we apply the double bootstrap method described in Section 2.3 using $n = 5000$, $B_1 = B_2 = 5000$, $\beta_0 = 0$ over the grid $\alpha_n \in \left\{n^{-1/3}, n^{-1/4}, n^{-1/6}, n^{-1/7}, n^{-1/8}, n^{-1/9}, n^{-1/10}\right\}$. The empirical coverage frequencies were $\{0.9512, 0.9502, 0.9506, 0.9478, 0.9432, 0.9368, 0.9338\}$. $\alpha_n = n^{-1/3}$ was the smallest value which achieved coverage at or above the nominal level of 0.95. We also tried using all the other values of α_n and found that the coverage was the same up to three decimal places across the different values of α_n. We did not constrain h when computing $\inf_{h \in \mathcal{B}_{\delta_n}} \hat{Q}_n\left(\beta + \frac{h}{\sqrt{n}}\right)$, which effectively sets δ_n to $\sqrt{n}\delta_n \to \infty$.

The coverage of the proximal bootstrap is close to the nominal level for all values of β_0 while the coverage of subsampling and the standard nonparametric bootstrap are far below the nominal level for drifting values of $\beta_0 \in \left\{n^{-1}, n^{-1/2}, n^{-1/3}, n^{-1/4}, n^{-1/6}\right\}$. The coverage is worst when $\beta_0 = n^{-1}$, where it can drop to around 50%. The average interval lengths of the proximal bootstrap are somewhat larger than the other methods.

3.3 Conditional Logit Model with Estimated Inequality Constraints

We generate data according to $y_{ij} = 1\left(y^*_{ij} > y^*_{ik} \forall k \neq j\right)$, where the utility of individual $i = 1...n$ from picking choice $j = 1...J$ is given by

$$y^*_{ij} = \beta_0 x_{ij} + \epsilon_{ij}, \text{ for } x_i \sim N\left(\begin{pmatrix} 1 \\ 2 \\ \vdots \\ J \end{pmatrix}, \begin{pmatrix} 1 & 0.5 & \dots & 0.5 \\ 0.5 & 1 & \dots & 0.5 \\ \vdots & \vdots & \vdots & \vdots \\ 0.5 & 0.5 & \dots & 1 \end{pmatrix}\right)$$

and $\epsilon_{ij} \overset{i.i.d.}{\sim}$ Type 1 Extreme Value. We set $\beta_0 = 0.1$. The constrained MLE estimator maximizes the log-likelihood subject to the constraints that the share of individuals who pick each choice cannot exceed the supply of that choice. These inequality constraints can be viewed as capacity constraints similar to the ones in de Palma et al. (2007) which state that the equilibrium demand

Table 2: Empirical Coverage Frequencies and Average Interval Lengths

β_0	0	n^{-1}	$n^{-1/2}$	$n^{-1/3}$	$n^{-1/4}$	$n^{-1/6}$	2
$n = 100$							
Proximal Bootstrap	0.946	0.946	0.946	0.947	0.947	0.947	0.947
	(0.380)	(0.380)	(0.381)	(0.381)	(0.381)	(0.381)	(0.381)
Subsampling	0.969	0.496	0.587	0.686	0.761	0.847	0.939
	(0.211)	(0.213)	(0.232)	(0.266)	(0.297)	(0.341)	(0.390)
Nonparametric Bootstrap	0.969	0.518	0.671	0.844	0.916	0.947	0.947
	(0.236)	(0.240)	(0.294)	(0.359)	(0.383)	(0.390)	(0.388)
$n = 500$							
Proximal Bootstrap	0.953	0.952	0.953	0.954	0.953	0.953	0.953
	(0.165)	(0.165)	(0.166)	(0.166)	(0.166)	(0.166)	(0.166)
Subsampling	0.974	0.490	0.539	0.646	0.773	0.897	0.952
	(0.092)	(0.093)	(0.099)	(0.115)	(0.132)	(0.162)	(0.175)
Nonparametric Bootstrap	0.971	0.490	0.666	0.885	0.941	0.944	0.941
	(0.106)	(0.107)	(0.133)	(0.169)	(0.175)	(0.175)	(0.175)
$n = 1000$							
Proximal Bootstrap	0.945	0.944	0.944	0.945	0.944	0.944	0.944
	(0.115)	(0.115)	(0.115)	(0.115)	(0.115)	(0.115)	(0.115)
Subsampling	0.981	0.519	0.562	0.686	0.797	0.923	0.962
	(0.065)	(0.065)	(0.069)	(0.080)	(0.093)	(0.118)	(0.124)
Nonparametric Bootstrap	0.969	0.497	0.681	0.917	0.940	0.943	0.943
	(0.076)	(0.076)	(0.095)	(0.122)	(0.124)	(0.124)	(0.124)
$n = 5000$							
Proximal Bootstrap	0.953	0.952	0.952	0.953	0.953	0.953	0.953
	(0.046)	(0.046)	(0.046)	(0.046)	(0.046)	(0.046)	(0.046)
Subsampling	0.973	0.558	0.522	0.648	0.785	0.944	0.959
	(0.028)	(0.028)	(0.029)	(0.035)	(0.042)	(0.055)	(0.055)
Nonparametric Bootstrap	0.975	0.601	0.685	0.953	0.954	0.952	0.954
	(0.034)	(0.034)	(0.042)	(0.055)	(0.055)	(0.055)	(0.055)

for each housing unit should not exceed the supply of that housing unit. For $P_{ij}(\beta) \equiv \frac{\exp(\beta x_{ij})}{\sum_l \exp(\beta x_{il})}$,

$$\hat{\beta}_n = \arg\max_{\beta} \ \ln L(\beta) = \frac{1}{nJ}\sum_{i=1}^{n}\sum_{j=1}^{J} y_{ij}\ln P_{ij}(\beta)$$
$$\text{s.t. } \frac{1}{n}\sum_{i=1}^{n} P_{ij}(\beta) \leqslant \bar{b}_j \text{ for all } j = 1...J$$

where $\bar{b}_j = \frac{1}{10^6}\sum_{i=1}^{10^6}\frac{\exp(\beta_0\tilde{x}_{ij})}{\sum_l \exp(\beta_0\tilde{x}_{il})}$ for $\tilde{x}_{ij}$ drawn independently from the same distribution as x_{ij}.

We examine the empirical coverage and average length of the proximal bootstrap confidence set $\mathcal{C}^*_{1-\alpha} = \left\{\beta : n\left(\hat{Q}_n(\beta) - \inf_{h\in\mathcal{B}_{\delta_n}}\hat{Q}_n\left(\beta + \frac{h}{\sqrt{n}}\right)\right) \leqslant \hat{c}^*_{1-\alpha}\right\}$, where $\hat{c}^*_{1-\alpha}$ is the $1-\alpha$ quantile of $-\frac{\inf_{\beta\in\mathbb{B}}\hat{Z}^*_n(\beta)}{\alpha_n^2}$ for $\hat{Z}^*_n(\beta) = \alpha_n\sqrt{n}\left(\hat{l}^*_n(\bar{\beta}_n) - \hat{l}_n(\bar{\beta}_n)\right)'(\beta - \bar{\beta}_n) + \frac{1}{2}\|\beta - \bar{\beta}_n\|^2_{\bar{H}_n}$ and $\bar{\beta}_n = \hat{\beta}_n$. We use analytic expressions for the components in the proximal bootstrap objective function:

$$\hat{l}_n(\beta) = -\frac{\partial\ln L(\beta)}{\partial\beta} = -\frac{1}{nJ}\sum_{i=1}^{n}\sum_{j=1}^{J}(y_{ij} - P_{ij}(\beta))\,x_{ij}$$
$$H_n(\beta) = -\frac{\partial^2\ln L(\beta)}{\partial\beta\partial\beta'} = \frac{1}{nJ}\sum_{i=1}^{n}\sum_{j=1}^{J}P_{ij}(\beta)\left(x_{ij} - \sum_l P_{il}(\beta)\,x_{il}\right)\left(x_{ij} - \sum_l P_{il}(\beta)\,x_{il}\right)'$$

We consider $n \in \{100, 500, 1000, 5000\}$, $J = 20$, $\alpha_n \in \left\{n^{-1/3}, n^{-1/4}, n^{-1/6}, n^{-1/8}, n^{-1/10}\right\}$, $B = 1000$ bootstrap iterations, and $R = 2000$ Monte Carlo simulations. Empirical coverage frequencies for the proximal bootstrap confidence set, subsampling equal-tailed interval, and standard nonparametric bootstrap equal-tailed interval, as well as average interval lengths are reported in Table 3. The proximal bootstrap coverage frequencies and average interval lengths are the same up to three decimal places across the different values of α_n. The proximal bootstrap coverage frequencies are very close to the nominal level of 95% for sufficiently large values of n. Both subsampling and the standard nonparametric bootstrap undercover for all values of n, with the standard nonparametric bootstrap having worse coverage than subsampling.

Table 3: Empirical Coverage Frequencies and Average Interval Lengths

n	100	500	1000	5000
$\alpha_n = n^{-1/3}$	0.936	0.951	0.949	0.951
	(0.073)	(0.032)	(0.022)	(0.009)
$\alpha_n = n^{-1/4}$	0.936	0.951	0.949	0.951
	(0.073)	(0.032)	(0.022)	(0.009)
$\alpha_n = n^{-1/6}$	0.936	0.951	0.949	0.951
	(0.073)	(0.032)	(0.022)	(0.009)
$\alpha_n = n^{-1/8}$	0.936	0.951	0.949	0.951
	(0.073)	(0.032)	(0.022)	(0.009)
$\alpha_n = n^{-1/10}$	0.936	0.951	0.949	0.951
	(0.073)	(0.032)	(0.022)	(0.009)
Subsampling	0.927	0.937	0.939	0.933
	(0.002)	(0.001)	(0.000)	(0.000)
Nonparametric Bootstrap	0.917	0.928	0.916	0.903
	(0.002)	(0.001)	(0.001)	(0.000)

4 Conclusion

We have demonstrated how to use a computationally efficient bootstrap procedure to conduct asymptotically valid inference for $\sqrt{n}$-consistent constrained optimization estimators with nonstandard asymptotic distributions. Our proximal bootstrap estimator can be expressed as the solution to a quadratic programming problem and relies on a scaling sequence that converges to zero at a slower than $\sqrt{n}$ rate. We have illustrated its empirical performance in boundary constrained MLE and GMM problems and a conditional logit model with capacity constraints.

