The Proximal Bootstrap for Constrained Estimators *

Jessie Li

August 26, 2022

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 that can be either non-random or estimated from the data at the $\sqrt{n}$ rate. The proximal bootstrap estimator is typically much faster to compute than the standard bootstrap 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 nonsmooth GMM model, a conditional logit model with estimated capacity constraints, and a mathematical programming with equilibrium constraints (MPEC) formulation of the Rust (1987) Bus Engine Replacement model proposed in Su and Judd (2012).

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

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 which we will focus on in this paper is estimators defined by the solution to a constrained optimization problem with a possibly nonsmooth and nonconvex sample objective function and either

*I would like to thank the participants of the University of Chicago Econometrics workshop (in particular Stéphane Bonhomme, Azeem Shaikh, and Alex Torgovitsky), the 2021 ASSA session Optimization-conscious Econometrics, the 2021 North American Summer Meeting of the Econometrics Society, UC Davis Econometrics workshop (in particular Bulat Gafarov and Takuya Ura), and the Yale Econometrics workshop (in particular Donald Andrews, Xiaohong Chen, Timothy Christensen, and Jean-Jacques Forneron) for helpful comments and suggestions.

estimated or non-random smooth inequality and/or equality constraints. 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 (Andrews (1999),Andrews (2000),Andrews (2002)).

Motivated by the optimization literature and recent contributions in computationally efficient bootstrap procedures (e.g. 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 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 conservatively valid inference on the entire parameter vector using a confidence set constructed by inverting the optimal value function. We can also conduct uniformly conservatively valid inference on subvectors of the parameter vector using two-sided intervals obtained through projection. The proximal bootstrap relies on a scaling sequence (labeled α_n in this paper) that converges to zero at a slower than $\sqrt{n}$ rate, 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.

Another novel part of this paper is that we provide a general asymptotic distribution for estimators defined by the solution to a constrained optimization problem with equality and/or inequality constraints which can be estimated from the data, while Hong and Li (2020) looked only at estimators with non-random constraints that do not depend on the data. The asymptotic distribution of constrained estimators with estimated constraints is derived using ideas from the optimization literature and encompasses as special cases the results in Geyer (1994), Andrews (1999),Andrews (2000), and Andrews (2002) for constrained estimators with non-random constraint sets and true parameters possibly lying on the boundaries of the constraint sets.

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, but 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 $\sqrt{n}\alpha_n \to \infty$. For estimators with estimated constraint sets, α_n will also serve as a selection device so that the active constraints are included in the asymptotic distribution while the nonactive, non-drifting constraints are not.

We were inspired to write this paper after reading a series of papers by Alexander Shapiro: Shapiro (1988), Shapiro (1989), Shapiro (1990), Shapiro (1991), Shapiro (1993), Shapiro (2000), and also by Keith Knight: 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), and Fang and Seo (2021). While many of these papers are concerned with either conducting inference on the optimal value of the constrained optimization problem or testing whether the parameter of interest satisfies the constraints, we are mainly interested in conducting inference on the optimal solution assuming that the constraints are valid. 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 allow for non-unique solutions. Our inference procedure is also different from theirs because we use resampling while they exploit the fact that the primal-dual formulation of the KKT conditions can be written as a set of moment inequalities and then apply test inversion.

The outline of our paper is as follows. Section 2 contains the main theoretical results, starting with Subsection 2.1 which contains the notation followed by Subsection 2.2 which briefly reviews the concept of proximal mappings from the optimization literature. Subsection 2.3 shows consistency of the proximal bootstrap for finite-dimensional constrained estimators with non-random constraints, and Subsection 2.4 shows consistency for estimated constraints. In both the non-random constraints case and the estimated constraints case, the proximal bootstrap can consistently replicate the asymptotic distribution when parameters are on the boundary of the constraint set, but not when parameters are drifting towards the boundary. Nevertheless, as demonstrated in Subsection 2.5, we can still use the proximal bootstrap to conduct asymptotically uniformly conservatively valid inference by inverting the optimal value function. Section 3 contains Monte Carlo simulation evidence demonstrating the validity of confidence intervals constructed using the proximal bootstrap for a boundary constrained nonsmooth GMM model, a conditional logit model with estimated capacity constraints, and the mathematical programming with equilibrium constraints (MPEC) formulation of the Rust (1987) Bus Engine Replacement model proposed in Su and Judd (2012). Section 4 concludes. Section 5 is the Appendix which contains proofs of the theorems and some auxiliary results.

2 Proximal Bootstrap

2.1 Notation

Consider a random sample $X_1, X_2, ..., X_n$ 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^*, which can refer to the multinomial, wild, or other exchangeable bootstraps. 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. 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 the law of X_n^* is governed by P_n and if $P_n(|X_n^*| > \epsilon) = o_P(1)$ for all $\epsilon > 0$. Also define $M_n^* = O_P^*(1)$ (hence also $O_P(1)$) if $\lim_{m \to \infty} \limsup_{n \to \infty} P(P_n(M_n^* > m) > \epsilon) \to 0$, $\forall \epsilon > 0$.

2.2 Proximal Mappings

Given an Euclidean space $\mathbb{D}$ and a function $r : \mathbb{D} \mapsto \mathbb{R}$, the proximal mapping of r is the operator given by

$$prox_r(z) = \underset{\beta \in \mathbb{D}}{\arg\min} \left\{ r(\beta) + \frac{1}{2} \|\beta - z\|_2^2 \right\} \text{ for any } z \in \mathbb{D}$$

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 $\|\beta - z\|_H^2 = (\beta - z)' H (\beta - z)$,

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

When r is a proper closed and convex function then $prox_r(z)$ is a singleton for any $z \in \mathbb{D}$ (Theorem 6.3 Beck (2017)). The same can be said for $prox_{H,r}(z)$ (Lee et al. (2014)). Although it is rarely the case that the scaled proximal map has a closed form solution, the solution can be efficiently computed using various proximal algorithms (see e.g. Lee et al. (2012), Lee et al. (2014), Parikh et al. (2014), Tran-Dinh et al. (2015), Ghanbari and Scheinberg (2016), Rodomanov and Kropotov (2016), Byrd et al. (2016)).

2.3 Constrained Estimators with Non-random Constraints

It is well known (see e.g. Andrews (2000)) that the standard bootstrap is inconsistent when the true parameters β_0 lie on the boundary of the constraint set C. 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. We first consider constrained estimators with non-random constraints $\hat{\beta}_n = \underset{\beta \in C}{\arg\min} \hat{Q}_n(\beta)$, where $C \subseteq \mathbb{B}$ is a non-random, closed constraint set that is a subset of the compact parameter space $\mathbb{B} \subset \mathbb{R}^d$, where d is fixed, 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 $\beta_0 = \underset{\beta \in C}{\arg\min} Q(\beta)$. We assume both $\hat{\beta}_n$ and β_0 are unique, which rules out partially identified models.

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 2.5 that the proximal bootstrap can be used to form a uniformly conservatively valid confidence set for either the whole parameter vector or subvectors of the parameter vector. Because the more general results in Section 2.5 cover the case of non-random constraints as a special case, we will defer discussion of drifting sequences in the case of non-random constraints until Section 2.5. Geyer (1994) shows that if $Q(\beta)$ achieves its minimum over C at some point β_0 where it has a local quadratic approximation $Q(\beta) = Q(\beta_0) + \frac{1}{2}(\beta-\beta_0)' H_0 (\beta-\beta_0) + o\left(\|\beta-\beta_0\|^2\right)$, where $H_0 = \left.\frac{\partial^2 Q(\beta)}{\partial\beta\partial\beta'}\right|_{\beta=\beta_0}$ is positive definite, then $\sqrt{n}\left(\hat{\beta}_n - \beta_0\right) \rightsquigarrow \mathcal{J} = \underset{h\in T_C(\beta_0)}{\arg\min}\left\{h'W_0 + \frac{1}{2}h'H_0h\right\}$, where W_0 is a Gaussian and $T_C(\beta_0) \equiv \limsup_{\tau\downarrow 0}\frac{C-\beta_0}{\tau}$ is the tangent cone of C at β_0. For closed sets C that are Chernoff Regular at β_0, the limit exists and $T_C(\beta_0) = \lim_{\tau\downarrow 0}\frac{C-\beta_0}{\tau}$.

Note that the assumption that $Q(\beta)$ has a local quadratic approximation at β_0 of the form $Q(\beta) = Q(\beta_0) + \frac{1}{2}(\beta-\beta_0)' H_0(\beta-\beta_0) + o\left(\|\beta-\beta_0\|^2\right)$ effectively assumes $l(\beta_0) = \left.\frac{\partial Q(\beta)}{\partial\beta}\right|_{\beta=\beta_0} = 0$ (this is noted on the top of page 2000 of Geyer (1994)). In other words, the constraints are not necessary for identification of β_0. We will relax this assumption in Section 2.4 to allow for $l(\beta_0) \neq 0$. When $l(\beta_0) = 0$, one way that we can define the proximal bootstrap estimator is for some $\alpha_n \to 0$ and $\alpha_n\sqrt{n} \to \infty$,

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

Here, $\bar{\beta}_n$ is an initial $\sqrt{n}$-consistent estimator of β_0. For example, we can use $\bar{\beta}_n = \hat{\beta}_n$. The sequence α_n ensures that $\bar{\beta}_n$'s estimation error does not enter into the proximal bootstrap approximation of $\hat{\beta}_n$'s asymptotic distribution. $\hat{l}_n\left(\bar{\beta}_n\right)$ is a consistent estimate of $l(\beta_0)$ using $\hat{\beta}_n$, 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 $\hat{\beta}_n$. More generally, to handle non-differentiable $\hat{Q}_n(\beta)$, $\hat{l}_n\left(\hat{\beta}_n\right)$ is a subgradient of $\hat{Q}_n(\beta)$ at $\hat{\beta}_n$. $\bar{H}_n$ is a consistent, symmetric, positive definite estimate of the population Hessian H_0 using $\hat{\beta}_n$.

If C is a convex set, then this formulation of the proximal bootstrap solves a convex optimization problem. If C is not convex, we can linearize the constraints to make the problem convex assuming that a constraint qualification is satisfied which ensures the linearized constraints sufficiently capture the geometry of the constraints around the solution. Let the constraint set be $C = \{\beta \in \mathbb{B} : f_j(\beta) = 0 \text{ for } j \in \mathcal{E}, f_j(\beta) \leqslant 0 \text{ for } j \in \mathcal{I}\}$. Then for $\bar{F}_j = \frac{\partial f_j(\beta)}{\partial \beta}\Big|_{\beta=\bar{\beta}_n}$, we can define an alternative proximal bootstrap estimator using a linearized constraint set as

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

$$C^* = \left\{ \beta \in \mathbb{B} : f_j\left(\bar{\beta}_n\right) + \bar{F}_j' \left(\beta - \bar{\beta}_n\right) = 0 \text{ for } j \in \mathcal{E}, f_j\left(\bar{\beta}_n\right) + \bar{F}_j' \left(\beta - \bar{\beta}_n\right) \leqslant 0 \text{ for } j \in \mathcal{I} \right\}$$

Because this version of the proximal bootstrap with a linearized constraint set C^* is a special case of the more general result in Subsection 2.4's Theorem 2, we do not prove it in this section. We will only consider the version with the nonlinearized constraint set C and show that $\frac{\hat{\beta}_n^* - \hat{\beta}_n}{\alpha_n}$ consistently estimates the asymptotic distribution of $\sqrt{n}\left(\hat{\beta}_n - \beta_0\right)$ when the parameters are not drifting towards the boundary and $l(\beta_0) = 0$. Before we present the theorem, we list a few assumptions needed for the theorem to hold.

The first assumption is needed to show consistency of $\hat{\beta}_n$ for β_0.

Assumption 1. *(i) $\mathbb{B} \subset \mathbb{R}^d$ is compact and d is fixed. (ii) $\hat{\beta}_n = \underset{\beta \in C \subseteq \mathbb{B}}{\arg\min} \hat{Q}_n(\beta)$ is uniformly tight and unique.* [1]

(iii) $\beta_0 = \underset{\beta \in C}{\arg\min}\ Q(\beta)$ is unique.

(iv) $Q(\beta)$ is a lower semicontinuous function that is twice continuously differentiable at β_0, and $\sup_{\beta \in K} \left| \hat{Q}_n(\beta) - Q(\beta) \right| = o_P(1)$ for every compact subset K of C.

The next assumption states that $\hat{Q}_n(\beta)$ admits a uniform local quadratic approximation around $\sqrt{n}$ neighborhoods of β_0. This assumption does not require $\hat{Q}_n(\beta)$ to be differentiable at β_0 since $\hat{l}_n(\beta)$ does not need to be the Jacobian of $\hat{Q}_n(\beta)$. This assumption is similar to the stochastic differentiability assumption in Pollard (1985) and is needed to derive the asymptotic distribution of $\sqrt{n}\left(\hat{\beta}_n - \beta_0\right)$.

[1] Uniform tightness means for every $\epsilon > 0$, there exists a compact $K \subset C$ with $P\left(\hat{\beta}_n \in K\right) \geqslant 1 - \epsilon$ for every n.

Assumption 2. *There exists a symmetric, positive definite* H_0 *and* $\sqrt{n}\left(\hat{l}_n(\beta_0) - l(\beta_0)\right) = O_P(1)$ *such that for any* $\delta_n \to 0$,

$$\sup_{\|h\| \leqslant \sqrt{n}\delta_n} \left| \frac{n\hat{Q}_n\left(\beta_0 + \frac{h}{\sqrt{n}}\right) - n\hat{Q}_n(\beta_0) - h'\sqrt{n}\left(\hat{l}_n(\beta_0) - l(\beta_0)\right) - \frac{1}{2}h'H_0 h}{1 + \|h\|^2} \right| = o_P(1)$$

The role of the shrinking sequence δ_n is localize the quadratic approximation around β_0. A similar assumption can be found in Gallant et al. (2022).

The next assumption is needed to 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.

Assumption 3. *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* $\epsilon > 0$.

The next assumption is needed to show stochastic equicontinuity and bootstrap equicontinuity results which will be used to show $\sqrt{n}\left(\hat{l}_n^*(\bar{\beta}_n) - \hat{l}_n(\bar{\beta}_n)\right)$ and $\sqrt{n}\left(\hat{l}_n^*(\beta_0) - \hat{l}_n(\beta_0)\right)$ have the same asymptotic distribution.

Assumption 4. *(i)* $\mathcal{G}_R \equiv \{g(\cdot, \beta) - g(\cdot, \beta_0) : \|\beta - \beta_0\| \leqslant R\}$ *is a Donsker class for some* $R > 0$ *and* $P\|g(\cdot, \beta) - g(\cdot, \beta_0)\|^2 \to 0$ *for* $\beta \to \beta_0$.

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

(i) will imply stochastic equicontinuity, which in combination with the envelope function integrability condition in (ii) will imply bootstrap equicontinuity. A sufficient condition for (ii) 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 $\delta_n \to 0$.

Our first theorem shows that the proximal bootstrap can consistently estimate the non-standard distribution of constrained estimators with non-random constraint sets when the parameters are not drifting towards the boundary. Of particular importance is the sequence α_n which converges to zero at a slower than $\sqrt{n}$ rate. The purpose of the slower than $\sqrt{n}$ rate is to offset the estimation error from the initial $\sqrt{n}$-consistent estimator $\hat{\beta}_n$.

Theorem 1. *Suppose Assumptions 1-4 are satisfied, $C \subset \mathbb{R}^d$ is a non-random closed set that is Chernoff Regular at $\beta_0 = \underset{\beta \in C}{\arg\min} Q(\beta)$, and $Q(\beta) = Q(\beta_0) + \frac{1}{2}(\beta - \beta_0)' H_0 (\beta - \beta_0) + o\left(\|\beta - \beta_0\|^2\right)$, where $H_0 > 0$. For any $\bar{\beta}_n$ such that $\sqrt{n}\left(\bar{\beta}_n - \beta_0\right) = O_P(1)$ and $\bar{H}_n \xrightarrow{p} H_0$, let*

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

For any 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 $\frac{\hat{\beta}_n^* - \hat{\beta}_n}{\alpha_n} \underset{\mathbb{W}}{\overset{\mathbb{P}}{\rightsquigarrow}} \mathcal{J}$, where $\mathcal{J} = \underset{h \in T_C(\beta_0)}{\arg\min} \left\{ h' W_0 + \frac{1}{2} h' H_0 h \right\}$, $T_C(\beta_0) = \lim_{\tau \downarrow 0} \frac{C - \beta_0}{\tau}$, and $W_0 \sim N\left(0, P\left(g(\cdot, \beta_0) - Pg(\cdot, \beta_0)\right)\left(g(\cdot, \beta_0) - Pg(\cdot, \beta_0)\right)'\right)$.

Remark 1. We can also show that the optimal value's asymptotic distribution can be consistently estimated by the proximal bootstrap when the parameters are not drifting towards the boundary. In particular, $n\left(\hat{Q}_n\left(\hat{\beta}_n\right) - \hat{Q}_n(\beta_0)\right) \rightsquigarrow q(\mathcal{J})$, where $q(h) \equiv h'W_0 + \frac{1}{2} h' H_0 h$, and $\frac{\hat{A}_n^*(\hat{\beta}_n^*) - \hat{A}_n^*(\hat{\beta}_n)}{\alpha_n^2} \underset{\mathbb{W}}{\overset{\mathbb{P}}{\rightsquigarrow}} q(\mathcal{J})$. This result follows from Theorem 4.4 in Geyer (1994) in combination with $\frac{\hat{A}_n^*(\hat{\beta}_n^*) - \hat{A}_n^*(\hat{\beta}_n)}{\alpha_n^2} = \sqrt{n}\left(\hat{l}_n^*(\bar{\beta}_n) - \hat{l}_n(\bar{\beta}_n)\right)' \left(\frac{\hat{\beta}_n^* - \beta_0}{\alpha_n}\right) + \frac{1}{2}\left\|\frac{\hat{\beta}_n^* - \beta_0}{\alpha_n}\right\|^2_{\bar{H}_n} + o_p(1) \underset{\mathbb{W}}{\overset{\mathbb{P}}{\rightsquigarrow}} q(\mathcal{J})$.

Remark 2. We can remove the assumption that C is a closed set by assuming instead that $\mathcal{J} = \underset{h \in T_C(\beta_0)}{\arg\min} \left\{ h' W_0 + \frac{1}{2} h' H_0 h \right\}$ is almost surely unique. This can happen for example if we strengthen the condition on C to Clarke Regularity at β_0 (see Geyer (1994) page 1997 or Rockafellar et al. (1998) Definition 6.4 page 199 for a definition), which implies that $T_C(\beta_0)$ is a convex cone. Every convex set is Clarke Regular, but Clarke Regularity is weaker than assuming convexity of C. See example 3 in Geyer (1994) for an example of a set that is Clarke Regular but not convex.

Remark 3. A special case is when $\beta_0 = \underset{\beta \in C}{\arg\min} Q(\beta)$ lies in the interior of C. Then as noted in several papers (e.g. Andrews (1999), Andrews (2002), Chen et al. (2018)), $T_C(\beta_0) = \mathbb{R}^d$ and $\mathcal{J}$ is multivariate normal. Another special case of C is when there are only equality constraints: $C = \left\{\beta \in \mathbb{R}^d : f(\beta) = 0\right\}$ where $f(\beta)$ are constraints that do not depend on the data. It is well known from Amemiya (1985) and Newey and McFadden (1994) that $\mathcal{J}$ is multivariate normal.

Remark 4. If $l(\beta_0) \neq 0$, then it is important to include the Lagrange multiplier weighted constraint

Hessians when defining the proximal bootstrap objective function:

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

where $C \equiv \{\beta \in \mathbb{B} : f_j(\beta) = 0 \text{ for } j \in \mathcal{E}, f_j(\beta) \leqslant 0 \text{ for } j \in \mathcal{I}\}$, $\bar{G}_{nj} \xrightarrow{p} \left.\frac{\partial^2 f_j(\beta)}{\partial \beta \partial \beta'}\right|_{\beta = \beta_0}$ for all $j \in \mathcal{E} \cup \mathcal{I}$, and $\bar{\lambda}_n$ are a set of optimal Lagrange multipliers for $\bar{\beta}_n$. The reason for including the extra term will be described in Section 2.4, where we provide a more general asymptotic distribution for when there are estimated constraints and $l(\beta_0)$ may be nonzero.

2.4 Constrained Estimators with Estimated Constraints

Now we consider 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 $\mathbb{B} \subset \mathbb{R}^d$.

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

We will define the population analog of C 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 $\beta_0 \equiv \underset{\beta \in C_0}{\arg\min} Q(\beta)$, which is assumed to be unique. $Q(\beta)$ is twice continuously differentiable at β_0 and $\sup_{\beta \in \mathbb{B}} \left| \hat{Q}_n(\beta) - Q(\beta) \right| = o_P(1)$.

For simplicity, we will impose that the population 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 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 that the gradients of the equality constraints are linearly independent.

Assumption 5. *Suppose Linear Independence Constraint Qualification (LICQ) holds at β_0 : the gradients of the active constraints $F_{0j} \equiv \left.\frac{\partial f_{0j}(\beta)}{\partial \beta}\right|_{\beta=\beta_0}$ for $j \in \mathcal{E} \cup \mathcal{I}^*$, where $\mathcal{I}^* \equiv \{j \in \mathcal{I} : f_{0j}(\beta_0) = 0\}$, are linearly independent.*

Instead of Assumption 2, we now require that the Lagrangian has a uniform local quadratic approximation in $\sqrt{n}$ neighborhoods of β_0. The importance of using the Lagrangian instead of the objective function is that it allows for the pseudo-true parameters to not be a solution of the unconstrained population optimization problem; in other words, we allow for the possibility that $l(\beta_0) \neq 0$.

Assumption 6. *Suppose $f_{nj} : \mathbb{B} \mapsto \mathbb{R}$ and $f_{0j} : \mathbb{B} \mapsto \mathbb{R}$ are twice continuously differentiable functions that satisfy $\sup_{\beta \in \mathbb{B}} |f_{nj}(\beta) - f_{0j}(\beta)| = o_P(1)$ for all $j \in \mathcal{E} \cup \mathcal{I}$. Define λ_{nj} to be a set of optimal Lagrange multipliers for $\hat{\beta}_n$. Define $\hat{\mathcal{L}}_n(\beta) \equiv \hat{Q}_n(\beta) + \sum_{j \in \mathcal{E} \cup \mathcal{I}} \lambda_{nj} f_{nj}(\beta)$, $F_{nj}(\beta_0) \equiv \left.\frac{\partial f_{nj}(\beta)}{\partial \beta}\right|_{\beta=\beta_0}$, and $G_{0j} \equiv \left.\frac{\partial^2 f_{0j}(\beta)}{\partial \beta \partial \beta'}\right|_{\beta=\beta_0}$. Suppose H_0 and $B_0 = H_0 + \sum_{j \in \mathcal{E} \cup \mathcal{I}} \lambda_{0j} G_{0j}$ are symmetric, positive definite. For any $\delta_n \to 0$,*

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

where λ_{0j} are 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 $\nabla \mathcal{L}(\beta_0, \lambda_0) \equiv l(\beta_0) + \sum_{j \in \mathcal{E} \cup \mathcal{I}} \lambda_{0j} F_{0j} = 0$.

Note that since $\nabla \mathcal{L}(\beta_0, \lambda_0) \equiv l(\beta_0) + \sum_{j \in \mathcal{E} \cup \mathcal{I}} \lambda_{0j} F_{0j} = 0$, Assumption 6 can also be written as follows: for any $\delta_n \to 0$,

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

When $l(\beta_0) = 0$ and $\max_{j \in \mathcal{E} \cup \mathcal{I}} |\lambda_{nj} - \lambda_{0j}| \xrightarrow{p} 0$, which follows from $\hat{\beta}_n$ being consistent for β_0, Assumption 6 will imply Assumption 2 because $\lambda_{0j} = 0$ for all $j \in \mathcal{E} \cup \mathcal{I}$. A more in-depth discussion of why $\lambda_{0j} = 0$ appears in Remark 6. However, if $l(\beta_0) \neq 0$, Assumption 6 will not necessarily imply Assumption 2 because some λ_{0j} may not be zero.

In principle, we do not require that there exists a set of unique optimal Lagrange multipliers λ_{nj} for $\hat{\beta}_n$, although in practice it is usually the case that λ_{nj} are unique. This is because the active constraint gradients in the sample are almost surely linearly independent when LICQ is satisfied in the population and the sample constraints converge uniformly to the population constraints.