5 Appendix

5.1 Proofs of Theorems

5.1.1 Proof of Theorem 1

Using the arguments in Theorem 2.1 of Shapiro (1988) and Lemma 3.1 of Shapiro (1989), when $\hat{\beta}_n$ lies in a neighborhood of β_0, $\hat{\beta}_n$ is almost surely a minimizer of $\tilde{\mathcal{L}}_n(\beta) = \hat{Q}_n(\beta) + \sum_{j \in \mathcal{E} \cup \mathcal{I}} \lambda_{0j} f_j(\beta)$ over $C(\lambda_0) = \left\{\beta \in \mathbb{B} : f_j(\beta) = 0 \text{ for } j \in \mathcal{E} \cup \mathcal{I}_+^*(\lambda_0), f_j(\beta) \leqslant 0 \text{ for } j \in \mathcal{I}_0^*(\lambda_0)\right\}$, where $\mathcal{I}_+^*(\lambda_0) \equiv \{j \in \mathcal{I}^* : \lambda_{0j} > 0\}$, $\mathcal{I}_0^*(\lambda_0) \equiv \{j \in \mathcal{I}^* : \lambda_{0j} = 0\}$, and $\mathcal{I}^* = \{j \in \mathcal{I} : f_j(\beta_0) = 0\}$.

Consistency of $\hat{\beta}_n$ for β_0 follows from Assumption 1 and Corollary 3.2.3 in van der Vaart and

Wellner (1996). We can show that consistency implies $\sqrt{n}$-consistency using a modified version of the first part of the proof of Theorem 5 on page 141 of Pollard (1984) to allow for constraints. We need to replace his population objective $F(\cdot)$ with the population Lagrangian $\mathcal{L}(\beta_0, \lambda_0) \equiv Q(\beta_0) + \sum_{j \in \mathcal{E} \cup \mathcal{I}} \lambda_{0j} f_j(\beta_0)$. The first order KKT condition $\nabla \mathcal{L}(\beta_0, \lambda_0) \equiv l(\beta_0) + \sum_{j \in \mathcal{E} \cup \mathcal{I}} \lambda_{0j} F_{0j} = 0$ implies the local quadratic expansion $\mathcal{L}(\beta, \lambda_0) = \mathcal{L}(\beta_0, \lambda_0) + \frac{1}{2} \|\beta - \beta_0\|^2_{\nabla^2 \mathcal{L}(\beta_0, \lambda_0)} + o\left(\|\beta - \beta_0\|^2\right)$ for β in a small neighborhood of β_0. This expansion in combination with the local quadratic approximation of the Lagrangian in Assumption 5 will imply a modified version of Pollard (1984)'s equation (6), where $F_n(\cdot)$ is replaced by $\tilde{\mathcal{L}}_n(\cdot)$ and the empirical process $E_n \Delta$ is replaced by $\sqrt{n}\left(\hat{l}_n(\beta_0) - l(\beta_0)\right)$.

We assumed in condition (iii) that $\mathcal{I}^*_+(\lambda_0) = \varnothing$, which means $\mathcal{I}^* = \mathcal{I}^*_0(\lambda_0)$, and $C(\lambda_0) = \{\beta \in \mathbb{B} : f_j(\beta) = 0 \text{ for } j \in \mathcal{E}, f_j(\beta) \leqslant 0 \text{ for } j \in \mathcal{I}^*\}$.

Denote the feasible direction set by

$$\mathcal{F}_n = \left\{h : f_j\left(\beta_0 + \frac{h}{\sqrt{n}}\right) = 0 \text{ for } j \in \mathcal{E}, f_j\left(\beta_0 + \frac{h}{\sqrt{n}}\right) \leqslant 0 \text{ for } j \in \mathcal{I}^*\right\}$$

Denote the linearized feasible direction set by

$$\Sigma_n = \left\{h : \sqrt{n} f_j(\beta_0) + F'_{0j} h = 0 \text{ for } j \in \mathcal{E}, \sqrt{n} f_j(\beta_0) + F'_{0j} h \leqslant 0 \text{ for } j \in \mathcal{I}^*\right\}$$

LICQ implies the linearized feasible direction set is sufficient to capture the geometry of the constraints near β_0 so that $\sqrt{n}\left(\hat{\beta}_n - \beta_0\right)$ is asymptotically equivalent to the minimizer of the Lagrangian over Σ_n:

$$\begin{aligned}
\sqrt{n}\left(\hat{\beta}_n - \beta_0\right) &= \underset{h \in \Sigma_n}{\arg\min} \left\{n \tilde{\mathcal{L}}_n\left(\beta_0 + \frac{h}{\sqrt{n}}\right) - n \tilde{\mathcal{L}}_n(\beta_0)\right\} + o_P(1) \\
&= \underset{h \in \Sigma_n}{\arg\min} \left\{n \hat{Q}_n\left(\beta_0 + \frac{h}{\sqrt{n}}\right) - n \hat{Q}_n(\beta_0) + \sum_{j \in \mathcal{E} \cup \mathcal{I}} \lambda_{0j} n \left(f_j\left(\beta_0 + \frac{h}{\sqrt{n}}\right) - f_j(\beta_0)\right)\right\} + o_P(1) \\
&\rightsquigarrow \underset{h \in \Sigma}{\arg\min} \left\{h' W_0 + \frac{1}{2} h' H_0 h + \frac{1}{2} \sum_{j \in \mathcal{E}} \lambda_{0j} h' G_{0j} h\right\} = \mathcal{J}
\end{aligned}$$

where the convergence result in the last line follows from the following arguments. First note that

Assumption 5 implies that for any $\delta_n \to 0$, and $\mathcal{B}_{\delta_n} = \left\{h \in \mathbb{R}^d : \frac{\|h\|}{\sqrt{n}} \leqslant \delta_n\right\}$,

$$\sup_{h \in \mathcal{B}_{\delta_n}} \left| \frac{n\tilde{\mathcal{L}}_n\left(\beta_0 + \frac{h}{\sqrt{n}}\right) - n\tilde{\mathcal{L}}_n\left(\beta_0\right) - h'\sqrt{n}\left(\hat{l}_n\left(\beta_0\right) - l\left(\beta_0\right)\right) - \frac{1}{2}h'H_0h - \frac{1}{2}\sum_{j \in \mathcal{E} \cup \mathcal{I}} \lambda_{0j} h' G_{0j} h}{1 + \|h\|^2} \right| = o_P(1)$$

Recall $\sqrt{n}\left(\hat{l}_n\left(\beta_0\right) - l\left(\beta_0\right)\right) \rightsquigarrow W_0$ and $\lambda_{0j} = 0$ for all $j \in \mathcal{I} \backslash \mathcal{I}_+^*\left(\lambda_0\right)$, where we have assumed $\mathcal{I}_+^*\left(\lambda_0\right) = \varnothing$. Since pointwise convergence implies uniform convergence over compact sets $K \subset \mathbb{R}^d$ for convex functions of h, we have that uniformly in $h \in \mathcal{B}_{\delta_n}$,

$$\begin{aligned}
& n\hat{Q}_n\left(\beta_0 + \frac{h}{\sqrt{n}}\right) - n\hat{Q}_n\left(\beta_0\right) + \sum_{j \in \mathcal{E} \cup \mathcal{I}} \lambda_{0j} n\left(f_j\left(\beta_0 + \frac{h}{\sqrt{n}}\right) - f_j\left(\beta_0\right)\right) \\
& = h'\sqrt{n}\left(\hat{l}_n\left(\beta_0\right) - l\left(\beta_0\right)\right) + \frac{1}{2}h'H_0h + \frac{1}{2}\sum_{j \in \mathcal{E} \cup \mathcal{I}} \lambda_{0j} h' G_{0j} h + o_P(1) \\
& \rightsquigarrow h'W_0 + \frac{1}{2}h'H_0h + \frac{1}{2}\sum_{j \in \mathcal{E}} \lambda_{0j} h' G_{0j} h
\end{aligned}$$

as a process indexed by h in the space of bounded functions on compact sets $\ell^\infty(K)$ for any compact $K \subset \mathbb{R}^d$.

Now consider the constraints. Since $\sqrt{n}f_j\left(\beta_0\right) + F_{0j}'h \xrightarrow{p} -\infty$ for $j \in \mathcal{I} \backslash \mathcal{I}^*$, the nonactive inequality constraints do not affect the asymptotic distribution. Also, $\sqrt{n}f_j\left(\beta_0\right) = 0$ for all $j \in \mathcal{E} \cup \mathcal{I}^*$. Condition (i) is a second order sufficient condition and guarantees that the argmin of $h'W_0 + \frac{1}{2}h'H_0h + \frac{1}{2}\sum_{j \in \mathcal{E}} \lambda_{0j} h' G_{0j} h$ over Σ is unique. Then by the argmin continuous mapping theorem (Theorem 1 of Knight (1999)), $\arg\min_h \hat{\mathbb{G}}_n(h) \to_{e-d} \arg\min_h \mathbb{G}_0(h)$, where

$$\begin{aligned}
\hat{\mathbb{G}}_n(h) &= n\hat{Q}_n\left(\beta_0 + \frac{h}{\sqrt{n}}\right) - n\hat{Q}_n\left(\beta_0\right) + \sum_{j \in \mathcal{E} \cup \mathcal{I}} \lambda_{0j} n\left(f_j\left(\beta_0 + \frac{h}{\sqrt{n}}\right) - f_j\left(\beta_0\right)\right) + \infty 1\left(h \notin \Sigma_n\right) \\
\mathbb{G}_0(h) &= h'W_0 + \frac{1}{2}h'H_0h + \frac{1}{2}\sum_{j \in \mathcal{E}} \lambda_{0j} h' G_{0j} h + \infty 1\left(h \notin \Sigma\right) \\
\Sigma &= \left\{h : F_{0j}'h = 0 \text{ for } j \in \mathcal{E}, F_{0j}'h \leqslant 0 \text{ for } j \in \mathcal{I}^*\right\}
\end{aligned}$$

Now we show consistency of the proximal bootstrap. $\alpha_n \to 0$ implies $\alpha_n\sqrt{n}\bar{H}_n\left(\hat{l}_n^*\left(\bar{\beta}_n\right) - \hat{l}_n\left(\bar{\beta}_n\right)\right) = o_p^*(1)$. Using convexity of the proximal bootstrap objective function, compactness of $C^* - \beta_0$, and the fact that $\bar{\beta}_n \in C^*$,

$$\begin{aligned}
\hat{\beta}_n^* - \beta_0 = &\underset{u \in (C^* - \beta_0)}{\arg\min} \left\{ \frac{1}{2} \left\| u + \beta_0 - \bar{\beta}_n + \alpha_n \sqrt{n} \bar{H}_n^{-1} \left(\hat{l}_n^* \left(\bar{\beta}_n \right) - \hat{l}_n \left(\bar{\beta}_n \right) \right) \right\|_{\bar{H}_n}^2 \right. \\
&\left. + \frac{1}{2} \sum_{j \in \mathcal{E} \cup \mathcal{I}} \bar{\lambda}_{nj} \left\| u + \beta_0 - \bar{\beta}_n \right\|_{\bar{G}_{nj}}^2 \right\} \\
= &\underset{u \in (C^* - \beta_0)}{\arg\min} \left\{ \frac{1}{2} \left\| u + \beta_0 - \bar{\beta}_n \right\|_{\bar{H}_n}^2 + \frac{1}{2} \sum_{j \in \mathcal{E} \cup \mathcal{I}} \bar{\lambda}_{nj} \left\| u + \beta_0 - \bar{\beta}_n \right\|_{\bar{G}_{nj}}^2 + o_p^*(1) \right\} \\
= &\bar{\beta}_n - \beta_0 + o_p(1) = o_p(1)
\end{aligned}$$