Next, we define the proximal bootstrap estimator. 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 $\bar{F}_{nj} \equiv F_{nj}\left(\bar{\beta}_n\right)$, $\bar{F}_{nj}^* \equiv F_{nj}^*\left(\bar{\beta}_n\right)$, $\bar{G}_{nj} \xrightarrow{p} \left.\frac{\partial^2 f_{0j}(\beta)}{\partial \beta \partial \beta'}\right|_{\beta = \bar{\beta}_n}$ 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 $\bar{\beta}_n$. Define $\hat{\beta}_n^* \equiv \underset{\beta \in C^*}{\arg\min} \hat{A}_n^*(\beta)$, for

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

Note that the proximal bootstrap estimator is the solution to a quadratic programming problem, which is a convex problem if $\bar{H}_n + \sum_{j \in \mathcal{E} \cup \mathcal{I}} \bar{\lambda}_{nj} \bar{G}_{nj}$ 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 (e.g. see the MPEC Rust (1987) example in the Monte Carlo simulations). We do not require that there exists a set of unique optimal Lagrange multipliers $\bar{\lambda}_{nj}$ for $\bar{\beta}_n$, although in practice it is usually the case that $\bar{\lambda}_{nj}$ are unique for $\bar{\beta}_n = \hat{\beta}_n$ because of our assumptions of LICQ and uniform convergence of the sample constraints to the population constraints.

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. The key for proximal bootstrap consistency lies in the scaling sequence α_n which converges to zero at a slower than $\sqrt{n}$ rate. Here, α_n serves the dual purpose of offsetting the estimation error from $\bar{\beta}_n$ and also selecting the active constraints to be included in the asymptotic distribution while dropping the nonactive constraints.

Theorem 2. *Suppose Assumption 1 (after setting $\beta_0 \equiv \arg\min_{\beta \in C_0} Q(\beta)$) and Assumptions 3 - 6 are satisfied in addition to the following:*

(i) Suppose $\hat{\beta}_n - \beta_0 \xrightarrow{p} 0$.

(ii) Suppose $\hat{\beta}_n^* - \hat{\beta}_n = o_P^*(1)$.

(iii) Suppose $\nabla^2 \mathcal{L}(\beta_0, \lambda_0) \equiv H_0 + \sum_{j \in \mathcal{E} \cup \mathcal{I}} \lambda_{0j} G_{0j}$ is positive definite.

(iv) Suppose $\sqrt{n}\left(f_n(\beta_0) - f_0(\beta_0)\right) \rightsquigarrow U_0$, a tight random vector, and $\sqrt{n}\left(\hat{l}_n(\beta_0) - l(\beta_0)\right) + \sum_{j \in \mathcal{E} \cup \mathcal{I}} \lambda_{0j} \sqrt{n}\left(F_{nj}(\beta_0) - F_{0j}\right) \rightsquigarrow W_0 + \sum_{j \in \mathcal{E} \cup \mathcal{I}} \lambda_{0j} V_{0j}$, a tight random vector.

(v) Suppose $\sqrt{n}\left(f_n^*(\beta_0) - f_n(\beta_0)\right) \underset{\mathbb{W}}{\overset{\mathbb{P}}{\rightsquigarrow}} U_0$, $\max_{j \in \mathcal{E} \cup \mathcal{I}} \left|\bar{\lambda}_{nj} - \lambda_{0j}\right| \xrightarrow{p} 0$,

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

$$\sup_{\|\beta - \beta_0\| \leqslant o(1)} \sqrt{n}\left(f_n^*(\beta) - f_n(\beta) - f_n^*(\beta_0) + f_n(\beta_0)\right) = o_P^*(1), \text{ and}$$

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

(vi) $\bar{H}_n \xrightarrow{p} H_0$, $\max_{j \in \mathcal{E} \cup \mathcal{I}} \left|\bar{G}_{nj} - G_{0j}\right| \xrightarrow{p} 0$, and $\bar{H}_n + \sum_{j \in \mathcal{E} \cup \mathcal{I}} \bar{\lambda}_{nj} \bar{G}_{nj}$ is symmetric, positive definite.

Suppose $f_{0j}(\beta_0)$ for $j \in \mathcal{I}$ does not depend on 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 $\frac{\hat{\beta}_n^* - \hat{\beta}_n}{\alpha_n} \underset{\mathbb{W}}{\overset{\mathbb{P}}{\rightsquigarrow}} \mathcal{J}$, where for $\mathcal{I}_+^*(\lambda_0) \equiv \{j \in \mathcal{I}^* : \lambda_{0j} > 0\}$ and $\mathcal{I}_0^*(\lambda_0) \equiv \{j \in \mathcal{I}^* : \lambda_{0j} = 0\}$,

$$\mathcal{J} = \arg\min_{h \in \Sigma} \left\{ h' W_0 + \frac{1}{2} h' H_0 h + \sum_{j \in \mathcal{E} \cup \mathcal{I}_+^*(\lambda_0)} \lambda_{0j} \left(h' V_{0j} + \frac{1}{2} h' G_{0j} h \right) \right\}$$

$$\Sigma = \left\{ h : U_{0j} + F_{0j}' h = 0 \text{ for } j \in \mathcal{E} \cup \mathcal{I}_+^*(\lambda_0), U_{0j} + F_{0j}' h \leqslant 0 \text{ for } j \in \mathcal{I}_0^*(\lambda_0) \right\}$$

A sufficient condition for $\sqrt{n}\left(\hat{l}_n(\beta_0) - l(\beta_0)\right) + \sum_{j\in\mathcal{E}\cup\mathcal{I}} \lambda_{0j}\sqrt{n}\left(F_{nj}(\beta_0) - F_{0j}\right) \rightsquigarrow W_0 + \sum_{j\in\mathcal{E}\cup\mathcal{I}} \lambda_{0j} V_{0j}$ is $\begin{pmatrix} \sqrt{n}\left(\hat{l}_n(\beta_0) - l(\beta_0)\right) \\ \sqrt{n}\left(F_n(\beta_0) - F_0\right) \end{pmatrix} \rightsquigarrow \begin{pmatrix} W_0 \\ V_0 \end{pmatrix}$, where $V_0 = (V_{0j} \text{ for } j \in \mathcal{E}\cup\mathcal{I})$. Similarly, a sufficient condition for $\sqrt{n}\left(\hat{l}_n^*(\beta_0) - \hat{l}_n(\beta_0)\right) + \sum_{j\in\mathcal{E}\cup\mathcal{I}} \lambda_{0j}\sqrt{n}\left(F_{nj}^*(\beta_0) - F_{nj}(\beta_0)\right) \underset{\mathbb{W}}{\overset{\mathbb{P}}{\rightsquigarrow}} W_0 + \sum_{j\in\mathcal{E}\cup\mathcal{I}} \lambda_{0j} V_{0j}$ is $\begin{pmatrix} \sqrt{n}\left(\hat{l}_n^*(\beta_0) - \hat{l}_n(\beta_0)\right) \\ \sqrt{n}\left(F_n^*(\beta_0) - F_n(\beta_0)\right) \end{pmatrix} \underset{\mathbb{W}}{\overset{\mathbb{P}}{\rightsquigarrow}} \begin{pmatrix} W_0 \\ V_0 \end{pmatrix}$. When $F_n(\beta) = P_n\pi(\cdot,\beta)$ and $F_n^*(\beta) = P_n^*\pi(\cdot,\beta)$ are sample averages, these joint weak convergence statements can be verified using a joint Lindeberg condition. For each $\epsilon > 0$,

$$\lim_{n\to\infty} P \left\| \begin{pmatrix} g(\cdot,\beta_0) \\ \pi(\cdot,\beta_0) \end{pmatrix} \right\|^2 1\left\{ \left\| \begin{pmatrix} g(\cdot,\beta_0) \\ \pi(\cdot,\beta_0) \end{pmatrix} \right\| > \epsilon\sqrt{n} \right\} = 0.$$

Theorem 2 tells us that both two-sided and one-sided confidence intervals constructed using the proximal bootstrap critical values will be asymptotically exact when the population inequality constraints $f_{0j}(\beta_0)$ are not drifting towards zero. Later, in Section 2.5, we will consider the case of drifting constraints and show how to construct a uniform confidence set for either the whole parameter vector or subvectors. Before we discuss drifting sequences, we make some remarks on special cases of the general asymptotic distribution in Theorem 2.

Remark 5. The optimal value's asymptotic distribution can also be consistently estimated by the proximal bootstrap under pointwise, non-drifting asymptotics. Specifically, $n\left(\hat{\mathcal{L}}_n\left(\hat{\beta}_n\right) - \hat{\mathcal{L}}_n(\beta_0)\right) \rightsquigarrow q(\mathcal{J})$, where $q(h) \equiv h'W_0 + \frac{1}{2}h'H_0h + \sum_{j\in\mathcal{E}\cup\mathcal{I}_+^*(\lambda_0)} \lambda_{0j}\left(h'V_{0j} + \frac{1}{2}h'G_{0j}h\right)$, and $\frac{\hat{A}_n^*(\hat{\beta}_n^*) - \hat{A}_n^*(\hat{\beta}_n)}{\alpha_n^2} \underset{\mathbb{W}}{\overset{\mathbb{P}}{\rightsquigarrow}} q(\mathcal{J})$. The first result follows from Assumption 6 and the continuous mapping theorem, and the second result follows from $\frac{\hat{A}_n^*(\hat{\beta}_n^*) - \hat{A}_n^*(\hat{\beta}_n)}{\alpha_n^2} = \sqrt{n}\left(\hat{l}_n^*(\bar{\beta}_n) - \hat{l}_n(\bar{\beta}_n)\right)'\left(\frac{\hat{\beta}_n^* - \beta_0}{\alpha_n}\right) + \frac{1}{2}\left\|\frac{\hat{\beta}_n^* - \beta_0}{\alpha_n}\right\|_{\bar{H}_n}^2 + \sum_{j\in\mathcal{E}\cup\mathcal{I}} \bar{\lambda}_{nj}\left(\sqrt{n}\left(\bar{F}_{nj}^* - \bar{F}_{nj}\right)'\left(\frac{\hat{\beta}_n^* - \beta_0}{\alpha_n}\right) + \frac{1}{2}\left\|\frac{\hat{\beta}_n^* - \beta_0}{\alpha_n}\right\|_{\bar{G}_{nj}}^2\right) + o_p(1) \underset{\mathbb{W}}{\overset{\mathbb{P}}{\rightsquigarrow}} q(\mathcal{J})$.

Remark 6. If $l(\beta_0) = 0$, which is implied by $Q(\beta) = Q(\beta_0) + \frac{1}{2}(\beta - \beta_0)' H_0 (\beta - \beta_0) + o\left(\|\beta - \beta_0\|^2\right)$, then $\mathcal{J}$ reduces down to

$$\mathcal{J} = \arg\min_{h\in\Sigma} \left\{ h'W_0 + \frac{1}{2}h'H_0h \right\}$$

$$\Sigma = \left\{h : U_{0j} + F_{0j}'h = 0 \text{ for } j \in \mathcal{E}, U_{0j} + F_{0j}'h \leqslant 0 \text{ for } j \in \mathcal{I}_0^*(\lambda_0)\right\}$$

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 $\lambda_{0j} f_{0j}(\beta_0) = 0$ 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 $\sum_{j\in\mathcal{E}\cup\mathcal{I}} \lambda_{0j} \left(h'V_{0j} + \frac{1}{2}h'G_{0j}h\right) = 0$, $\mathcal{I}_+^*(\lambda_0) = \varnothing$, and $\mathcal{I}^* = \mathcal{I}_0^*(\lambda_0)$.

In this case, it is easy to extend our theory to the case where the number of constraints is growing with n, assuming that the dimension of β is fixed. We redefine the proximal bootstrap estimator as $\hat{\beta}_n^* \equiv \underset{\beta\in C^*}{\arg\min} \hat{A}_n^*(\beta)$, where

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

Remark 7. If there are only equality constraints, then the asymptotic distribution becomes $\mathcal{J} = \underset{h\in\Sigma}{\arg\min}\left\{h'W_0 + \frac{1}{2}h'\left(H_0 + \sum_{j\in\mathcal{E}} \lambda_{0j} G_{0j}\right)h\right\}$ for $\Sigma = \left\{h : U_{0j} + F_{0j}'h = 0 \text{ for } j \in \mathcal{E}\right\}$. Using standard arguments in Amemiya (1985) section 1.4.1 or Newey and McFadden (1994) section 9.1 (which are repeated in Lemma 5.1 in the Appendix), $\mathcal{J} = -B_0^{-1}\left(I - F_0\left(F_0'B_0^{-1}F_0\right)^{-1}F_0'B_0^{-1}\right)W_0 - B_0^{-1}F_0\left(F_0'B_0^{-1}F_0\right)^{-1}U_0$. If W_0 and U_0 are 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 $\mathcal{J} = -H_0^{-1}\left(I - F_0\left(F_0'H_0^{-1}F_0\right)^{-1}F_0'H_0^{-1}\right)W_0 - H_0^{-1}F_0\left(F_0'H_0^{-1}F_0\right)^{-1}U_0$.

Remark 8. If strict complementarity holds, meaning $\lambda_{0j} > 0$ whenever $f_{0j}(\beta_0) = 0$, then $\mathcal{I}^* =$

$\mathcal{I}_+^*(\lambda_0)$ and the asymptotic distribution reduces down to

$$\mathcal{J} = \underset{h \in \Sigma}{\arg\min} \left\{ h'W_0 + \frac{1}{2}h'H_0h + \sum_{j \in \mathcal{E} \cup \mathcal{I}_+^*(\lambda_0)} \lambda_{0j} \left(h'V_{0j} + \frac{1}{2}h'G_{0j}h \right) \right\}$$

for $\Sigma = \left\{ h : U_{0j} + F_{0j}'h = 0 \text{ for } j \in \mathcal{E} \cup \mathcal{I}_+^*(\lambda_0) \right\}$. Just like in the previous remark, we can express $\mathcal{J} = -B_0^{-1}\left(I - F_0\left(F_0'B_0^{-1}F_0\right)^{-1}F_0'B_0^{-1}\right)\left(W_0 + \sum_{j \in \mathcal{E} \cup \mathcal{I}_+^*(\lambda_0)} \lambda_{0j}V_{0j}\right) - B_0^{-1}F_0\left(F_0'B_0^{-1}F_0\right)^{-1}U_0$, where $B_0 = H_0 + \sum_{j \in \mathcal{E} \cup \mathcal{I}_+^*(\lambda_0)} \lambda_{0j}G_{0j}$. If W_0, V_0, and U_0 are multivariate normal, then $\mathcal{J}$ will also be multivariate normal.

If $l(\beta_0) = 0$, then $\sum_{j \in \mathcal{E} \cup \mathcal{I}^*} \lambda_{0j}\left(h'V_{0j} + \frac{1}{2}h'G_{0j}h\right) = 0$ and $\mathcal{I}_+^*(\lambda_0) = \varnothing$, so $\mathcal{J}$ reduces down to $\mathcal{J} = \underset{h \in \Sigma}{\arg\min}\left\{h'W_0 + \frac{1}{2}h'H_0h\right\}$, for $\Sigma = \left\{h : U_{0j} + F_{0j}'h = 0 \text{ for } j \in \mathcal{E}\right\}$.

Remark 9. If there are only inequality constraints, we can also obtain a closed form expression for $\mathcal{J}$. Because $\Sigma = \left\{h : U_{0j} + F_{0j}'h = 0 \text{ for } j \in \mathcal{I}_+^*(\lambda_0), U_{0j} + F_{0j}'h \leqslant 0 \text{ for } j \in \mathcal{I}_0^*(\lambda_0)\right\}$ in this case, it follows from Lemma 5.2 in the Appendix that

$$\begin{aligned}\mathcal{J} = \max \Bigg\{ & -B_0^{-1}\left(I - F_{0+}\left(F_{0+}'B_0^{-1}F_{0+}\right)^{-1}F_{0+}'B_0^{-1}\right)\left(W_0 + \sum_{j \in \mathcal{I}_+^*(\lambda_0)} \lambda_{0j}V_{0j}\right) \\ & -B_0^{-1}F_{0+}\left(F_{0+}'B_0^{-1}F_{0+}\right)^{-1}U_{0+}, -B_0^{-1}\left(W_0 + \sum_{j \in \mathcal{I}_+^*(\lambda_0)} \lambda_{0j}V_{0j}\right)\Bigg\}\end{aligned}$$

where F_{0+} is the matrix of F_{0j} for $j \in \mathcal{I}_+^*(\lambda_0)$, U_{0+} is the vector of U_{0j} for $j \in \mathcal{I}_+^*(\lambda_0)$, and $B_0 = H_0 + \sum_{j \in \mathcal{I}_+^*(\lambda_0)} \lambda_{0j}G_{0j}$.

If there are no strongly active (binding) inequality constraints, meaning $\mathcal{I}_+^*(\lambda_0) = \varnothing$, then $\sum_{j \in \mathcal{I}_+^*(\lambda_0)} \lambda_{0j}G_{0j} = 0$, and $H_0 = B_0$, so the asymptotic distribution reduces down to $\mathcal{J} = -H_0^{-1}W_0$, which will be multivariate normal if W_0 is multivariate normal.

Remark 10. In the case of non-random constraints $f_n(\beta) = f_0(\beta)$ that do not depend on the data, if $l(\beta_0)$ may not be zero, the proximal bootstrap estimator is

$$\hat{\beta}_n^* = \underset{\beta \in C^*}{\arg\min}\ \alpha_n\sqrt{n}\left(\hat{l}_n^*(\bar{\beta}_n) - \hat{l}_n(\bar{\beta}_n)\right)'(\beta - \bar{\beta}_n) + \frac{1}{2}\left\|\beta - \bar{\beta}_n\right\|_{\bar{H}_n}^2 + \frac{1}{2}\sum_{j \in \mathcal{E} \cup \mathcal{I}} \bar{\lambda}_{nj}\left\|\beta - \bar{\beta}_n\right\|_{\bar{G}_{0j}}^2$$

$$C^* \equiv \left\{\beta \in \mathbb{B} : f_{0j}\left(\bar{\beta}_n\right) + \bar{F}_{0j}'\left(\beta - \bar{\beta}_n\right) = 0 \text{ for } j \in \mathcal{E}, f_{0j}\left(\bar{\beta}_n\right) + \bar{F}_{0j}'\left(\beta - \bar{\beta}_n\right) \leqslant 0 \text{ for } j \in \mathcal{I}\right\}$$

If $l(\beta_0) = 0$, which is implied by $Q(\beta) = Q(\beta_0) + \frac{1}{2}(\beta - \beta_0)' H_0 (\beta - \beta_0) + o\left(\|\beta - \beta_0\|^2\right)$, then the proximal bootstrap estimator can be defined as

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

$$C^* \equiv \left\{\beta \in \mathbb{B} : f_{0j}\left(\bar{\beta}_n\right) + \bar{F}_{0j}'\left(\beta - \bar{\beta}_n\right) = 0 \text{ for } j \in \mathcal{E}, f_{0j}\left(\bar{\beta}_n\right) + \bar{F}_{0j}'\left(\beta - \bar{\beta}_n\right) \leqslant 0 \text{ for } j \in \mathcal{I}\right\}$$

The asymptotic distribution when $l(\beta_0)$ may not be zero can be derived as follows:

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

Furthermore since $\sqrt{n} f_{nj}(\beta_0) = \sqrt{n} f_{0j}(\beta_0) = 0$ for all $j \in \mathcal{E} \cup \mathcal{I}^*$,

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

$$\Sigma = \left\{h : F_{0j}'h = 0 \text{ for } j \in \mathcal{E} \cup \mathcal{I}_+^*(\lambda_0), F_{0j}'h \leqslant 0 \text{ for } j \in \mathcal{I}_0^*(\lambda_0)\right\}$$

When $l(\beta_0) = 0$, since $\lambda_{0j} = 0$ for all $j \in \mathcal{E} \cup \mathcal{I}$,

$$\mathcal{J} = \arg\min_{h \in \Sigma}\left\{h'W_0 + \frac{1}{2}h'H_0h\right\}$$

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

Since LICQ is satisfied (which implies the Tangent cone $T_C(\beta_0)$ is equal to the linearized feasible set Σ), $\mathcal{J}$ is equivalent to the asymptotic distribution in Theorem 1. A special case of this is the constrained maximum likelihood example in Andrews (2000). He imposes a nonnegativity constraint

$\mu \geqslant 0$ for a normal mean model (with variance 1) and shows that the asymptotic distribution of the maximum likelihood estimator is $\mathcal{J} = \max\{Z, 0\}$ (where $Z \sim N(0,1)$) if the true mean equals 0. We can obtain this asymptotic distribution by setting $F_0 = -1$, $H_0 = 1$, and $W_0 = Z$.

Remark 11. Alternatively, we can define the proximal bootstrap estimator as $\hat{\beta}_n^* = \underset{\beta \in C^*}{\arg\min} \hat{A}_n^*(\beta)$, where

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

The feasible direction set is

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

and the linearized feasible direction set is, for some $\tilde{\beta}$ in between $\bar{\beta}_n$ and β_0,

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

Note that since $\frac{f_{nj}(\beta_0)}{\alpha_n} + F_{nj}\left(\tilde{\beta}\right)' h + \sqrt{n}\left(f_{nj}^*\left(\bar{\beta}_n\right) - f_{nj}\left(\bar{\beta}_n\right)\right) \xrightarrow{p} -\infty$ for $j \in \mathcal{I} \backslash \mathcal{I}^*$, the nonactive inequality constraints do not affect the asymptotic distribution, under the assumption of no drifting constraints. Since $F_{nj}\left(\tilde{\beta}\right) \xrightarrow{p} F_{0j}$ for all $j \in \mathcal{E} \cup \mathcal{I}$, $\frac{f_{nj}(\beta_0)}{\alpha_n} = \frac{\sqrt{n}(f_{nj}(\beta_0) - f_{0j}(\beta_0))}{\sqrt{n}\alpha_n} = o_P(1)$ for all $j \in \mathcal{E} \cup \mathcal{I}^*$, $\sqrt{n}\left(f_{nj}^*\left(\beta_0\right) - f_{nj}\left(\beta_0\right)\right) \underset{\mathbb{W}}{\overset{\mathbb{P}}{\rightsquigarrow}} U_{0j}$, jointly, for all $j \in \mathcal{E} \cup \mathcal{I}^*$, and $\sqrt{n}\left(f_{nj}^*\left(\bar{\beta}_n\right) - f_{nj}\left(\bar{\beta}_n\right)\right) \underset{\mathbb{W}}{\overset{\mathbb{P}}{\rightsquigarrow}} U_{0j}$, jointly, for all $j \in \mathcal{E} \cup \mathcal{I}^*$, it follows that $\infty 1\left(h \notin \Sigma_n^*\right) \underset{e-d}{\xrightarrow{p}} \infty 1\left(h \notin \Sigma\right)$.

Therefore, this nonlinearized bootstrap estimator has the same asymptotic distribution as the linearized version in Theorem 2.

Remark 12. The choice of α_n is a difficult problem. 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 bootstrap estimates $\hat{\beta}_n^{(*,b_1)}$ for $b_1 = 1...B_1$. Conditional on each of these bootstrap samples $b_1 = 1...B_1$, draw B_2 bootstrap samples and compute bootstrap estimates $\hat{\beta}_n^{(*,b_1,b_2)}$ for $b_2 = 1...B_2$. Pick some nominal frequency $1-\tau$. Define $\hat{c}^*_{1-\tau}$ to be the $1-\tau$ quantile of $\frac{\hat{\beta}_n^{(*,b_1,b_2)}-\hat{\beta}_n^{(*,b_1)}}{\alpha_n}$. Compute the empirical frequency with which equal-tailed intervals $\left[\hat{\beta}_n^{(*,b_1)} - \frac{\hat{c}^*_{1-\tau/2}}{\sqrt{n}}, \hat{\beta}_n^{(*,b_1)} - \frac{\hat{c}^*_{\tau/2}}{\sqrt{n}}\right]$ cover $\hat{\beta}_n$. 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. In the absence of drifting constraints, this procedure should find the optimal value of α_n that asymptotically achieves coverage closest to the nominal level.

2.5 Uniformity

In the case of inequality constraints that are drifting towards the boundary, the proximal bootstrap will typically not consistently replicate the estimator's asymptotic distribution; however we can still obtain a uniformly conservatively asymptotically valid confidence set for β_0. We use the fact that $n\left(\hat{\mathcal{L}}_n(\beta_0) - \hat{\mathcal{L}}_n\left(\hat{\beta}_n\right)\right) \rightsquigarrow \min_{h\in\Omega} q(h)$, where $q(h) \equiv -h'W_0 + \frac{1}{2}h'H_0h + \sum_{j\in\mathcal{E}\cup\mathcal{I}_+^*(\lambda_0)} \lambda_{0j}\left(-h'V_{0j} + \frac{1}{2}h'G_{0j}h\right)$, and $\frac{\hat{A}_n^*(\hat{\beta}_n)-\hat{A}_n^*(\hat{\beta}_n^*)}{\alpha_n^2} \underset{\mathbb{W}}{\overset{\mathbb{P}}{\rightsquigarrow}} \min_{h\in\Omega^*} q(h)$, where $\hat{A}_n^*(\beta)$ is defined in equation 1. We will show that $\Omega^* \subseteq \Omega$ for all drifting sequences, which implies that the asymptotic distribution of $\frac{\hat{A}_n^*(\hat{\beta}_n)-\hat{A}_n^*(\hat{\beta}_n^*)}{\alpha_n^2}$ uniformly first order stochastically dominates the asymptotic distribution of $n\left(\hat{\mathcal{L}}_n(\beta_0) - \hat{\mathcal{L}}_n\left(\hat{\beta}_n\right)\right)$. We show $\liminf_{n\to\infty} \inf_{P\in\mathcal{P}} P\left(n\left(\hat{\mathcal{L}}_n(\beta_0) - \hat{\mathcal{L}}_n\left(\hat{\beta}_n\right)\right) \leqslant \hat{c}^*_{1-\alpha}\right) \geqslant 1-\alpha$, which implies $\mathcal{C}^*_{1-\alpha} = \left\{\beta : n\left(\hat{\mathcal{L}}_n(\beta) - \hat{\mathcal{L}}_n\left(\hat{\beta}_n\right)\right) \leqslant \hat{c}^*_{1-\alpha}\right\}$ will be a uniformly conservatively valid nominal $1-\alpha$ confidence set for $\beta_0 \equiv \beta_0(P)$. The next theorem formalizes these arguments by drawing on insights from Chen et al. (2018).