Note that since C^* is already a linearized constraint set, the linearized feasible direction set is simply

$$\begin{aligned}
\Sigma_n^* = &\left\{ h : f_j \left(\bar{\beta}_n \right) + \bar{F}_{nj}' \left(\beta_0 - \bar{\beta}_n + \alpha_n h \right) = 0 \text{ for } j \in \mathcal{E}, \right. \\
&\left. f_j \left(\bar{\beta}_n \right) + \bar{F}_{nj}' \left(\beta_0 - \bar{\beta}_n + \alpha_n h \right) \leqslant 0 \text{ for } j \in \mathcal{I} \right\} \\
= &\left\{ h : \frac{f_j \left(\bar{\beta}_n \right)}{\alpha_n} + \bar{F}_{nj}' h + \bar{F}_{nj}' \left(\frac{\beta_0 - \bar{\beta}_n}{\alpha_n} \right) = 0 \text{ for } j \in \mathcal{E}, \right. \\
&\left. \frac{f_j \left(\bar{\beta}_n \right)}{\alpha_n} + \bar{F}_{nj}' h + \bar{F}_{nj}' \left(\frac{\beta_0 - \bar{\beta}_n}{\alpha_n} \right) \leqslant 0 \text{ for } j \in \mathcal{I} \right\}
\end{aligned}$$

Using the local parameter $h \in \frac{C^* - \beta_0}{\alpha_n}$, we can derive the asymptotic distribution of the proximal bootstrap.

$$\begin{aligned}
\frac{\hat{\beta}_n^* - \beta_0}{\alpha_n} = &\underset{h \in \Sigma_n^*}{\arg\min} \left\{ \alpha_n \sqrt{n} \left(\hat{l}_n^* \left(\bar{\beta}_n \right) - \hat{l}_n \left(\bar{\beta}_n \right) \right)' \left(\beta_0 - \bar{\beta}_n + \alpha_n h \right) + \frac{1}{2} \left\| \beta_0 - \bar{\beta}_n + \alpha_n h \right\|_{\bar{H}_n}^2 \right. \\
&\left. + \frac{1}{2} \sum_{j \in \mathcal{E} \cup \mathcal{I}} \bar{\lambda}_{nj} \left\| \beta_0 - \bar{\beta}_n + \alpha_n h \right\|_{\bar{G}_{nj}}^2 \right\} \\
= &\underset{h \in \Sigma_n^*}{\arg\min} \left\{ \sqrt{n} \left(\hat{l}_n^* \left(\bar{\beta}_n \right) - \hat{l}_n \left(\bar{\beta}_n \right) \right)' \left(\frac{\beta_0 - \bar{\beta}_n}{\alpha_n} + h \right) + \frac{1}{2} \left\| \frac{\beta_0 - \bar{\beta}_n}{\alpha_n} + h \right\|_{\bar{H}_n}^2 \right. \\
&\left. + \frac{1}{2} \sum_{j \in \mathcal{E} \cup \mathcal{I}} \bar{\lambda}_{nj} \left\| \frac{\beta_0 - \bar{\beta}_n}{\alpha_n} + h \right\|_{\bar{G}_{nj}}^2 \right\} \\
= &\underset{h \in \Sigma_n^*}{\arg\min} \left\{ h' \sqrt{n} \left(\hat{l}_n^* \left(\bar{\beta}_n \right) - \hat{l}_n \left(\bar{\beta}_n \right) \right) + \frac{1}{2} h' \bar{H}_n h + \frac{1}{2} \sum_{j \in \mathcal{E} \cup \mathcal{I}} \bar{\lambda}_{nj} h' \bar{G}_{nj} h + o_P^*(1) \right\}
\end{aligned}$$

$$\overset{\mathbb{P}}{\underset{\mathbb{W}}{\rightsquigarrow}} \arg\min_{h \in \Sigma} \left\{ h' W_0 + \frac{1}{2} h' H_0 h + \frac{1}{2} \sum_{j \in \mathcal{E}} \lambda_{0j} h' G_{0j} h \right\} = \mathcal{J}$$

where the last line follows from the following arguments. First, note that under the envelope integrability assumption 3, Lemma 4.2 in Wellner and Zhan (1996) implies that for any compact $K \subset \mathbb{R}^d$,

$$\left\| \sqrt{n} \left(P_n^* - P_n \right) \left(g\left(\cdot, \bar{\beta}_n\right) - g\left(\cdot, \beta_0\right) \right) \right\| = o_p^* \left(1 + \sqrt{n} \left\| \bar{\beta}_n - \beta_0 \right\| \right) = o_p^*(1)$$

This bootstrap equicontinuity result implies $\sqrt{n} \left(\hat{l}_n^* \left(\bar{\beta}_n \right) - \hat{l}_n \left(\bar{\beta}_n \right) \right)$ and $\sqrt{n} \left(\hat{l}_n^* \left(\beta_0 \right) - \hat{l}_n \left(\beta_0 \right) \right)$ have the same asymptotic distribution. Additionally, since $\bar{H}_n \xrightarrow{p} H_0$, $\bar{G}_{nj} \xrightarrow{p} G_{0j}$ for all j, $\max_{j \in \mathcal{E} \cup \mathcal{I}} \left| \bar{\lambda}_{nj} - \lambda_{0j} \right| \xrightarrow{p} 0$, and and the proximal bootstrap Lagrangian is convex in h, we have that uniformly over compact sets $K \subset \mathbb{R}^d$,

$$\begin{aligned}
& h' \sqrt{n} \left(\hat{l}_n^* \left(\bar{\beta}_n \right) - \hat{l}_n \left(\bar{\beta}_n \right) \right) + \frac{1}{2} h' \bar{H}_n h + \frac{1}{2} \sum_{j \in \mathcal{E} \cup \mathcal{I}} \bar{\lambda}_{nj} h' \bar{G}_{nj} h \\
&= h' \sqrt{n} \left(\hat{l}_n^* \left(\bar{\beta}_n \right) - \hat{l}_n \left(\bar{\beta}_n \right) \right) + \frac{1}{2} h' H_0 h + \frac{1}{2} \sum_{j \in \mathcal{E} \cup \mathcal{I}} \lambda_{0j} h' G_{0j} h + o_P(1) \\
&\overset{\mathbb{P}}{\underset{\mathbb{W}}{\rightsquigarrow}} h' W_0 + \frac{1}{2} h' H_0 h + \frac{1}{2} \sum_{j \in \mathcal{E} \cup \mathcal{I}} \lambda_{0j} h' G_{0j} h \\
&= h' W_0 + \frac{1}{2} h' H_0 h + \frac{1}{2} \sum_{j \in \mathcal{E}} \lambda_{0j} h' G_{0j} h
\end{aligned}$$

as a process indexed by h in the space of bounded functions on compact sets $\ell^\infty(K)$ for any compact $K \subset \mathbb{R}^d$.

For the proximal bootstrap constraint set, note that $\frac{f_j\left(\bar{\beta}_n\right)}{\alpha_n} \xrightarrow{p} -\infty$ for $j \in \mathcal{I} \backslash \mathcal{I}^*$ while $\frac{f_j\left(\bar{\beta}_n\right)}{\alpha_n} = \frac{\sqrt{n}\left(f_j\left(\bar{\beta}_n\right) - f_j\left(\beta_0\right)\right)}{\sqrt{n}\alpha_n} = o_P(1)$ for $j \in \mathcal{E} \cup \mathcal{I}^*$. Additionally, $\bar{F}_{nj}' \left(\frac{\beta_0 - \bar{\beta}_n}{\alpha_n} \right) = o_P(1)$ and $\bar{F}_{nj} = F_{0j} + o_P(1)$ for all $j \in \mathcal{E} \cup \mathcal{I}$. Then, by a modification of the bootstrap argmin continuous mapping lemma 14.2 in Hong and Li (2020) that replaces weak convergence with epi-convergence, $\arg\min_h \hat{\mathbb{G}}_n^*(h) \underset{e-d}{\xrightarrow{p}} \arg\min_h \mathbb{G}_0(h)$ for

$$\hat{\mathbb{G}}_n^*(h) = h' \sqrt{n} \left(\hat{l}_n^* \left(\bar{\beta}_n \right) - \hat{l}_n \left(\bar{\beta}_n \right) \right) + \frac{1}{2} h' \bar{H}_n h + \frac{1}{2} \sum_{j \in \mathcal{E} \cup \mathcal{I}} \bar{\lambda}_{nj} h' \bar{G}_{nj} h + \infty 1 \left(h \notin \Sigma_n^* \right)$$

$$\mathbb{G}_0(h) = h'W_0 + \frac{1}{2}h'H_0h + \frac{1}{2}\sum_{j\in\mathcal{E}}\lambda_{0j}h'G_{0j}h + \infty 1(h \notin \Sigma)$$

Here, $\underset{e-d}{\overset{p}{\to}}$ denotes epi-convergence of the conditional law of $\hat{\mathbb{G}}_n^*$ to $\mathbb{G}_0$, which can be equivalently stated as $\sup_{f\in BL_1}\left|E_{\mathbb{W}}f\left(\hat{\mathbb{G}}_n^*\right) - Ef\left(\mathbb{G}_0\right)\right| \xrightarrow{p} 0$ and $E_{\mathbb{W}}f\left(\hat{\mathbb{G}}_n^*\right)^* - E_{\mathbb{W}}f\left(\hat{\mathbb{G}}_n^*\right)_* \xrightarrow{p} 0$ for all $f \in BL_1$, where BL_1 is the class of Lipschitz norm 1 functions with respect to the metric of epi-convergence defined as $d\left(\hat{\mathbb{G}}_n^*, \mathbb{G}_0\right) = \int_0^\infty \max\left\{\left|d_{\text{epi }\hat{\mathbb{G}}_n^*}(v) - d_{\text{epi }\mathbb{G}_0}(v)\right| : |v| \leqslant \rho\right\}\exp(-\rho)\,d\rho$, where $d_C(v) = \inf\{|v-u| : u \in C\}$ for a non-empty closed subset of $\mathbb{R}^{d+1}$, and $\text{epi } G(h) = \{(h,\alpha) : G(h) \leqslant \alpha\}$ is the epigraph of $G : \mathbb{R}^d \mapsto \mathbb{R}$.