Theorem 3. *Let $\mathcal{P}$ be a class of distributions for which Assumptions 1 and 3- 6 and Conditions (i)-(vi) of Theorem 2 are satisfied. For each $P \in \mathcal{P}$, let $J_n(\cdot, P)$ denote the CDF of $n\left(\hat{\mathcal{L}}_n(\beta_0) - \hat{\mathcal{L}}_n\left(\hat{\beta}_n\right)\right)$ under P, and assume $\lim_{n\to\infty} \sup_{P\in\mathcal{P}} \sup_{x\in\mathbb{R}} |J_n(x,P) - J(x,P)| = 0$, where the limiting distributions $\{J(\cdot,P) : P \in \mathcal{P}\}$ are equicontinuous at their $1-\alpha$ quantiles. Let $J^*_{\alpha_n}(\cdot,P)$ denote the conditional CDF of $\frac{\hat{A}_n^*(\hat{\beta}_n)-\hat{A}_n^*(\hat{\beta}_n^*)}{\alpha_n^2}$ under P, and assume for all $\epsilon > 0$, $\lim_{n\to\infty} \sup_{P\in\mathcal{P}} P\left(\sup_{x\in\mathbb{R}} \left|J^*_{\alpha_n}(x,P) - J^*(x,P)\right| > \epsilon\right) = 0$, where the limiting distributions $\{J^*(\cdot,P) : P \in \mathcal{P}\}$ are equicontinuous at their $1-\alpha$ quantiles.*

Then, for any sequence α_n such that $\alpha_n \to 0$ and $\sqrt{n}\alpha_n \to \infty$, $\liminf_{n\to\infty} \inf_{P\in\mathcal{P}} P\left(\beta_0(P) \in \mathcal{C}^*_{1-\alpha}\right) \geqslant 1-\alpha$, where $\mathcal{C}^*_{1-\alpha} = \left\{\beta : n\left(\hat{\mathcal{L}}_n(\beta) - \hat{\mathcal{L}}_n\left(\hat{\beta}_n\right)\right) \leqslant \hat{c}^*_{1-\alpha}\right\}$ and $\hat{c}^*_{1-\alpha}$ is the $1-\alpha$ quantile of $\frac{\hat{A}^*_n(\hat{\beta}_n) - \hat{A}^*_n(\hat{\beta}^*_n)}{\alpha_n^2}$.

Remark 13. If we would like to construct a nominal $1-\alpha$ confidence set for $\gamma_0 = a'\beta_0$, where a is a known unit vector, we can 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 from the uniform asymptotic validity of $\mathcal{C}^*_{1-\alpha}$.

Remark 14. In the case of $\lambda_{0j} = 0$ for all $j \in \mathcal{E} \cup \mathcal{I}$, which occurs when $l(\beta_0) = 0$, we can replace $\hat{\mathcal{L}}_n$ by $\hat{Q}_n$. The simultaneous confidence set becomes $\mathcal{C}^*_{1-\alpha} = \left\{\beta : n\left(\hat{Q}_n(\beta) - \hat{Q}_n\left(\hat{\beta}_n\right)\right) \leqslant \hat{c}^*_{1-\alpha}\right\}$, where $\hat{c}^*_{1-\alpha}$ is the $1-\alpha$ quantile of $\frac{\hat{A}^*_n(\hat{\beta}_n) - \hat{A}^*_n(\hat{\beta}^*_n)}{\alpha_n^2}$ and $\hat{A}^*_n(\beta) = \alpha_n\sqrt{n}\left(\hat{l}^*_n(\bar{\beta}_n) - \hat{l}_n(\bar{\beta}_n)\right)'(\beta - \bar{\beta}_n) + \frac{1}{2}\|\beta - \bar{\beta}_n\|^2_{\bar{H}_n}$.

Remark 15. We illustrate the intuition for this result by looking at an example with non-random constraints, some of which are drifting at the $\sqrt{n}$ rate to zero. Suppose we have equality constraints $\mathcal{E}$, active inequality constraints $\mathcal{I}^* = \{j \in \mathcal{I} : f_j(\beta_0) = 0\}$, and non-active inequality constraints that are drifting towards the boundary at a $\sqrt{n}$ rate: $\mathcal{I}^{\#}_{1/2} = \{j \in \mathcal{I} : f_j(\beta_0) = c/\sqrt{n}\}$, for some $c < 0$. We allow for other rates of drift in Theorem 3, but we do not present them here for simplicity. Suppose $l(\beta_0) = 0$ (we do not require this in Theorem 3). Then, $n\left(\hat{Q}_n(\beta_0) - \hat{Q}_n\left(\hat{\beta}_n\right)\right) \rightsquigarrow \min_{h\in\Omega} q(h)$ and $\frac{\hat{A}^*_n(\hat{\beta}_n) - \hat{A}^*_n(\hat{\beta}^*_n)}{\alpha_n^2} \underset{\mathbb{W}}{\overset{\mathbb{P}}{\rightsquigarrow}} \min_{h\in\Omega^*} q(h)$, where $\hat{A}^*_n(\beta) \equiv \alpha_n\sqrt{n}\left(\hat{l}^*_n(\bar{\beta}_n) - \hat{l}_n(\bar{\beta}_n)\right)'(\beta - \bar{\beta}_n) + \frac{1}{2}\|\beta - \bar{\beta}_n\|^2_{\bar{H}_n}$ and $q(h) = -h'W_0 + \frac{1}{2}h'H_0h$. For $F_{0j} \equiv \left.\frac{\partial f_j(\beta)}{\partial \beta}\right|_{\beta=\beta_0}$,

$$\Omega^* = \left\{h : -F'_{0j}h = 0 \text{ for } j \in \mathcal{E}, -F'_{0j}h \leqslant 0 \text{ for } j \in \mathcal{I}^* \cup \mathcal{I}^{\#}_{1/2}\right\}$$
$$\Omega = \left\{h : -F'_{0j}h = 0 \text{ for } j \in \mathcal{E}, -F'_{0j}h \leqslant 0 \text{ for } j \in \mathcal{I}^*, -F'_{0j}h \leqslant -c \text{ for } j \in \mathcal{I}^{\#}_{1/2}\right\}$$

Since $\Omega^* \subseteq \Omega$ and $q(h) = -h'W_0 + \frac{1}{2}h'H_0h$ is a strictly convex function of h when $H_0 > 0$, $\min_{h\in\Omega^*} q(h)$ first order stochastically dominates $\min_{h\in\Omega} q(h)$. Then, $\liminf_{n\to\infty} \inf_{P\in\mathcal{P}} P\left(n\left(\hat{Q}_n(\beta_0) - \hat{Q}_n\left(\hat{\beta}_n\right)\right) \leqslant \hat{c}^*_{1-\alpha}\right) \geqslant 1-\alpha$, which implies $\mathcal{C}^*_{1-\alpha} = \left\{\beta : n\left(\hat{Q}_n(\beta) - \hat{Q}_n\left(\hat{\beta}_n\right)\right) \leqslant \hat{c}^*_{1-\alpha}\right\}$ will be a uniformly conservatively valid nominal $1-\alpha$ confidence set for β_0.

Remark 16. In order to determine the optimal value of α_n when there are drifting constraints, we can change the procedure in Remark 12 to instead compute the empirical frequency with which $\left\{\beta : n\left(\hat{\mathcal{L}}_n(\beta) - \hat{\mathcal{L}}_n\left(\hat{\beta}_n^{(*,b_1)}\right)\right) \leqslant \hat{c}^*_{1-\tau}\right\}$ covers $\hat{\beta}_n$, where $\hat{c}^*_{1-\tau}$ is the $1-\tau$ quantile of $\frac{\hat{A}^*_n\left(\hat{\beta}_n^{(*,b_1)}\right) - \hat{A}^*_n\left(\hat{\beta}_n^{(*,b_1,b_2)}\right)}{\alpha_n^2}$ and $\hat{A}^*_n(\beta)$ is defined in equation 1. When $l(\beta_0) = 0$, we can use $\left\{\beta : n\left(\hat{Q}_n(\beta) - \hat{Q}_n\left(\hat{\beta}_n^{(*,b_1)}\right)\right) \leqslant \hat{c}^*_{1-\tau}\right\}$ and $\hat{A}^*_n(\beta) \equiv \alpha_n\sqrt{n}\left(\hat{l}^*_n(\bar{\beta}_n) - \hat{l}_n(\bar{\beta}_n)\right)'(\beta - \bar{\beta}_n) + \frac{1}{2}\|\beta - \bar{\beta}_n\|^2_{\bar{H}_n}$.

3 Monte Carlo Simulations

3.1 Boundary Constrained Maximum Likelihood

We consider a two sample location model with i.i.d data:

$$\begin{array}{l} y_{1i} = \beta_{01} + \epsilon_{1i} \\ y_{2i} = \beta_{02} + \epsilon_{2i} \end{array}, \quad \begin{pmatrix} \epsilon_{1i} \\ \epsilon_{2i} \end{pmatrix} \overset{i.i.d.}{\sim} N(0, I_2)$$

We would like to impose a non-positivity constraint on β_{01} and a non-negativity constraint on β_{02}:

$$\hat{\beta}_n = \begin{pmatrix} \hat{\beta}_{n1} \\ \hat{\beta}_{n2} \end{pmatrix} = \underset{\beta_1 \leqslant 0, \beta_2 \geqslant 0}{\arg\min} \frac{1}{2n}\left(\sum_{i=1}^n (y_{1i} - \beta_1)^2 + \sum_{i=1}^n (y_{2i} - \beta_2)^2\right)$$

We use Matlab's built-in fmincon solver to compute the original estimator $\bar{\beta}_n = \hat{\beta}_n$ and also the proximal bootstrap estimator $\hat{\beta}^*_n = \underset{\beta \in C}{\arg\min}\left\{\alpha_n\sqrt{n}\left(\hat{l}^*_n(\bar{\beta}_n) - \hat{l}_n(\bar{\beta}_n)\right)'(\beta - \bar{\beta}_n) + \frac{1}{2}\|\beta - \bar{\beta}_n\|^2_{\bar{H}_n}\right\}$, where $\bar{H}_n = I_2$ and $\hat{l}_n(\bar{\beta}_n) = \left[-(\bar{y}_{1n} - \bar{\beta}_{n1}), -(\bar{y}_{2n} - \bar{\beta}_{n2})\right]'$ for $\bar{y}_{1n} = \frac{1}{n}\sum_{i=1}^n y_{1i}$ and $\bar{y}_{2n} = \frac{1}{n}\sum_{i=1}^n y_{2i}$.

The goal of this simulation is to examine the coverage properties of the proximal bootstrap projection confidence intervals $CI^{Proj}_{1-\alpha} = \left[\inf_{\beta \in \mathcal{C}^*_{1-\alpha}} a'\beta, \sup_{\beta \in \mathcal{C}^*_{1-\alpha}} a'\beta\right]$, where a is either $(1,0)'$ or $(0,1)'$, and $\mathcal{C}^*_{1-\alpha} = \left\{\beta : n\left(\hat{Q}_n(\beta) - \hat{Q}_n\left(\hat{\beta}_n\right)\right) \leqslant \hat{c}^*_{1-\alpha}\right\}$, where $\hat{c}^*_{1-\alpha}$ is the $1-\alpha$ quantile of $\frac{\hat{A}^*_n(\hat{\beta}_n) - \hat{A}^*_n(\hat{\beta}^*_n)}{\alpha_n^2}$ and $\hat{A}^*_n(\beta) = \alpha_n\sqrt{n}\left(\hat{l}^*_n(\bar{\beta}_n) - \hat{l}_n(\bar{\beta}_n)\right)'(\beta - \bar{\beta}_n) + \frac{1}{2}\|\beta - \bar{\beta}_n\|^2_{\bar{H}_n}$. The true parameters β_{02} and $\beta_{01} = -\beta_{02}$ are drifting towards zero at five different rates: $\beta_{02} \in \left\{n^{-1/6}, n^{-1/4}, n^{-1/3}, n^{-1/2}, n^{-1}\right\}$. We consider three different sample sizes $n \in \{100, 500, 1000\}$ and seven different α_n's for each n: $\alpha_n \in \left\{n^{-1/2.1}, n^{-1/2.5}, n^{-1/3}, n^{-1/4}, n^{-1/6}, n^{-1/8}, n^{-1/10}\right\}$. Empirical coverage frequencies and average interval lengths (in parentheses) of nominal 95% confidence intervals for β_{01} are reported in Table 1 and those for β_{02} are reported in Table 2. We use 5000 bootstrap iterations and 2000

Monte Carlo simulations. The coverage frequencies are very close to 95% for all rates of drift except for n^{-1}, in which case the coverage is around 99% for the smaller values of α_n and around 98% for the larger values of α_n. The fact that smaller values of α_n can lead to overcoverage is also evident for the $n^{-1/2}$ drift rate.

Table 1: Proximal β_{01} Projection Interval Empirical Coverage Frequencies

	α_n	$n^{-1/2.1}$	$n^{-1/2.5}$	$n^{-1/3}$	$n^{-1/4}$	$n^{-1/6}$	$n^{-1/8}$	$n^{-1/10}$
$\beta_{01} = -n^{-1}$	$n = 100$	0.988	0.986	0.984	0.981	0.981	0.980	0.980
		(0.450)	(0.444)	(0.441)	(0.438)	(0.437)	(0.436)	(0.436)
	$n = 500$	0.992	0.989	0.987	0.986	0.985	0.984	0.984
		(0.198)	(0.195)	(0.194)	(0.193)	(0.193)	(0.192)	(0.192)
	$n = 1000$	0.988	0.986	0.984	0.983	0.982	0.982	0.982
		(0.137)	(0.135)	(0.134)	(0.133)	(0.133)	(0.133)	(0.133)
$\beta_{01} = -n^{-1/2}$	$n = 100$	0.988	0.987	0.984	0.980	0.977	0.976	0.976
		(0.444)	(0.434)	(0.426)	(0.418)	(0.413)	(0.412)	(0.411)
	$n = 500$	0.983	0.980	0.978	0.974	0.970	0.970	0.970
		(0.193)	(0.187)	(0.183)	(0.180)	(0.179)	(0.178)	(0.178)
	$n = 1000$	0.974	0.970	0.965	0.959	0.957	0.956	0.956
		(0.134)	(0.129)	(0.126)	(0.124)	(0.123)	(0.123)	(0.123)
$\beta_{01} = -n^{-1/3}$	$n = 100$	0.978	0.977	0.974	0.970	0.964	0.963	0.963
		(0.465)	(0.455)	(0.443)	(0.427)	(0.416)	(0.413)	(0.411)
	$n = 500$	0.978	0.976	0.971	0.966	0.959	0.958	0.957
		(0.206)	(0.201)	(0.193)	(0.183)	(0.179)	(0.178)	(0.177)
	$n = 1000$	0.966	0.962	0.956	0.948	0.941	0.939	0.939
		(0.144)	(0.140)	(0.133)	(0.126)	(0.123)	(0.122)	(0.122)
$\beta_{01} = -n^{-1/4}$	$n = 100$	0.982	0.981	0.979	0.976	0.969	0.965	0.963
		(0.475)	(0.470)	(0.461)	(0.443)	(0.426)	(0.420)	(0.417)
	$n = 500$	0.981	0.981	0.978	0.973	0.960	0.959	0.959
		(0.209)	(0.208)	(0.204)	(0.193)	(0.183)	(0.180)	(0.179)
	$n = 1000$	0.978	0.978	0.975	0.963	0.950	0.945	0.945
		(0.144)	(0.144)	(0.142)	(0.133)	(0.126)	(0.124)	(0.124)
$\beta_{01} = -n^{-1/6}$	$n = 100$	0.986	0.985	0.985	0.980	0.976	0.974	0.972
		(0.477)	(0.477)	(0.474)	(0.462)	(0.443)	(0.433)	(0.428)
	$n = 500$	0.983	0.983	0.983	0.981	0.974	0.970	0.969
		(0.209)	(0.209)	(0.209)	(0.205)	(0.193)	(0.187)	(0.184)
	$n = 1000$	0.976	0.976	0.976	0.975	0.962	0.955	0.951
		(0.145)	(0.145)	(0.145)	(0.143)	(0.133)	(0.129)	(0.127)

In contrast to the moderately conservative coverage of the proximal bootstrap confidence sets, standard bootstrap confidence intervals produce severe undercoverage for several rates of drift. Table 3 shows the empirical coverage frequencies and average interval lengths (in parentheses) of standard multinomial bootstrap two-sided equal-tailed confidence intervals, using 5000 bootstrap

Table 2: Proximal β_{02} Projection Interval Empirical Coverage Frequencies

	α_n	$n^{-1/2.1}$	$n^{-1/2.5}$	$n^{-1/3}$	$n^{-1/4}$	$n^{-1/6}$	$n^{-1/8}$	$n^{-1/10}$
$\beta_{02} = n^{-1}$	$n = 100$	0.993	0.990	0.986	0.983	0.982	0.982	0.982
		(0.449)	(0.444)	(0.441)	(0.438)	(0.436)	(0.436)	(0.436)
	$n = 500$	0.989	0.987	0.984	0.981	0.981	0.980	0.980
		(0.198)	(0.195)	(0.194)	(0.193)	(0.193)	(0.192)	(0.192)
	$n = 1000$	0.990	0.989	0.986	0.984	0.984	0.984	0.983
		(0.137)	(0.135)	(0.134)	(0.133)	(0.133)	(0.133)	(0.133)
$\beta_{02} = n^{-1/2}$	$n = 100$	0.990	0.988	0.984	0.981	0.977	0.977	0.976
		(0.444)	(0.434)	(0.425)	(0.418)	(0.414)	(0.412)	(0.411)
	$n = 500$	0.979	0.977	0.974	0.969	0.966	0.965	0.964
		(0.193)	(0.187)	(0.183)	(0.180)	(0.179)	(0.178)	(0.178)
	$n = 1000$	0.970	0.967	0.965	0.960	0.958	0.958	0.958
		(0.134)	(0.129)	(0.126)	(0.124)	(0.123)	(0.123)	(0.123)
$\beta_{02} = n^{-1/3}$	$n = 100$	0.977	0.974	0.972	0.967	0.963	0.960	0.960
		(0.465)	(0.455)	(0.443)	(0.427)	(0.416)	(0.413)	(0.411)
	$n = 500$	0.970	0.966	0.964	0.958	0.952	0.950	0.949
		(0.206)	(0.201)	(0.193)	(0.183)	(0.179)	(0.178)	(0.177)
	$n = 1000$	0.967	0.964	0.957	0.953	0.947	0.947	0.946
		(0.144)	(0.140)	(0.133)	(0.126)	(0.123)	(0.122)	(0.122)
$\beta_{02} = n^{-1/4}$	$n = 100$	0.982	0.981	0.978	0.973	0.968	0.965	0.964
		(0.475)	(0.470)	(0.461)	(0.443)	(0.426)	(0.420)	(0.417)
	$n = 500$	0.974	0.973	0.970	0.964	0.956	0.953	0.952
		(0.209)	(0.208)	(0.204)	(0.193)	(0.183)	(0.180)	(0.179)
	$n = 1000$	0.976	0.975	0.974	0.961	0.952	0.948	0.945
		(0.145)	(0.144)	(0.142)	(0.134)	(0.126)	(0.124)	(0.124)
$\beta_{02} = n^{-1/6}$	$n = 100$	0.986	0.984	0.984	0.981	0.973	0.969	0.969
		(0.477)	(0.477)	(0.474)	(0.462)	(0.443)	(0.433)	(0.428)
	$n = 500$	0.975	0.975	0.975	0.973	0.966	0.963	0.960
		(0.209)	(0.209)	(0.209)	(0.205)	(0.193)	(0.187)	(0.184)
	$n = 1000$	0.976	0.976	0.976	0.975	0.964	0.956	0.953
		(0.145)	(0.145)	(0.145)	(0.143)	(0.134)	(0.129)	(0.127)

iterations and 2000 Monte Carlo simulations. Especially for the quicker rates of drift, the coverage can be far below 95%. We also examined one-sided intervals and they were also not able to get close to 95% coverage for both parameters. The average interval lengths of the proximal bootstrap projection intervals are wider than the standard bootstrap intervals, but the difference becomes less pronounced as the sample size increases.

Table 3: Standard Bootstrap Equal-tailed Empirical Coverage Frequencies

n	100	500	1000	5000	10000
$\beta_0 = \pm n^{-1}$	0.493	0.494	0.521	0.582	0.684
	(0.204)	(0.090)	(0.063)	(0.027)	(0.020)
	0.495	0.491	0.490	0.608	0.694
	(0.205)	(0.090)	(0.063)	(0.028)	(0.020)
$\beta_0 = \pm n^{-1/2}$	0.659	0.674	0.672	0.653	0.665
	(0.286)	(0.129)	(0.091)	(0.040)	(0.029)
	0.664	0.672	0.651	0.673	0.673
	(0.285)	(0.129)	(0.091)	(0.041)	(0.029)
$\beta_0 = \pm n^{-1/3}$	0.835	0.900	0.909	0.948	0.943
	(0.359)	(0.170)	(0.122)	(0.055)	(0.039)
	0.827	0.912	0.917	0.953	0.953
	(0.358)	(0.170)	(0.122)	(0.055)	(0.039)
$\beta_0 = \pm n^{-1/4}$	0.911	0.946	0.949	0.950	0.943
	(0.384)	(0.175)	(0.124)	(0.055)	(0.039)
	0.909	0.957	0.947	0.956	0.954
	(0.383)	(0.175)	(0.124)	(0.055)	(0.039)
$\beta_0 = \pm n^{-1/6}$	0.951	0.946	0.950	0.949	0.943
	(0.389)	(0.175)	(0.124)	(0.055)	(0.039)
	0.942	0.956	0.947	0.955	0.954
	(0.389)	(0.175)	(0.124)	(0.055)	(0.039)

3.2 Boundary Constrained Nonsmooth GMM

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

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

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:

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

We use Matlab's built-in fmincon solver to compute $\bar{\beta}_n = \hat{\beta}_n$ and also
$\hat{\beta}^*_n = \underset{\beta \in C}{\arg\min}\left\{\alpha_n \sqrt{n}\left(\hat{l}^*_n(\bar{\beta}_n) - \hat{l}_n(\bar{\beta}_n)\right)'(\beta - \bar{\beta}_n) + \frac{1}{2}\|\beta - \bar{\beta}_n\|^2_{\bar{H}_n}\right\}$, where $\bar{H}_n = \hat{G}_n'\hat{G}_n + \hat{L}_n'\hat{\pi}_n(\bar{\beta}_n)$,

$\hat{l}_n(\bar{\beta}_n) = \hat{G}_n' \hat{\pi}_n(\bar{\beta}_n)$, $\hat{l}_n^*(\bar{\beta}_n) = \hat{G}_n^{*\prime} \hat{\pi}_n^*(\bar{\beta}_n)$, 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 five different sample sizes $n \in \{100, 500, 1000, 5000, 10000\}$ and three different α_n's for each n: $\alpha_n \in \left\{n^{-1/3}, n^{-1/4}, n^{-1/6}, n^{-1/8}, n^{-1/10}\right\}$. We use 5000 bootstrap iterations and 2000 Monte Carlo simulations. Empirical coverage frequencies and average interval lengths for two-sided equal-tailed nominal 95% proximal bootstrap confidence intervals $\left[\hat{\beta}_n - \frac{\hat{c}^*_{0.975}}{\sqrt{n}}, \hat{\beta}_n - \frac{\hat{c}^*_{0.025}}{\sqrt{n}}\right]$, where $\hat{c}^*_\tau$ is the τth-percentile of $\frac{\hat{\beta}^*_n - \hat{\beta}_n}{\alpha_n}$, are reported in Table 4. There is slight overcoverage but the intervals are not particularly wide. For this example, there is practically no difference in coverage for the different values of α_n.