■

5.1.2 Proof of Theorem 2

Consider any sequence $\{P_n \in \mathcal{P} : n \geqslant 1\}$ that determines $\beta_n = \beta(P_n)$ and the laws of all random variables. If LICQ is satisfied and $l(\beta_0) = 0$, then $\lambda_{0j} = 0$ for all $j \in \mathcal{E} \cup \mathcal{I}$ so that Assumption 5 implies that uniformly over $h \in \mathcal{B}_{\delta_n}$,

$$\begin{aligned} n\hat{Q}_n\left(\beta_n + \frac{h}{\sqrt{n}}\right) - n\hat{Q}_n(\beta_n) &= h'\sqrt{n}\left(\hat{l}_n(\beta_n) - l(\beta_n)\right) + \frac{1}{2}h'H_0h + o_{P_n}(1) \\ &\rightsquigarrow h'W_0 + \frac{1}{2}h'H_0h \end{aligned}$$

as a process indexed by h in the space of bounded functions on compact sets $\ell^\infty(K)$ for any compact $K \subset \mathbb{R}^d$. These results in combination with the continuous mapping results in Lemma 10.11 of Kosorok (2007) imply that for $q(h) \equiv h'W_0 + \frac{1}{2}h'H_0h$,

$$\begin{aligned} & n\left(\hat{Q}_n(\beta_n) - \inf_{h\in\mathcal{B}_{\delta_n}}\hat{Q}_n\left(\beta_n + \frac{h}{\sqrt{n}}\right)\right) \\ &= -\inf_{h\in\mathcal{B}_{\delta_n}} n\left(\hat{Q}_n\left(\beta_n + \frac{h}{\sqrt{n}}\right) - \hat{Q}_n(\beta_n)\right) + o_{P_n}(1) \\ &= -\inf_{h\in\mathcal{B}_{\delta_n}}\left\{\sqrt{n}\left(\hat{l}_n(\beta_n) - l(\beta_n)\right)' h + \frac{1}{2}h'H_0h\right\} + o_{P_n}(1) \\ &\rightsquigarrow -\inf_{h\in\mathcal{B}_\kappa} q(h) \end{aligned}$$

where $\mathcal{B}_\kappa = \left\{h \in \mathbb{R}^d : \|h\| \leqslant \kappa\right\}$ for $\sqrt{n}\delta_n \to \kappa \in (0, \infty]$. We already showed in the proof of Theorem 1 that

$$\begin{aligned}\frac{\hat{Z}_n^*\left(\beta_n + \alpha_n h\right)}{\alpha_n^2} &= h'\sqrt{n}\left(\hat{l}_n^*\left(\bar{\beta}_n\right) - \hat{l}_n\left(\bar{\beta}_n\right)\right) + \frac{1}{2}h'\bar{H}_n h + o_{P_n}^*(1)\\ &\underset{\mathbb{W}}{\overset{\mathbb{P}}{\rightsquigarrow}} h'W_0 + \frac{1}{2}h'H_0 h\end{aligned}$$

Then the continuous mapping results in Lemma 10.11 of Kosorok (2007) imply

$$\begin{aligned}&-\frac{\inf_{\beta\in\mathbb{B}}\hat{Z}_n^*\left(\beta\right)}{\alpha_n^2}\\ &= -\frac{\inf_{h\in\frac{\mathbb{B}-\beta_n}{\alpha_n}}\hat{Z}_n^*\left(\beta_n + \alpha_n h\right)}{\alpha_n^2}\\ &= -\inf_{h\in\frac{\mathbb{B}-\beta_n}{\alpha_n}}\left\{\sqrt{n}\left(\hat{l}_n^*\left(\bar{\beta}_n\right) - \hat{l}_n\left(\bar{\beta}_n\right)\right)' h + \frac{1}{2}h'\bar{H}_n h\right\} + o_{P_n}(1)\\ &\underset{\mathbb{W}}{\overset{\mathbb{P}}{\rightsquigarrow}} -\inf_{h\in\mathbb{R}^d} q\left(h\right)\end{aligned}$$

Therefore, $\limsup_{n\to\infty}\sup_{P\in\mathcal{P}}\sup_{x\in\mathbb{R}}\left|J_n\left(x,P\right) - J\left(x,P\right)\right| = 0$, and since $\left\{J\left(\cdot,P\right) : P \in \mathcal{P}\right\}$ is equicontinuous at $J_n^{-1}\left(1-\alpha,P\right)$, we have for any P_n and ϵ small enough, $J_n\left(x_n,P_n\right) - J\left(x_n,P_n\right) = o(1)$ where $x_n = J_n^{-1}\left(1-\alpha-\epsilon,P_n\right)$. Similarly, $\limsup_{n\to\infty}\sup_{P\in\mathcal{P}} P\left(\sup_{x\in\mathbb{R}}\left|J_{\alpha_n}^*\left(x,P\right) - J^*\left(x,P\right)\right| > \epsilon\right) = 0$ for all $\epsilon > 0$, and since $\left\{J^*\left(\cdot,P\right) : P \in \mathcal{P}\right\}$ is equicontinuous at $J_n^{-1}\left(1-\alpha,P\right)$, for any P_n and ϵ small enough, $J_{\alpha_n}^*\left(x_n,P_n\right) - J^*\left(x_n,P_n\right) = o_{P_n}(1)$. Note that $-\inf_{h\in\left\{h\in\mathbb{R}^d:\|h\|\leqslant\kappa\right\}} q\left(h\right) \leqslant -\inf_{h\in\mathbb{R}^d} q\left(h\right)$ for any realizations of the random variables in the limiting distributions. Then, for all $\epsilon > 0$ and n large enough, there exists $\delta > 0$ such that $P_n\left(J_{\alpha_n}^*\left(x_n,P_n\right) - J_n\left(x_n,P_n\right) > \epsilon\right) \leqslant \delta$. If $J_{\alpha_n}^*\left(x_n,P_n\right) - J_n\left(x_n,P_n\right) \leqslant \epsilon$, then $J_n^{-1}\left(1-\alpha-\epsilon,P_n\right) \leqslant J_{\alpha_n}^{*-1}\left(1-\alpha,P_n\right)$. Take $\left\{\epsilon_n\right\}_{n=1}^\infty$ and $\left\{\delta_n\right\}_{n=1}^\infty$ to be positive sequences such that $\epsilon_n \to 0$ and $\delta_n \to 0$. Then, using arguments similar to those in Lemma A.1 (vi) of Romano and Shaikh (2012), for all $\epsilon > 0$ and n large enough,

$$\begin{aligned}&P_n\left(n\left(\hat{Q}_n\left(\beta_n\right) - \inf_{h\in\mathcal{B}_{\delta_n}}\hat{Q}_n\left(\beta_n + \frac{h}{\sqrt{n}}\right)\right) \leqslant J_{\alpha_n}^{*-1}\left(1-\alpha,P_n\right)\right)\\ &\geqslant P_n\left(n\left(\hat{Q}_n\left(\beta_n\right) - \inf_{h\in\mathcal{B}_{\delta_n}}\hat{Q}_n\left(\beta_n + \frac{h}{\sqrt{n}}\right)\right) \leqslant J_{\alpha_n}^{*-1}\left(1-\alpha,P_n\right) \cap J_{\alpha_n}^*\left(x_n,P_n\right) - J_n\left(x_n,P_n\right) \leqslant \epsilon\right)\\ &\geqslant P_n\left(n\left(\hat{Q}_n\left(\beta_n\right) - \inf_{h\in\mathcal{B}_{\delta_n}}\hat{Q}_n\left(\beta_n + \frac{h}{\sqrt{n}}\right)\right) \leqslant J_n^{-1}\left(1-\alpha-\epsilon,P_n\right) \cap J_{\alpha_n}^*\left(x_n,P_n\right) - J_n\left(x_n,P_n\right) \leqslant \epsilon\right)\end{aligned}$$

$$\geqslant P_n\left(n\left(\hat{Q}_n(\beta_n)-\inf_{h\in\mathcal{B}_{\delta_n}}\hat{Q}_n\left(\beta_n+\frac{h}{\sqrt{n}}\right)\right)\leqslant J_n^{-1}(1-\alpha-\epsilon,P_n)\right)-P_n\left(J^*_{\alpha_n}(x_n,P_n)-J_n(x_n,P_n)>\epsilon\right)$$

$$\geqslant 1-\alpha-\epsilon-\delta$$

Since ϵ and δ can be arbitrarily small, $\liminf_{n\to\infty} P_n\left(n\left(\hat{Q}_n(\beta_n)-\inf_{h\in\mathcal{B}_{\delta_n}}\hat{Q}_n\left(\beta_n+\frac{h}{\sqrt{n}}\right)\right)\leqslant \hat{c}^*_{1-\alpha}\right)\geqslant 1-\alpha$. For $\rho=\liminf_{n\to\infty}\inf_{P\in\mathcal{P}} P\left(n\left(\hat{Q}_n(\beta_n)-\inf_{h\in\mathcal{B}_{\delta_n}}\hat{Q}_n\left(\beta_n+\frac{h}{\sqrt{n}}\right)\right)\leqslant \hat{c}^*_{1-\alpha}\right)$, we can find a sequence $\{P_n\in\mathcal{P}\}$ such that $\rho=\liminf_{n\to\infty} P_n\left(n\left(\hat{Q}_n(\beta_n)-\inf_{h\in\mathcal{B}_{\delta_n}}\hat{Q}_n\left(\beta_n+\frac{h}{\sqrt{n}}\right)\right)\leqslant \hat{c}^*_{1-\alpha}\right)$. Find a subsequence n_k of n for which β_n converges, with its limit denoted β. The same arguments above applied to such a subsequence imply $\liminf_{n_k\to\infty} P_{n_k}\left(n_k\left(\hat{Q}_n(\beta_{n_k})-\inf_{h\in\mathcal{B}_{\delta_{n_k}}}\hat{Q}_n\left(\beta_{n_k}+\frac{h}{\sqrt{n_k}}\right)\right)\leqslant \hat{c}^*_{1-\alpha}\right)\geqslant 1-\alpha$. Since $\{P_{n_k},\beta_{n_k}\}$ is a subsequence of $\{P_n,\beta_n\}$,
$\rho=\liminf_{n_k\to\infty} P_{n_k}\left(n_k\left(\hat{Q}_n(\beta_{n_k})-\inf_{h\in\mathcal{B}_{\delta_{n_k}}}\hat{Q}_n\left(\beta_{n_k}+\frac{h}{\sqrt{n_k}}\right)\right)\leqslant \hat{c}^*_{1-\alpha}\right)\geqslant 1-\alpha$.

■

5.1.3 Proof of Theorem 3

Using similar arguments to Theorem 2.1 of Shapiro (1988) and Lemma 3.1 of Shapiro (1989), when $\hat{\beta}_n$ lies in a neighborhood of β_0, $\hat{\beta}_n$ is almost surely the minimizer of $\tilde{\mathcal{L}}_n(\beta)=\hat{Q}_n(\beta)+\sum_{j\in\mathcal{E}\cup\mathcal{I}}\lambda_{0j}f_{nj}(\beta)$ over $C(\lambda_0)=\left\{\beta\in\mathbb{B}:f_{nj}(\beta)=0\text{ for }j\in\mathcal{E}\cup\mathcal{I}^*_{n,+}(\lambda_0),f_{nj}(\beta)\leqslant 0\text{ for }j\in\mathcal{I}^*_{n,0}(\lambda_0)\right\}$, where $\mathcal{I}^*_{n,+}(\lambda_0)\equiv\{j\in\mathcal{I}^*_n:\lambda_{0j}>0\}$, $\mathcal{I}^*_{n,0}(\lambda_0)\equiv\{j\in\mathcal{I}^*_n:\lambda_{0j}=0\}$, and $\mathcal{I}^*_n\equiv\{j\in\mathcal{I}:f_{nj}(\beta_0)=0\}$. We assumed in condition (iii) that $\mathcal{I}^*_{n,+}(\lambda_0)=\varnothing$, which means $\mathcal{I}^*_n=\mathcal{I}^*_{n,0}(\lambda_0)$, and $C(\lambda_0)=\{\beta\in\mathbb{B}:f_{nj}(\beta)=0\text{ for }j\in\mathcal{E},f_{nj}(\beta)\leqslant 0\text{ for }j\in\mathcal{I}^*_n\}$.