Table 4: Proximal Bootstrap Equal-Tailed Coverage Frequencies and Interval Lengths, $\beta_0 = 0$

n	100	500	1000	5000	10000
$\alpha_n = n^{-1/3}$	0.969	0.975	0.971	0.972	0.968
	(0.216)	(0.095)	(0.067)	(0.029)	(0.021)
$\alpha_n = n^{-1/4}$	0.969	0.975	0.971	0.972	0.968
	(0.210)	(0.092)	(0.065)	(0.029)	(0.020)
$\alpha_n = n^{-1/6}$	0.969	0.975	0.971	0.972	0.968
	(0.206)	(0.091)	(0.064)	(0.028)	(0.020)
$\alpha_n = n^{-1/8}$	0.969	0.975	0.971	0.972	0.968
	(0.204)	(0.090)	(0.063)	(0.028)	(0.020)
$\alpha_n = n^{-1/10}$	0.969	0.975	0.971	0.972	0.968
	(0.203)	(0.090)	(0.063)	(0.028)	(0.020)

We now compare the proximal bootstrap with the centered standard bootstrap estimator $\hat{\beta}_n^{**} = \underset{\beta \in C}{\arg\min} \left(\hat{\pi}_n^*(\beta) - \hat{\pi}_n\left(\hat{\beta}_n\right)\right)' \left(\hat{\pi}_n^*(\beta) - \hat{\pi}_n\left(\hat{\beta}_n\right)\right)$. Empirical coverage frequencies for equal-tailed nominal 95% confidence intervals and average interval lengths are reported in Table 5. Interestingly, the coverage frequencies are similar, although the intervals are wider.

Table 5: Standard Bootstrap Equal-Tailed Coverage Frequencies and Interval Lengths, $\beta_0 = 0$

n	100	500	1000	5000	10000
	0.968	0.976	0.974	0.974	0.967
	(0.236)	(0.107)	(0.076)	(0.034)	(0.024)

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

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

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

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

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

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 use Matlab's built-in fmincon solver to compute $\bar{\beta}_n = \hat{\beta}_n$ and also $\hat{\beta}_n^* \equiv \arg\min_{\beta \in C^*} \hat{A}_n^*(\beta)$, where

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

We use analytic expressions for the components in the proximal bootstrap objective function and constraints:

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

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

$$F_{nj}(\beta) = \frac{1}{n}\sum_{i=1}^{n}\frac{\partial P_{ij}}{\partial \beta} = \frac{1}{n}\sum_{i=1}^{n} P_{ij}\frac{\partial \ln P_{ij}}{\partial \beta} = \frac{1}{n}\sum_{i=1}^{n} P_{ij}\left(x_{ij} - \sum_l P_{il}x_{il}\right)$$

$$\begin{aligned} G_{nj}(\beta) = \frac{1}{n}\sum_{i=1}^{n}\frac{\partial^2 P_{ij}}{\partial \beta \partial \beta'} &= \frac{1}{n}\sum_{i=1}^{n}\frac{\partial P_{ij}}{\partial \beta}\left(x_{ij} - \sum_l P_{il}x_{il}\right)' - \frac{1}{n}\sum_{i=1}^{n} P_{ij}\sum_l \frac{\partial P_{il}}{\partial \beta}x_{il}' \\ &= \frac{1}{n}\sum_{i=1}^{n} P_{ij}\left(x_{ij} - \sum_l P_{il}x_{il}\right)\left(x_{ij} - \sum_l P_{il}x_{il}\right)' \\ &\quad - \frac{1}{n}\sum_{i=1}^{n}\sum_l P_{ij}P_{il}\left(x_{il} - \sum_m P_{im}x_{im}\right)x_{il}' \end{aligned}$$

Because $l(\beta_0) = 0$ in this model, we can in principle also use an alternative formulation of the proximal bootstrap with $\hat{A}_n^*(\beta) = \alpha_n\sqrt{n}\left(\hat{l}_n^*(\bar{\beta}_n) - \hat{l}_n(\bar{\beta}_n)\right)'(\beta - \bar{\beta}_n) + \frac{1}{2}\|\beta - \bar{\beta}_n\|_{\bar{H}_n}^2$. However we found that especially for the smaller sample sizes, including the term involving the Lagrange multipliers $\bar{\lambda}_{nj}$ helps with the coverage. We consider $n \in \{100, 500, 1000\}$, $J = 20$, and $\alpha_n \in \left\{n^{-1/3}, n^{-1/4}, n^{-1/6}, n^{-1/8}, n^{-1/10}\right\}$. Empirical coverage frequencies for equal-tailed nominal 95% confidence intervals and average interval lengths are reported in table 6. We use $B = 2000$ bootstrap iterations and $R = 1000$ Monte Carlo simulations. While the proximal bootstrap intervals undercover somewhat for smaller values of n and α_n, the coverage is very close to the nominal level for $n = 2000$ and larger values of α_n. We also consider larger values of J. Proximal bootstrap empirical coverage frequencies for equal-tailed nominal 95% confidence intervals and average interval lengths are reported in table 7. The results are computed using $B = 2000, R = 1000$. The coverage is slightly below the nominal level for $n = 100$ but very close to the nominal level for $n = 500$. Standard bootstrap empirical coverage frequencies for equal-tailed nominal 95% confidence intervals and average interval lengths are reported in table 8. The standard bootstrap undercovers, and its coverage is less than that of the proximal bootstrap for all values of n. The standard bootstrap intervals are also wider than the proximal bootstrap intervals for smaller values of n.

Table 6: Proximal Bootstrap Empirical Coverage Frequencies and Average Interval Lengths

	$n = 100$	$n = 500$	$n = 1000$	$n = 2000$
	$J = 20$	$J = 20$	$J = 20$	$J = 20$
$\alpha_n = n^{-1/3}$	0.917	0.923	0.946	0.929
	(0.0016)	(0.0007)	(0.0005)	(0.0004)
$\alpha_n = n^{-1/4}$	0.924	0.935	0.952	0.946
	(0.0016)	(0.0007)	(0.0005)	(0.0004)
$\alpha_n = n^{-1/6}$	0.922	0.939	0.952	0.951
	(0.0016)	(0.0007)	(0.0005)	(0.0004)
$\alpha_n = n^{-1/8}$	0.922	0.927	0.945	0.953
	(0.0015)	(0.0007)	(0.0005)	(0.0004)
$\alpha_n = n^{-1/10}$	0.918	0.920	0.944	0.952
	(0.0015)	(0.0007)	(0.0005)	(0.0004)

Table 7: Proximal Bootstrap Empirical Coverage Frequencies and Average Interval Lengths

	$n = 100$	$n = 500$	$n = 100$	$n = 500$
	$J = 50$	$J = 50$	$J = 100$	$J = 100$
$\alpha_n = n^{-1/3}$	0.928	0.936	0.932	0.945
	(0.0007)	(0.0003)	(0.0006)	(0.0003)
$\alpha_n = n^{-1/4}$	0.936	0.940	0.939	0.949
	(0.0007)	(0.0003)	(0.0006)	(0.0003)
$\alpha_n = n^{-1/6}$	0.941	0.946	0.944	0.954
	(0.0007)	(0.0003)	(0.0006)	(0.0003)
$\alpha_n = n^{-1/8}$	0.939	0.946	0.945	0.950
	(0.0007)	(0.0003)	(0.0006)	(0.0003)
$\alpha_n = n^{-1/10}$	0.938	0.947	0.947	0.950
	(0.0007)	(0.0003)	(0.0006)	(0.0003)

3.4 Rust (1987) Bus Engine Replacement Model

We apply our method to conduct inference for the Mathematical Programming with Equilibrium Constraints (MPEC) formulation of the Rust (1987) Bus Engine Replacement model. Su and Judd (2012) indicate that the MPEC estimator can be bootstrapped, although they do not provide an analysis of the empirical coverage frequencies of bootstrap confidence intervals. We find that our proximal bootstrap method performs equally good in terms of coverage and is more than twice as fast as the standard bootstrap.

Using the code accompanying Su and Judd (2012), we generate data using the following parameters used in their paper: discount factor $\beta = 0.975$ which is assumed to be known by the researcher and thus not estimated, replacement cost $RC = 11.7257$, operating cost parameter $\theta_1 = 2.4569$,

Table 8: Standard Bootstrap Empirical Coverage Frequencies and Average Interval Lengths

	$n = 100$	$n = 500$	$n = 1000$	$n = 2000$
	$J = 20$	$J = 20$	$J = 20$	$J = 20$
$B = 2000$	0.922	0.911	0.919	0.909
$R = 1000$	(0.0018)	(0.0008)	(0.0006)	(0.0004)
$B = 2000$	0.926	0.907	0.910	0.910
$R = 2000$	(0.0018)	(0.0008)	(0.0006)	(0.0004)
	$n = 100$	$n = 500$	$n = 100$	$n = 500$
	$J = 50$	$J = 50$	$J = 100$	$J = 100$
$B = 2000$	0.922	0.928	0.911	0.940
$R = 1000$	(0.0008)	(0.0004)	(0.0007)	(0.0003)
$B = 2000$	0.929	0.926	0.921	0.934
$R = 2000$	(0.0008)	(0.0004)	(0.0007)	(0.0003)

and transition probabilities $\theta_3' = \begin{pmatrix} 0.0937, & 0.4475, & 0.4459, & 0.0127, & 0.0002 \end{pmatrix}$. The MPEC objective function is a log likelihood which is a function of both the structural parameters and the choice-specific value functions $EV(x,d)$ given the data $\left(\left(x_t^i, d_t^i\right)_{t=1}^T\right)_{i=1}^M$, where x_t^i is the mileage of bus i in period t and d_t^i is an indicator for whether bus i's engine is replaced in period t.

$$\begin{aligned}\mathcal{L}(\theta_1, \theta_3, RC, EV) = {} & \frac{1}{M}\sum_{i=1}^{M}\sum_{t=2}^{T}\log\left(\frac{\exp\left[\nu\left(x_t^i, d_t^i; \theta_1, RC\right) + \beta EV\left(x_t^i, d_t^i\right)\right]}{\sum_{d'\in\{0,1\}}\exp\left[\nu\left(x_t^i, d'; \theta_1, RC\right) + \beta EV\left(x_t^i, d'\right)\right]}\right) \\ & + \frac{1}{M}\sum_{i=1}^{M}\sum_{t=2}^{T}\log\left(p_3\left(x_t^i \middle| x_{t-1}^i, d_{t-1}^i, \theta_3\right)\right)\end{aligned}$$

The constraints are the fixed point equations defining the discretized choice-specific value functions $EV(x,d)$ for mileage constrained to lie on a grid $\hat{\mathbf{x}} = \{\hat{x}_1, \hat{x}_2, ..., \hat{x}_K\}$:

$$EV(\hat{x}_k, d) = \sum_{x'}\log\left(\sum_{d'\in\{0,1\}}\exp\left[\nu\left(x', d'; \theta_1, RC\right) + \beta EV\left(x', d'\right)\right]\right)p_3\left(x' \middle| \hat{x}_k, d, \theta_3\right) \tag{3}$$

Given the current state $\hat{x}_k$, the next period mileage $x' \in \{\hat{x}_k, \hat{x}_{k+1}, \hat{x}_{k+2}, \hat{x}_{k+3}, \hat{x}_{k+4}\}$ can move up at most 4 grid points if the engine is not replaced. If the engine is replaced, the mileage first resets to $\hat{x}_1$ before transitioning to a different mileage. Su and Judd (2012)'s code chooses the mileage grid to be $\hat{\mathbf{x}} = \{1, 2, 3, ..., 175\}$. The utility function in their code is defined as

$$\nu(x, d; \theta_1, RC) = \begin{cases} -0.001x\theta_1 & , d = 0 \\ -RC - 0.001\theta_1 & , d = 1 \end{cases}$$

If the engine is replaced, the transition probabilities are $p_3\left(x' = \hat{x}_{1+j} \middle| \hat{x}_k, 1, \theta_3\right) = \theta_{3j}$. If the engine is not replaced, the transition probabilities are $p_3\left(x' = \hat{x}_{k+j} \middle| \hat{x}_k, 0, \theta_3\right) = \theta_{3j}$. The only values of the choice-specific value functions we need to estimate are the ones corresponding to no replacement $EV = [EV(\hat{x}_1, 0), EV(\hat{x}_2, 0), ..., EV(\hat{x}_K, 0)]$ because $EV(\hat{x}_k, 1) = EV(\hat{x}_1, 0)$ for all k, as pointed out in footnote 9 of Su and Judd (2012). Notice that because the mileage grid is fixed, the constraints do not depend on the data $\left(\left(x_t^i, d_t^i\right)_{t=1}^T\right)_{i=1}^M$. Define $\theta \equiv (\theta_1, \theta_3', RC, EV)'$ and $C = \{f_j(\theta) = 0 \text{ for } j \in \mathcal{E}, f_j(\theta) \leqslant 0 \text{ for } j \in \mathcal{I}\}$, where $f_j(\theta)$ includes the EV fixed point equations (3) as well as the constraints on the transition probabilities satisfying $0 \leqslant \theta_3 \leqslant 1$ and $\sum_j \theta_{3j} = 1$. Because our asymptotics are large M, fixed T, the rate of convergence of our estimator is $\sqrt{M}$. For some $\alpha_M \to 0$ and $\sqrt{M}\alpha_M \to \infty$, and a $\sqrt{M}$-consistent estimator $\bar{\theta}_M$, the proximal bootstrap estimator is given by

$$\hat{\theta}_M^* \equiv \underset{\theta \in C*}{\arg\min}\ \alpha_M \sqrt{M} \left(\hat{l}_M^*\left(\bar{\theta}_M\right) - \hat{l}_n\left(\bar{\theta}_M\right)\right)' \left(\theta - \bar{\theta}_M\right) + \frac{1}{2} \left\|\theta - \bar{\theta}_M\right\|_{\bar{H}_M}^2$$

$$C^* = \left\{f_j\left(\bar{\theta}_M\right) + F_j'\left(\theta - \bar{\theta}_M\right) = 0 \text{ for } j \in \mathcal{E}, f_j\left(\bar{\theta}_M\right) + F_j'\left(\theta - \bar{\theta}_M\right) \leqslant 0 \text{ for } j \in \mathcal{I}\right\}$$

We follow Su and Judd (2012) and use Knitro to compute $\bar{\theta}_M = \hat{\theta}_M$ as well as $\hat{\theta}_M^*$, although in principle the built-in Matlab nonlinear optimization solvers should also find the solution given enough time to search the parameter space. Because $l(\theta_0) = 0$ in this model, we do not need to include the Lagrange multiplier term in the objective function.