Denote the feasible direction set by

$$\mathcal{F}_n=\left\{h:f_{nj}\left(\beta_0+\frac{h}{\sqrt{n}}\right)=0\text{ for }j\in\mathcal{E},f_{nj}\left(\beta_0+\frac{h}{\sqrt{n}}\right)\leqslant 0\text{ for }j\in\mathcal{I}^*_n\right\}$$

Denote the linearized feasible direction set by

$$\Sigma_n=\left\{h:\sqrt{n}f_{nj}(\beta_0)+F_{nj}(\beta_0)'h=0\text{ for }j\in\mathcal{E},\sqrt{n}f_{nj}(\beta_0)+F_{nj}(\beta_0)'h\leqslant 0\text{ for }j\in\mathcal{I}^*_n\right\}$$

LICQ implies the linearized feasible direction set is sufficient to capture the geometry of the constraints near β_0 so that $\sqrt{n}\left(\hat{\beta}_n-\beta_0\right)$ is asymptotically equivalent to the minimizer of the Lagrangian over Σ_n:

$$\begin{aligned}
\sqrt{n}\left(\hat{\beta}_n-\beta_0\right) &= \underset{h\in\Sigma_n}{\arg\min}\left\{n\tilde{\mathcal{L}}_n\left(\beta_0+\frac{h}{\sqrt{n}}\right)-n\tilde{\mathcal{L}}_n\left(\beta_0\right)\right\}+o_P(1)\\
&= \underset{h\in\Sigma_n}{\arg\min}\left\{n\hat{Q}_n\left(\beta_0+\frac{h}{\sqrt{n}}\right)-n\hat{Q}_n\left(\beta_0\right)+\sum_{j\in\mathcal{E}\cup\mathcal{I}}\lambda_{0j}n\left(f_{nj}\left(\beta_0+\frac{h}{\sqrt{n}}\right)-f_{nj}\left(\beta_0\right)\right)\right\}+o_P(1)\\
&\rightsquigarrow \underset{h\in\Sigma}{\arg\min}\left\{h'W_0+\frac{1}{2}h'H_0h+\sum_{j\in\mathcal{E}}\lambda_{0j}\left(h'V_{0j}+\frac{1}{2}h'G_{0j}h\right)\right\}=\mathcal{J}
\end{aligned}$$

where the convergence result follows from the following arguments. First note that Assumption 5′ implies that for any $\delta_n\to 0$, and $\mathcal{B}_{\delta_n}=\left\{h\in\mathbb{R}^d:\frac{\|h\|}{\sqrt{n}}\leqslant\delta_n\right\}$,

$$\sup_{h\in\mathcal{B}_{\delta_n}}\left|\frac{n\tilde{\mathcal{L}}_n\left(\beta_0+\frac{h}{\sqrt{n}}\right)-n\tilde{\mathcal{L}}_n\left(\beta_0\right)-h'\sqrt{n}\left(\hat{l}_n\left(\beta_0\right)-l\left(\beta_0\right)\right)-\frac{1}{2}h'H_0h-\sum_{j\in\mathcal{E}\cup\mathcal{I}}\lambda_{0j}\left(\sqrt{n}\left(F_{nj}\left(\beta_0\right)-F_{0j}\right)'h+\frac{1}{2}h'G_{0j}h\right)}{1+\|h\|^2}\right|$$
$$=o_P(1)$$

Therefore, uniformly in $h\in\mathcal{B}_{\delta_n}$,

$$\begin{aligned}
&n\hat{Q}_n\left(\beta_0+\frac{h}{\sqrt{n}}\right)-n\hat{Q}_n\left(\beta_0\right)+\sum_{j\in\mathcal{E}\cup\mathcal{I}}\lambda_{0j}n\left(f_{nj}\left(\beta_0+\frac{h}{\sqrt{n}}\right)-f_{nj}\left(\beta_0\right)\right)\\
&=h'\sqrt{n}\left(\hat{l}_n\left(\beta_0\right)-l\left(\beta_0\right)\right)+\frac{1}{2}h'H_0h+\sum_{j\in\mathcal{E}\cup\mathcal{I}}\lambda_{0j}\left(\sqrt{n}\left(F_{nj}\left(\beta_0\right)-F_{0j}\right)'h+\frac{1}{2}h'G_{0j}h\right)+o_P(1)
\end{aligned}$$

Recall $\sqrt{n}\left(\hat{l}_n\left(\beta_0\right)-l\left(\beta_0\right)\right)+\sum_{j\in\mathcal{E}\cup\mathcal{I}}\lambda_{0j}\sqrt{n}\left(F_{nj}\left(\beta_0\right)-F_{0j}\right)\rightsquigarrow W_0+\sum_{j\in\mathcal{E}\cup\mathcal{I}}\lambda_{0j}V_{0j}$, and $\lambda_{0j}=0$ for all $j\in\mathcal{I}\backslash\mathcal{I}_+^*\left(\lambda_0\right)$, where we have assumed $\mathcal{I}_+^*\left(\lambda_0\right)=\varnothing$. and $\lambda_{0j}=0$ for all $j\in\mathcal{I}\backslash\mathcal{I}_+^*\left(\lambda_0\right)$. Since the last line is a convex function of h, pointwise convergence implies uniform convergence over compact sets $K\subset\mathbb{R}^d$ (Pollard (1991)). Therefore,

$$\begin{aligned}
&h'\sqrt{n}\left(\hat{l}_n\left(\beta_0\right)-l\left(\beta_0\right)\right)+\frac{1}{2}h'H_0h+\sum_{j\in\mathcal{E}\cup\mathcal{I}}\lambda_{0j}\left(\sqrt{n}\left(F_{nj}\left(\beta_0\right)-F_{0j}\right)'h+\frac{1}{2}h'G_{0j}h\right)+o_P(1)\\
&\rightsquigarrow h'W_0+\frac{1}{2}h'H_0h+\sum_{j\in\mathcal{E}\cup\mathcal{I}}\lambda_{0j}\left(h'V_{0j}+\frac{1}{2}h'G_{0j}h\right)\\
&=h'W_0+\frac{1}{2}h'H_0h+\sum_{j\in\mathcal{E}}\lambda_{0j}\left(h'V_{0j}+\frac{1}{2}h'G_{0j}h\right)
\end{aligned}$$

as a process indexed by h in the space of bounded functions on compact sets $\ell^\infty(K)$ for any compact

$K \subset \mathbb{R}^d$.

Now consider the constraints. $\sqrt{n} f_{nj}(\beta_0) + F_{nj}(\beta_0)' h \xrightarrow{p} -\infty$ for $j \in \mathcal{I} \backslash \mathcal{I}^*$, so the nonactive inequality constraints do not affect the asymptotic distribution. Additionally, $\sqrt{n} f_{nj}(\beta_0) \rightsquigarrow U_{0j}$, jointly, for all $j \in \mathcal{E} \cup \mathcal{I}^*$, and $F_{nj}(\beta_0) = F_{0j} + o_P(1)$. Condition (i) is a second order sufficient condition and guarantees that the argmin of $h'W_0 + \frac{1}{2} h' H_0 h + \sum_{j \in \mathcal{E}} \lambda_{0j} \left(h' V_{0j} + \frac{1}{2} h' G_{0j} h \right)$ over Σ is unique. Then by the argmin continuous mapping theorem (Theorem 1 of Knight (1999)), $\arg\min_h \hat{\mathbb{G}}_n(h) \to_{e-d} \arg\min_h \mathbb{G}_0(h)$, where

$$\hat{\mathbb{G}}_n(h) = n \hat{Q}_n\left(\beta_0 + \frac{h}{\sqrt{n}}\right) - n \hat{Q}_n(\beta_0) + \sum_{j \in \mathcal{E} \cup \mathcal{I}} \lambda_{0j} n \left(f_{nj}\left(\beta_0 + \frac{h}{\sqrt{n}}\right) - f_{nj}(\beta_0) \right) + \infty 1 (h \notin \Sigma_n)$$

$$\mathbb{G}_0(h) = h' W_0 + \frac{1}{2} h' H_0 h + \sum_{j \in \mathcal{E}} \lambda_{0j} \left(h' V_{0j} + \frac{1}{2} h' G_{0j} h \right) + \infty 1 (h \notin \Sigma)$$

$$\Sigma = \left\{ h : U_{0j} + F_{0j}' h = 0 \text{ for } j \in \mathcal{E}, U_{0j} + F_{0j}' h \leqslant 0 \text{ for } j \in \mathcal{I}^* \right\}$$

Note that since C^* is already a linearized constraint set, the linearized feasible direction set is simply

$$\begin{aligned}
\Sigma_n^* &= \{ h : f_{nj}(\bar{\beta}_n) + \bar{F}_{nj}' (\beta_0 - \bar{\beta}_n + \alpha_n h) + \alpha_n \sqrt{n} \left(f_{nj}^*(\bar{\beta}_n) - f_{nj}(\bar{\beta}_n) \right) = 0 \text{ for } j \in \mathcal{E} \\
&\qquad f_{nj}(\bar{\beta}_n) + \bar{F}_{nj}' (\beta_0 - \bar{\beta}_n + \alpha_n h) + \alpha_n \sqrt{n} \left(f_{nj}^*(\bar{\beta}_n) - f_{nj}(\bar{\beta}_n) \right) \leqslant 0 \text{ for } j \in \mathcal{I} \} \\
&= \left\{ h : \frac{f_{nj}(\bar{\beta}_n)}{\alpha_n} + \bar{F}_{nj}' h + \sqrt{n} \left(f_{nj}^*(\bar{\beta}_n) - f_{nj}(\bar{\beta}_n) \right) + \bar{F}_{nj}' \left(\frac{\beta_0 - \bar{\beta}_n}{\alpha_n} \right) = 0 \text{ for } j \in \mathcal{E}, \right. \\
&\qquad \left. \frac{f_{nj}(\bar{\beta}_n)}{\alpha_n} + \bar{F}_{nj}' h + \sqrt{n} \left(f_{nj}^*(\bar{\beta}_n) - f_{nj}(\bar{\beta}_n) \right) + \bar{F}_{nj}' \left(\frac{\beta_0 - \bar{\beta}_n}{\alpha_n} \right) \leqslant 0 \text{ for } j \in \mathcal{I} \right\}
\end{aligned}$$

Using the local parameter $h \in \frac{C^* - \beta_0}{\alpha_n}$, we can derive the asymptotic distribution of the proximal bootstrap.

$$\begin{aligned}
\frac{\hat{\beta}_n^* - \beta_0}{\alpha_n} &= \arg\min_{h \in \Sigma_n^*} \left\{ \alpha_n \sqrt{n} \left(\hat{l}_n^*(\bar{\beta}_n) - \hat{l}_n(\bar{\beta}_n) \right)' (\beta_0 - \bar{\beta}_n + \alpha_n h) + \frac{1}{2} \left\| \beta_0 - \bar{\beta}_n + \alpha_n h \right\|_{\bar{H}_n}^2 \right. \\
&\qquad \left. + \sum_{j \in \mathcal{E} \cup \mathcal{I}} \bar{\lambda}_{nj} \left(\alpha_n \sqrt{n} \left(\bar{F}_{nj}^* - \bar{F}_{nj} \right)' (\beta_0 - \bar{\beta}_n + \alpha_n h) + \frac{1}{2} \left\| \beta_0 - \bar{\beta}_n + \alpha_n h \right\|_{\bar{G}_{nj}}^2 \right) \right\} \\
&= \arg\min_{h \in \Sigma_n^*} \left\{ \sqrt{n} \left(\hat{l}_n^*(\bar{\beta}_n) - \hat{l}_n(\bar{\beta}_n) \right)' \left(\frac{\beta_0 - \bar{\beta}_n}{\alpha_n} + h \right) + \frac{1}{2} \left\| \frac{\beta_0 - \bar{\beta}_n}{\alpha_n} + h \right\|_{\bar{H}_n}^2 \right\}
\end{aligned}$$

$$
\begin{aligned}
&+\sum_{j\in\mathcal{E}\cup\mathcal{I}}\bar{\lambda}_{nj}\left(\sqrt{n}\left(\bar{F}^*_{nj}-\bar{F}_{nj}\right)'\left(\frac{\beta_0-\bar{\beta}_n}{\alpha_n}+h\right)+\frac{1}{2}\left\|\frac{\beta_0-\bar{\beta}_n}{\alpha_n}+h\right\|^2_{\bar{G}_{nj}}\right)\Bigg\}\\
=\underset{h\in\Sigma^*_n}{\arg\min}&\left\{h'\sqrt{n}\left(\hat{l}^*_n\left(\bar{\beta}_n\right)-\hat{l}_n\left(\bar{\beta}_n\right)\right)+\frac{1}{2}h'\bar{H}_nh\right.\\
&\left.+\sum_{j\in\mathcal{E}\cup\mathcal{I}}\bar{\lambda}_{nj}\left(h'\sqrt{n}\left(\bar{F}^*_{nj}-\bar{F}_{nj}\right)+\frac{1}{2}h'\bar{G}_{nj}h\right)\right\}+o^*_P(1)\\
\overset{\mathbb{P}}{\underset{\mathbb{W}}{\rightsquigarrow}}\underset{h\in\Sigma}{\arg\min}&\left\{h'W_0+\frac{1}{2}h'H_0h+\sum_{j\in\mathcal{E}}\lambda_{0j}\left(h'V_{0j}+\frac{1}{2}h'G_{0j}h\right)\right\}=\mathcal{J}
\end{aligned}
$$

where the last line follows from the following arguments. First, note that since $\bar{H}_n\xrightarrow{p}H_0$, $\bar{G}_{nj}\xrightarrow{p}G_{0j}$ for all j, and the proximal bootstrap Lagrangian is convex in h, we have that uniformly over compact sets $K\subset\mathbb{R}^d$,

$$
\begin{aligned}
&h'\sqrt{n}\left(\hat{l}^*_n\left(\bar{\beta}_n\right)-\hat{l}_n\left(\bar{\beta}_n\right)\right)+\frac{1}{2}h'\bar{H}_nh+\sum_{j\in\mathcal{E}\cup\mathcal{I}}\bar{\lambda}_{nj}\left(h'\sqrt{n}\left(\bar{F}^*_{nj}-\bar{F}_{nj}\right)+\frac{1}{2}h'\bar{G}_{nj}h\right)\\
=&h'\sqrt{n}\left(\hat{l}^*_n\left(\bar{\beta}_n\right)-\hat{l}_n\left(\bar{\beta}_n\right)\right)+\frac{1}{2}h'H_0h+\sum_{j\in\mathcal{E}\cup\mathcal{I}}\bar{\lambda}_{nj}\left(h'\sqrt{n}\left(\bar{F}^*_{nj}-\bar{F}_{nj}\right)+\frac{1}{2}h'G_{0j}h\right)+o_P(1)
\end{aligned}
$$

Next, note that Assumption 3, $\max_{j\in\mathcal{E}\cup\mathcal{I}}\left|\bar{\lambda}_{nj}-\lambda_{0j}\right|\xrightarrow{p}0$, and $\sup_{\|\beta-\beta_0\|\leqslant o(1)}\sqrt{n}\left(F^*_n\left(\beta\right)-F_n\left(\beta\right)-F^*_n\left(\beta_0\right)+F_n\left(\beta_0\right)\right)=o^*_P(1)$ imply $\sqrt{n}\left(\hat{l}^*_n\left(\bar{\beta}_n\right)-\hat{l}_n\left(\bar{\beta}_n\right)\right)+\sum_{j\in\mathcal{E}\cup\mathcal{I}}\bar{\lambda}_{nj}\sqrt{n}\left(\bar{F}^*_{nj}-\bar{F}_{nj}\right)\overset{\mathbb{P}}{\underset{\mathbb{W}}{\rightsquigarrow}}W_0+\sum_{j\in\mathcal{E}\cup\mathcal{I}}\lambda_{0j}V_{0j}$ because

$$
\begin{aligned}
&\sqrt{n}\left(\hat{l}^*_n\left(\bar{\beta}_n\right)-\hat{l}_n\left(\bar{\beta}_n\right)\right)+\sum_{j\in\mathcal{E}\cup\mathcal{I}}\bar{\lambda}_{nj}\sqrt{n}\left(\bar{F}^*_{nj}-\bar{F}_{nj}\right)\\
&=\sqrt{n}\left(\hat{l}^*_n\left(\beta_0\right)-\hat{l}_n\left(\beta_0\right)\right)+\sqrt{n}\left(\hat{l}^*_n\left(\bar{\beta}_n\right)-\hat{l}_n\left(\bar{\beta}_n\right)-\left(\hat{l}^*_n\left(\beta_0\right)-\hat{l}_n\left(\beta_0\right)\right)\right)\\
&+\sum_{j\in\mathcal{E}\cup\mathcal{I}}\lambda_{0j}\sqrt{n}\left(F^*_{nj}\left(\beta_0\right)-F_{nj}\left(\beta_0\right)\right)+\sum_{j\in\mathcal{E}\cup\mathcal{I}}\left(\bar{\lambda}_{nj}-\lambda_{0j}\right)\sqrt{n}\left(\bar{F}^*_{nj}-\bar{F}_{nj}\right)\\
&+\sum_{j\in\mathcal{E}\cup\mathcal{I}}\lambda_{0j}\sqrt{n}\left(\bar{F}^*_{nj}-\bar{F}_{nj}-\left(F^*_{nj}\left(\beta_0\right)-F_{nj}\left(\beta_0\right)\right)\right)\\
&=\sqrt{n}\left(\hat{l}^*_n\left(\beta_0\right)-\hat{l}_n\left(\beta_0\right)\right)+\sum_{j\in\mathcal{E}\cup\mathcal{I}}\lambda_{0j}\sqrt{n}\left(F^*_{nj}\left(\beta_0\right)-F_{nj}\left(\beta_0\right)\right)+o^*_P(1)
\end{aligned}
$$

and we assumed $\sqrt{n}\left(\hat{l}^*_n\left(\beta_0\right)-\hat{l}_n\left(\beta_0\right)\right)+\sum_{j\in\mathcal{E}\cup\mathcal{I}}\lambda_{0j}\sqrt{n}\left(F^*_{nj}\left(\beta_0\right)-F_{nj}\left(\beta_0\right)\right)\overset{\mathbb{P}}{\underset{\mathbb{W}}{\rightsquigarrow}}W_0+\sum_{j\in\mathcal{E}\cup\mathcal{I}}\lambda_{0j}V_{0j}$. Additionally, $\max_{j\in\mathcal{E}\cup\mathcal{I}}\left|\bar{G}_{nj}-G_{0j}\right|\xrightarrow{p}0$ and $\max_{j\in\mathcal{E}\cup\mathcal{I}}\left|\bar{\lambda}_{nj}-\lambda_{0j}\right|\xrightarrow{p}0$ imply that $\sum_{j\in\mathcal{E}\cup\mathcal{I}}\bar{\lambda}_{nj}\bar{G}_{nj}\xrightarrow{p}\sum_{j\in\mathcal{E}\cup\mathcal{I}}\lambda_{0j}G_{0j}$. By convexity of the bootstrap Lagrangian in h, pointwise convergence implies uniform convergence over compact sets $K \subset \mathbb{R}^d$; therefore,

$$
\begin{aligned}
& h'\sqrt{n}\left(\hat{l}_n^*\left(\bar{\beta}_n\right)-\hat{l}_n\left(\bar{\beta}_n\right)\right)+\frac{1}{2}h'H_0h+\sum_{j\in\mathcal{E}\cup\mathcal{I}}\bar{\lambda}_{nj}\left(h'\sqrt{n}\left(\bar{F}_{nj}^*-\bar{F}_{nj}\right)+\frac{1}{2}h'G_{0j}h\right)\\
& \overset{\mathbb{P}}{\underset{\mathbb{W}}{\rightsquigarrow}} h'W_0+\frac{1}{2}h'H_0h+\sum_{j\in\mathcal{E}\cup\mathcal{I}}\lambda_{0j}\left(h'V_{0j}+\frac{1}{2}h'G_{0j}h\right)\\
& = h'W_0+\frac{1}{2}h'H_0h+\sum_{j\in\mathcal{E}}\lambda_{0j}\left(h'V_{0j}+\frac{1}{2}h'G_{0j}h\right)
\end{aligned}
$$

as a process indexed by h in the space of bounded functions on compact sets $\ell^\infty(K)$ for any compact $K \subset \mathbb{R}^d$.

Note that $\frac{f_{nj}\left(\bar{\beta}_n\right)}{\alpha_n}\xrightarrow{p}-\infty$ for $j \in \mathcal{I}\backslash\mathcal{I}^*$ while $\frac{f_{nj}\left(\bar{\beta}_n\right)}{\alpha_n}=\frac{\sqrt{n}\left(f_{nj}\left(\bar{\beta}_n\right)-f_{0j}(\beta_0)\right)}{\sqrt{n}\alpha_n}=\frac{\sqrt{n}\left(f_{nj}\left(\bar{\beta}_n\right)-f_{nj}(\beta_0)\right)}{\sqrt{n}\alpha_n}+\frac{\sqrt{n}(f_{nj}(\beta_0)-f_{0j}(\beta_0))}{\sqrt{n}\alpha_n}=o_P(1)$ for $j \in \mathcal{E}\cup\mathcal{I}^*$. Additionally, $\bar{F}_{nj}'\left(\frac{\beta_0-\bar{\beta}_n}{\alpha_n}\right)=o_P(1)$ and $\bar{F}_{nj}=F_{0j}+o_P(1)$ for all $j \in \mathcal{E}\cup\mathcal{I}$. Since $\sqrt{n}\left(f_n^*\left(\beta_0\right)-f_n\left(\beta_0\right)\right)\overset{\mathbb{P}}{\underset{\mathbb{W}}{\rightsquigarrow}}U_0$ and $\sup_{\|\beta-\beta_0\|\leqslant o(1)}\sqrt{n}\left(f_n^*\left(\beta\right)-f_n\left(\beta\right)-f_n^*\left(\beta_0\right)+f_n\left(\beta_0\right)\right)=o_P^*(1)$, it follows that $\sqrt{n}\left(f_n^*\left(\bar{\beta}_n\right)-f_n\left(\bar{\beta}_n\right)\right)\overset{\mathbb{P}}{\underset{\mathbb{W}}{\rightsquigarrow}}U_0$. Then, by the bootstrap argmin continuous mapping lemma 14.2 in Hong and Li (2020) (after replacing weak convergence with epi-convergence), $\arg\min_h \hat{\mathbb{G}}_n^*(h)\underset{e-d}{\xrightarrow{p}}\arg\min_h \mathbb{G}_0(h)$ for

$$
\begin{aligned}
\hat{\mathbb{G}}_n^*(h) &= h'\sqrt{n}\left(\hat{l}_n^*\left(\bar{\beta}_n\right)-\hat{l}_n\left(\bar{\beta}_n\right)\right)+\frac{1}{2}h'\bar{H}_nh\\
&\quad+\sum_{j\in\mathcal{E}\cup\mathcal{I}}\bar{\lambda}_{nj}\left(h'\sqrt{n}\left(\bar{F}_{nj}^*-\bar{F}_{nj}\right)+\frac{1}{2}h'\bar{G}_{nj}h\right)+\infty 1\left(h\notin\Sigma_n^*\right)\\
\mathbb{G}_0(h) &= h'W_0+\frac{1}{2}h'H_0h+\sum_{j\in\mathcal{E}\cup\mathcal{I}_+^*(\lambda_0)}\lambda_{0j}\left(h'V_{0j}+\frac{1}{2}h'G_{0j}h\right)+\infty 1\left(h\notin\Sigma\right)\\
\Sigma &= \left\{h: U_{0j}+F_{0j}'h=0 \text{ for } j\in\mathcal{E}, U_{0j}+F_{0j}'h\leqslant 0 \text{ for } j\in\mathcal{I}^*\right\}
\end{aligned}
$$

■

5.2 Verification of Assumptions

We first verify that Assumptions 2 and 3 are satisfied for the boundary constrained GMM example (example 2). In this example, $\hat{l}_n\left(\bar{\beta}_n\right)=\hat{G}_n'\hat{\pi}_n\left(\bar{\beta}_n\right)$, $\hat{l}_n^*\left(\bar{\beta}_n\right)=\hat{G}_n^{*\prime}\hat{\pi}_n^*\left(\bar{\beta}_n\right)$, $\hat{\pi}_n\left(\beta\right)=$

$\left[\frac{1}{n}\sum_{i=1}^{n} 1\left(y_i \leqslant \beta\right) - 0.5, \frac{1}{n}\sum_{i=1}^{n} y_i - \beta\right]'$, $\hat{\pi}_n^*\left(\beta\right) = \left[\frac{1}{n}\sum_{i=1}^{n} 1\left(y_i^* \leqslant \beta\right) - 0.5, \frac{1}{n}\sum_{i=1}^{n} y_i^* - \beta\right]'$, and

$$\hat{G}_n = \left[\begin{array}{c} \frac{1}{nh}\sum_{i=1}^{n} K_h\left(y_i - \hat{\beta}_n\right) \\ -1 \end{array}\right], \hat{G}_n^* = \left[\begin{array}{c} \frac{1}{nh}\sum_{i=1}^{n} K_h\left(y_i^* - \hat{\beta}_n\right) \\ -1 \end{array}\right], G = \left[\begin{array}{c} f(\beta_0) \\ -1 \end{array}\right],$$

where $f\left(\cdot\right)$ is the density of y and $K_h\left(x\right) = K\left(x/h\right)$ for some kernel function $K\left(\cdot\right)$ and bandwidth h. We can express $\sqrt{n}\left(\hat{l}_n^*\left(\bar{\beta}_n\right) - \hat{l}_n\left(\bar{\beta}_n\right)\right)$ as

$$\begin{aligned}\sqrt{n}\left(\hat{l}_n^*\left(\bar{\beta}_n\right) - \hat{l}_n\left(\bar{\beta}_n\right)\right) &= \sqrt{n}\left(\hat{G}_n^{*\prime}\hat{\pi}_n^*\left(\bar{\beta}_n\right) - \hat{G}_n'\hat{\pi}_n\left(\bar{\beta}_n\right)\right) \\ &= G'\sqrt{n}\left(\hat{\pi}_n^*\left(\bar{\beta}_n\right) - \hat{\pi}_n\left(\bar{\beta}_n\right)\right) + \left(\hat{G}_n^* - G\right)'\sqrt{n}\left(\hat{\pi}_n^*\left(\bar{\beta}_n\right) - \hat{\pi}_n^*\left(\beta_0\right)\right) \\ &- \left(\hat{G}_n - G\right)'\sqrt{n}\left(\hat{\pi}_n\left(\bar{\beta}_n\right) - \hat{\pi}_n\left(\beta_0\right)\right) + \left(\hat{G}_n^* - G\right)'\sqrt{n}\left(\hat{\pi}_n^*\left(\beta_0\right) - \hat{\pi}_n\left(\beta_0\right)\right) \\ &+ \left(\hat{G}_n^* - \hat{G}_n\right)'\sqrt{n}\left(\hat{\pi}_n\left(\beta_0\right) - \pi\left(\beta_0\right)\right) \\ &= G'\sqrt{n}\left(P_n^* - P_n\right)\pi\left(\cdot,\beta_0\right) + G'\sqrt{n}\left(P_n^* - P_n\right)\left(\pi\left(\cdot,\bar{\beta}_n\right) - \pi\left(\cdot,\beta_0\right)\right) + o_p(1)\end{aligned}$$

where we have used $\sqrt{n}\left(\hat{\pi}_n^*\left(\bar{\beta}_n\right) - \hat{\pi}_n^*\left(\beta_0\right)\right) = O_p(1)$, $\sqrt{n}\left(\hat{\pi}_n\left(\bar{\beta}_n\right) - \hat{\pi}_n\left(\beta_0\right)\right) = O_p(1)$, $\sqrt{n}\left(\hat{\pi}_n^*\left(\beta_0\right) - \hat{\pi}_n\left(\beta_0\right)\right) = O_p(1)$, $\sqrt{n}\left(\hat{\pi}_n\left(\beta_0\right) - \pi\left(\beta_0\right)\right) = O_p(1)$, $\hat{G}_n^* - G = o_p(1)$, and $\hat{G}_n^* - G = o_p(1)$. We can express $G'\left(\pi\left(\cdot,\bar{\beta}_n\right) - \pi\left(\cdot,\beta_0\right)\right) = f\left(\beta_0\right)\left(1\left(y_i \leqslant \bar{\beta}_n\right) - 1\left(y_i \leqslant \beta_0\right)\right) + \left(\bar{\beta}_n - \beta_0\right) = g\left(\cdot,\bar{\beta}_n\right) - g\left(\cdot,\beta_0\right)$ for $g\left(\cdot,\beta\right) = f\left(\beta_0\right)\left(1\left(y_i \leqslant \beta\right) - \tau\right) - \left(y_i - \beta\right)$. Note that $\mathcal{G}_R \equiv \left\{g\left(\cdot,\beta\right) - g\left(\cdot,\beta_0\right) : \left|\beta - \beta_0\right| \leqslant R\right\}$ is a Donsker class for some $R > 0$ because $\left\{1\left(y_i \leqslant \beta\right) : \left|\beta - \beta_0\right| \leqslant R\right\}$ and $\left\{1\left(y_i \leqslant \beta_0\right)\right\}$ are bounded Donsker classes, $f\left(\beta_0\right)$ is bounded between 0 and 1, and $\beta - \beta_0$ is bounded between $-R$ and R on $\mathcal{G}_R$. Using the Donsker preservation properties for sums and products of bounded Donsker classes, $\mathcal{G}_R$ is a Donsker class. Additionally, $P\left|g\left(\cdot,\beta\right) - g\left(\cdot,\beta_0\right)\right|^2 \to 0$ as $\beta \to \beta_0$ because

$$\begin{aligned}&P\left|g\left(\cdot,\beta\right) - g\left(\cdot,\beta_0\right)\right|^2 \\ \leqslant& f\left(\beta_0\right)^2 E\left|1\left(y_i \leqslant \beta\right) - 1\left(y_i \leqslant \beta_0\right)\right|^2 + \left|\beta - \beta_0\right|^2 + 2E\left|1\left(y_i \leqslant \beta\right) - 1\left(y_i \leqslant \beta_0\right)\right|\left|\beta - \beta_0\right| \\ =& f\left(\beta_0\right)^2 E\left|1\left(y_i \leqslant \beta\right) - 1\left(y_i \leqslant \beta_0\right)\right| + \left|\beta - \beta_0\right|^2 + 2E\left|1\left(y_i \leqslant \beta\right) - 1\left(y_i \leqslant \beta_0\right)\right|\left|\beta - \beta_0\right| \\ \leqslant& f\left(\beta_0\right)^2\left(E\left|1\left(\beta_0 \leqslant y_i \leqslant \beta\right)\right| + E\left|1\left(\beta \leqslant y_i \leqslant \beta_0\right)\right|\right) + 2\left|\beta - \beta_0\right| \\ =& f\left(\beta_0\right)^2\left(P\left(\beta_0 \leqslant y_i \leqslant \beta\right) + P\left(\beta \leqslant y_i \leqslant \beta_0\right)\right) + 2\left|\beta - \beta_0\right|\end{aligned}$$

The envelope integrability condition in Assumption 3 is satisfied because $f\left(\beta_0\right)\left(1\left(y_i \leqslant \bar{\beta}_n\right)-1\left(y_i \leqslant \beta_0\right)\right)$ is bounded between -1 and 1, which implies $\sup\limits_{g(\cdot,\beta)\in\mathcal{G}_{\delta_n}} \left|\frac{g(\cdot,\beta)-g(\cdot,\beta_0)}{1+\sqrt{n}\|\beta-\beta_0\|}\right| \leqslant 1$ for any $n>1$. Therefore, $G'\sqrt{n}\left(P_n^*-P_n\right)\left(\pi\left(\cdot,\bar{\beta}_n\right)-\pi\left(\cdot,\beta_0\right)\right) = o_p(1)$, and $\sqrt{n}\left(\hat{l}_n^*\left(\bar{\beta}_n\right)-\hat{l}_n\left(\bar{\beta}_n\right)\right)$ converges to the same asymptotic distribution as $\sqrt{n}\left(\hat{l}_n\left(\beta_0\right)-l\left(\beta_0\right)\right)$.

We can also check Assumptions 2 and 3 are satisfied for the conditional logit example (example 3) under additional assumptions. In that example, $\hat{l}_n\left(\beta\right) = \frac{1}{nJ}\sum_{i=1}^n\sum_{j=1}^J\left(y_{ij}-P_{ij}\left(\beta\right)\right)x_{ij} = P_n g\left(\cdot,\beta\right)$ and $\hat{l}_n^*\left(\beta\right) = \frac{1}{nJ}\sum_{i=1}^n\sum_{j=1}^J\left(y_{ij}^*-P_{ij}^*\left(\beta\right)\right)x_{ij}^* = P_n^* g\left(\cdot,\beta\right)$ for $g\left(\cdot,\beta\right) = \frac{1}{J}\sum_{j=1}^J\left(y_{ij}-P_{ij}\left(\beta\right)\right)x_{ij}$ and $P_{ij}\left(\beta\right) \equiv \frac{\exp(\beta x_{ij})}{\sum_l \exp(\beta x_{il})}$, where $J \ll n$ is fixed. If $E\left|x_{ij}\right|^4 < \infty$, we can show that $P\left|g\left(\cdot,\beta\right)-g\left(\cdot,\beta_0\right)\right|^2 \to 0$ as $\beta \to \beta_0$ because

$$\begin{aligned} P\left|g\left(\cdot,\beta\right)-g\left(\cdot,\beta_0\right)\right|^2 \leqslant & \frac{1}{J}\sum_{j=1}^J E\left|P_{ij}\left(\beta_0\right)-P_{ij}\left(\beta\right)\right|^2\left|x_{ij}\right|^2 \\ \leqslant & \frac{1}{J}\sum_{j=1}^J \sqrt{E\left|P_{ij}\left(\beta_0\right)-P_{ij}\left(\beta\right)\right|^4 E\left|x_{ij}\right|^4} \end{aligned}$$

If $E\left(\sup\limits_b \frac{1}{J}\sum_{j=1}^J\left|P_{ij}'\left(b\right)x_{ij}\right|\right)^2 < \infty$, $\mathcal{G}_R \equiv \left\{g\left(\cdot,\beta\right)-g\left(\cdot,\beta_0\right) : \left|\beta-\beta_0\right| \leqslant R\right\}$ is a Donsker class for some $R>0$ because $\left|g\left(\cdot,\beta\right)-g\left(\cdot,\beta_0\right)\right| = \left|\frac{1}{J}\sum_{j=1}^J\left(P_{ij}\left(\beta_0\right)-P_{ij}\left(\beta\right)\right)x_{ij}\right| \leqslant \sup\limits_b \frac{1}{J}\sum_{j=1}^J\left|P_{ij}'\left(b\right)x_{ij}\right|\left|\beta-\beta_0\right|$ is Lipschitz with a square-integrable Lipschitz constant. The envelope integrability condition will be satisfied if the envelope function for $\mathcal{G}_R$ is uniformly integrable or if the x_{ij} are uniformly bounded.

References

Amemiya, T. (1985): *Advanced Econometrics*, Harvard University Press. 17

Andrews, D. W. (1999): “Estimation when a parameter is on a boundary,” *Econometrica*, 67, 1341–1383. 3

——— (2000): “Inconsistency of the bootstrap when a parameter is on the boundary of the parameter space,” *Econometrica*, 68, 399–405. 2, 3, 5, 9, 13, 18

——— (2002a): “Generalized method of moments estimation when a parameter is on a boundary,” *Journal of Business & Economic Statistics*, 20, 530–544. 3

——— (2002b): "Higher-Order Improvements of a Computationally Attractive k-Step Bootstrap for Extremum Estimators," *Econometrica*, 70, 119–162. 3

Andrews, D. W. and P. Guggenberger (2009): "Validity of subsampling and "plug-in asymptotic" inference for parameters defined by moment inequalities," *Econometric Theory*, 25, 669–709. 2

——— (2010): "Asymptotic size and a problem with subsampling and with the m out of n bootstrap," *Econometric Theory*, 26, 426–468. 2

Andrews, D. W. and S. Han (2009): "Invalidity of the bootstrap and the m out of n bootstrap for confidence interval endpoints defined by moment inequalities," *The Econometrics Journal*, 12, S172–S199. 2

Andrews, D. W. and G. Soares (2010): "Inference for parameters defined by moment inequalities using generalized moment selection," *Econometrica*, 78, 119–157. 2

Armstrong, T. B., M. Bertanha, and H. Hong (2014): "A fast resample method for parametric and semiparametric models," *Journal of Econometrics*, 179, 128–133. 2

Bugni, F. A. (2010): "Bootstrap inference in partially identified models defined by moment inequalities: Coverage of the identified set," *Econometrica*, 78, 735–753. 2

Canay, I. A. (2010): "EL inference for partially identified models: Large deviations optimality and bootstrap validity," *Journal of Econometrics*, 156, 408–425. 2

Chakraborty, B., E. B. Laber, and Y. Zhao (2013): "Inference for optimal dynamic treatment regimes using an adaptive m-out-of-n bootstrap scheme," *Biometrics*, 69, 714–723. 18

Chen, X., T. M. Christensen, and E. Tamer (2018): "Monte Carlo confidence sets for identified sets," *Econometrica*, 86, 1965–2018. 4

Chernozhukov, V. and H. Hong (2003): "A MCMC Approach to Classical Estimation," *Journal of Econometrics*, 115, 293–346. 14

Chernozhukov, V., H. Hong, and E. Tamer (2007): "Estimation and confidence regions for parameter sets in econometric models," *Econometrica*, 75, 1243–1284. 2

Chernozhukov, V., W. K. Newey, and A. Santos (2023): "Constrained Conditional Moment Restriction Models," *Econometrica*, 91, 709–736. 4

Davidson, R. and J. G. MacKinnon (1999): "Bootstrap testing in nonlinear models," *International Economic Review*, 40, 487–508. 3

de Palma, A., N. Picard, and P. Waddell (2007): "Discrete choice models with capacity constraints: An empirical analysis of the housing market of the greater Paris region," *Journal of Urban Economics*, 62, 204–230. 6, 27

Fang, Z. and A. Santos (2019): "Inference on directionally differentiable functions," *The Review of Economic Studies*, 86, 377–412. 4, 23, 24, 26, 27

Fang, Z. and J. Seo (2021): "A projection framework for testing shape restrictions that form convex cones," *Econometrica*, 89, 2439–2458. 4

Forneron, J.-J. and S. Ng (2019): "Estimation and Inference by Stochastic Optimization," *arXiv preprint arXiv:2004.09627*. 2

Gafarov, B. (2016): "Inference on scalar parameters in set–identified affine models," Tech. rep., Mimeo: UC Davis. 4

Geyer, C. J. (1994): "On the asymptotics of constrained M-estimation," *The Annals of Statistics*, 1993–2010. 3, 17

Hall, P. and M. A. Martin (1988): "On bootstrap resampling and iteration," *Biometrika*, 75, 661–671. 19

Hong, H. and J. Li (2020): "The numerical bootstrap," *The Annals of Statistics*, 48, 397–412. 3, 34, 41

Horowitz, J. L. and S. Lee (2019): "Non-asymptotic inference in a class of optimization problems," . 4

Hsieh, Y.-W., X. Shi, and M. Shum (2022): "Inference on estimators defined by mathematical programming," *Journal of Econometrics*, 226, 248–268. 4, 24, 26, 27

Kaido, H. (2016): "A dual approach to inference for partially identified econometric models," *Journal of Econometrics*, 192, 269–290. 4

Kaido, H., F. Molinari, and J. Stoye (2019): "Confidence intervals for projections of partially identified parameters," *Econometrica*, 87, 1397–1432. 4

——— (2021): "Constraint qualifications in partial identification," *Econometric Theory*, 1–24. 4, 13

Kaido, H. and A. Santos (2014): "Asymptotically efficient estimation of models defined by convex moment inequalities," *Econometrica*, 82, 387–413. 4

Kline, P. and A. Santos (2012): "A score based approach to wild bootstrap inference," *Journal of Econometric Methods*, 1, 23–41. 2, 3

Knight, K. (1999): "Epi-convergence in distribution and stochastic equi-semicontinuity," *Unpublished manuscript*, 37. 32, 39

——— (2001): "Limiting distributions of linear programming estimators," *Extremes*, 4, 87–103. 4

——— (2006): "Asymptotic theory for M-estimators of boundaries," in *The Art of Semiparametrics*, Springer, 1–21. 4

——— (2010): "On the asymptotic distribution of the analytic center estimator," in *Nonparametrics and Robustness in Modern Statistical Inference and Time Series Analysis: A Festschrift in honor of Professor Jana Jurečková*, Institute of Mathematical Statistics, 123–133. 4

Kosorok, M. R. (2007): *Introduction to empirical processes and semiparametric inference*, Springer. 6, 7, 12, 35, 36

Li, J. (2021): "The Proximal Bootstrap for Finite-Dimensional Regularized Estimators," *American Economic Association Papers and Proceedings*, 111, 616–620. 1

Moon, H. R. and F. Schorfheide (2009): "Estimation with overidentifying inequality moment conditions," *Journal of Econometrics*, 153, 136–154. 4

Newey, W. and D. McFadden (1994): "Large Sample Estimation and Hypothesis Testing," in *Handbook of Econometrics, Vol. 4*, ed. by R. Engle and D. McFadden, North Holland, 2113–2241. 17

NOCEDAL, J. AND S. WRIGHT (2006): *Numerical optimization*, Springer Science & Business Media. 13

POLLARD, D. (1984): "Convergence of stochastic processes," . 31

——— (1985): "New ways to prove central limit theorems," *Econometric Theory*, 1, 295–313. 14

——— (1991): "Asymptotics for least absolute deviation regression estimators," *Econometric Theory*, 7, 186–199. 38

ROMANO, J. P. AND A. M. SHAIKH (2008): "Inference for identifiable parameters in partially identified econometric models," *Journal of Statistical Planning and Inference*, 138, 2786–2807. 2

——— (2012): "On the uniform asymptotic validity of subsampling and the bootstrap," *The Annals of Statistics*, 40, 2798–2822. 36

SHAPIRO, A. (1988): "Sensitivity analysis of nonlinear programs and differentiability properties of metric projections," *SIAM Journal on Control and Optimization*, 26, 628–645. 4, 17, 30, 37

——— (1989): "Asymptotic properties of statistical estimators in stochastic programming," *The Annals of Statistics*, 841–858. 4, 17, 30, 37

——— (1990): "On differential stability in stochastic programming," *Mathematical Programming*, 47, 107–116. 4

VAN DER VAART, A. AND J. A. WELLNER (1996): *Weak Convergence and Empirical Processes*, Springer. 30

WACHSMUTH, G. (2013): "On LICQ and the uniqueness of Lagrange multipliers," *Operations Research Letters*, 41, 78–80. 13

WELLNER, J. A. AND Y. ZHAN (1996): "Bootstrapping Z-estimators," *University of Washington Department of Statistics Technical Report*, 308. 13, 34