Tables 9-11 show the empirical coverage frequencies and average interval lengths for two-sided equal tailed nominal 95% proximal bootstrap confidence intervals computed using $B = 1000$ bootstrap iterations and $R = 2000$ Monte Carlo simulations. We consider 6 different values of $M \in \{500, 1000, 2000, 4000, 5000, 6000\}$ and three different values of $\alpha_M \in \left\{M^{-1/3}, M^{-1/4}, M^{-1/6}\right\}$. The number of time periods is $T = 120$. Most of the parameters have coverage very close to the nominal level for sufficiently large values of M, and the coverage is very similar for the three different values of α_M. Due to time constraints on the server, we were unable to obtain results for the standard bootstrap using the same values of M, B, and R, but the results should be similar given that the standard bootstrap is consistent in this example.

Table 9: Proximal Bootstrap Coverage Frequencies and Average Interval Lengths for $\alpha_M = M^{-1/3}$

M	500	1000	2000	4000	5000	6000
θ_1	0.925	0.946	0.949	0.942	0.948	0.949
	(0.520)	(0.373)	(0.264)	(0.187)	(0.167)	(0.152)
θ_{30}	0.951	0.947	0.945	0.933	0.932	0.935
	(0.005)	(0.003)	(0.002)	(0.002)	(0.001)	(0.001)
θ_{31}	0.955	0.944	0.951	0.948	0.94	0.947
	(0.008)	(0.006)	(0.004)	(0.003)	(0.003)	(0.002)
θ_{32}	0.949	0.952	0.944	0.942	0.942	0.952
	(0.008)	(0.006)	(0.004)	(0.003)	(0.003)	(0.002)
θ_{33}	0.957	0.95	0.949	0.951	0.96	0.957
	(0.002)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)
RC	0.927	0.95	0.949	0.946	0.946	0.947
	(1.683)	(1.204)	(0.853)	(0.604)	(0.540)	(0.492)

Table 10: Proximal Bootstrap Coverage Frequencies and Average Interval Lengths for $\alpha_M = M^{-1/4}$

M	500	1000	2000	4000	5000	6000
θ_1	0.923	0.949	0.95	0.941	0.949	0.95
	(0.520)	(0.372)	(0.264)	(0.187)	(0.167)	(0.153)
θ_{30}	0.952	0.948	0.94	0.935	0.934	0.937
	(0.005)	(0.003)	(0.002)	(0.002)	(0.001)	(0.001)
θ_{31}	0.954	0.942	0.95	0.946	0.944	0.948
	(0.008)	(0.006)	(0.004)	(0.003)	(0.003)	(0.002)
θ_{32}	0.952	0.95	0.941	0.943	0.939	0.948
	(0.008)	(0.006)	(0.004)	(0.003)	(0.003)	(0.002)
θ_{33}	0.959	0.95	0.95	0.949	0.958	0.958
	(0.002)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)
RC	0.927	0.952	0.95	0.945	0.949	0.952
	(1.683)	(1.204)	(0.853)	(0.604)	(0.540)	(0.493)

Table 11: Proximal Bootstrap Coverage Frequencies and Average Interval Lengths for $\alpha_M = M^{-1/6}$

M	500	1000	2000	4000	5000	6000
θ_1	0.924	0.947	0.949	0.943	0.948	0.95
	(0.520)	(0.372)	(0.264)	(0.187)	(0.167)	(0.152)
θ_{30}	0.952	0.95	0.941	0.933	0.933	0.94
	(0.005)	(0.003)	(0.002)	(0.002)	(0.001)	(0.001)
θ_{31}	0.955	0.942	0.949	0.949	0.944	0.953
	(0.008)	(0.006)	(0.004)	(0.003)	(0.003)	(0.002)
θ_{32}	0.951	0.951	0.942	0.944	0.94	0.946
	(0.008)	(0.006)	(0.004)	(0.003)	(0.003)	(0.002)
θ_{33}	0.96	0.951	0.949	0.951	0.959	0.954
	(0.002)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)
RC	0.925	0.951	0.95	0.947	0.948	0.944
	(1.666)	(1.201)	(0.852)	(0.603)	(0.540)	(0.493)

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 applicability in a boundary constrained GMM problem, a conditional logit model with capacity constraints, and a MPEC formulation of the Rust (1987) model.

5 Appendix

5.1 Proofs of Theorems

5.1.1 Proof of Theorem 1

Assumption 1 implies that $\hat{\beta}_n \xrightarrow{p} \beta_0 = \arg\min_{\beta \in C} Q(\beta)$ (see e.g. Corollary 3.2.3 in van der Vaart and Wellner (1996)). Assumption 2, $Q(\beta) = Q(\beta_0) + \frac{1}{2}(\beta - \beta_0)' H_0 (\beta - \beta_0) + o\left(\|\beta - \beta_0\|^2\right)$, and $\hat{\beta}_n \xrightarrow{p} \beta_0$ imply that the conditions of Lemma 4.3 in Geyer (1994) are satisfied, and therefore $\sqrt{n}\left(\hat{\beta}_n - \beta_0\right) = O_P(1)$.

To derive its asymptotic distribution, use the centered and scaled parameter $h = \sqrt{n}(\beta - \beta_0)$:

$$\begin{aligned}
\sqrt{n}\left(\hat{\beta}_n - \beta_0\right) &= \underset{h \in \sqrt{n}(C-\beta_0)}{\arg\min} \left\{ n\hat{Q}_n\left(\beta_0 + \frac{h}{\sqrt{n}}\right) - n\hat{Q}_n\left(\beta_0\right) \right\} \\
&= \underset{h \in \sqrt{n}(C-\beta_0)}{\arg\min} \left\{ h'\sqrt{n}\left(\hat{l}_n\left(\beta_0\right) - l\left(\beta_0\right)\right) + \frac{1}{2}h'H_0 h + o_P(1) \right\}
\end{aligned}$$

The second line is due to the uniform in h local quadratic expansion of $n\hat{Q}_n\left(\beta_0 + \frac{h}{\sqrt{n}}\right) - n\hat{Q}_n\left(\beta_0\right)$, which follows from Assumption 2.

Then Assumption 3 implies the Lindeberg Condition is satisfied and $\sqrt{n}\left(P_n - P\right) g\left(\cdot, \beta_0\right) \rightsquigarrow W_0$. Therefore,

$$n\hat{Q}_n\left(\beta_0 + \frac{h}{\sqrt{n}}\right) - n\hat{Q}_n\left(\beta_0\right) \rightsquigarrow h'W_0 + \frac{1}{2}h'H_0 h$$

as a process indexed by h in the space of bounded functions on compact sets $\ell^\infty(K)$ for any compact $K \subset \mathbb{R}^d$. Since $h'W_0 + \frac{1}{2}h'H_0 h$ has a continuous sample path, according to page 5 of Knight (1999),

$$n\hat{Q}_n\left(\beta_0 + \frac{h}{\sqrt{n}}\right) - n\hat{Q}_n\left(\beta_0\right) \rightarrow_{u-d} h'W_0 + \frac{1}{2}h'H_0 h$$

where $\rightarrow_{u-d}$ denotes convergence in distribution with respect to the topology of uniform convergence on compact sets. Chernoff Regularity implies that

$$\infty 1\left(h \notin \sqrt{n}\left(C - \beta_0\right)\right) \xrightarrow{e} \infty 1\left(h \notin T_C\left(\beta_0\right)\right)$$

where $\xrightarrow{e}$ denotes epigraphical convergence as defined in Geyer (1994), page 1997. Therefore, by Theorem 4 of Knight (1999),

$$n\hat{Q}_n\left(\beta_0 + \frac{h}{\sqrt{n}}\right) - n\hat{Q}_n\left(\beta_0\right) + \infty 1\left(h \notin \sqrt{n}\left(C - \beta_0\right)\right) \rightarrow_{e-d} h'W_0 + \frac{1}{2}h'H_0 h + \infty 1\left(h \notin T_C\left(\beta_0\right)\right)$$

where $\rightarrow_{e-d}$ denotes epi-convergence in distribution as defined on page 5 of Knight (1999). Then by Theorem 1 of Knight (1999), whose conditions are satisfied because $h'W_0 + \frac{1}{2}h'H_0 h$ almost surely has a unique minimizer over $T_C\left(\beta_0\right)$ due to C being a closed set (see Proposition 4.2 and Theorem

4.4 of Geyer (1994)),

$$\sqrt{n}\left(\hat{\beta}_n-\beta_0\right)=\underset{h\in\mathbb{R}^d}{\arg\min}\left\{n\hat{Q}_n\left(\beta_0+\frac{h}{\sqrt{n}}\right)-n\hat{Q}_n\left(\beta_0\right)+\infty 1\left(h\notin\sqrt{n}\left(C-\beta_0\right)\right)\right\}$$
$$\rightsquigarrow\underset{h\in\mathbb{R}^d}{\arg\min}\left\{h'W_0+\frac{1}{2}h'H_0h+\infty 1\left(h\notin T_C\left(\beta_0\right)\right)\right\}=\mathcal{J}$$

Now we show $\hat{\beta}_n^*\xrightarrow{p}\beta_0$. Since $\alpha_n\to 0$ implies $\alpha_n\sqrt{n}\bar{H}_n\left(\hat{l}_n^*\left(\bar{\beta}_n\right)-\hat{l}_n\left(\bar{\beta}_n\right)\right)=o_p^*(1)$,

$$\begin{aligned}\hat{\beta}_n^*-\beta_0&=\underset{u\in(C-\beta_0)}{\arg\min}\left\{\frac{1}{2}\left\|u+\beta_0-\bar{\beta}_n+\alpha_n\sqrt{n}\bar{H}_n^{-1}\left(\hat{l}_n^*\left(\bar{\beta}_n\right)-\hat{l}_n\left(\bar{\beta}_n\right)\right)\right\|_{\bar{H}_n}^2\right\}\\&=\underset{u\in(C-\beta_0)}{\arg\min}\left\{\frac{1}{2}u'H_0u+u'H_0\left(\beta_0-\bar{\beta}_n\right)+\frac{1}{2}\left\|\beta_0-\bar{\beta}_n\right\|_{H_0}^2\right\}+o_p(1)\\&=\bar{\beta}_n-\beta_0+o_p(1)=o_p(1)\end{aligned}$$

where the second line follows from the convexity in h of the proximal bootstrap objective function and compactness of $C-\beta_0$.

Next, to derive the asymptotic distribution, since $\sqrt{n}\alpha_n\to\infty$ and $\sqrt{n}\left(\hat{\beta}_n-\beta_0\right)=O_P(1)$, $\frac{\hat{\beta}_n^*-\hat{\beta}_n}{\alpha_n}=\frac{\hat{\beta}_n^*-\beta_0}{\alpha_n}+o_P^*(1)$, where

$$\begin{aligned}\frac{\hat{\beta}_n^*-\beta_0}{\alpha_n}&=\underset{h\in\mathbb{R}^d}{\arg\min}\left\{\infty 1\left(h\notin\frac{C-\beta_0}{\alpha_n}\right)+\alpha_n\sqrt{n}\left(\hat{l}_n^*\left(\bar{\beta}_n\right)-\hat{l}_n\left(\bar{\beta}_n\right)\right)'\left(\beta_0-\bar{\beta}_n+\alpha_nh\right)+\frac{1}{2}\left\|\beta_0-\bar{\beta}_n+\alpha_nh\right\|_{\bar{H}_n}^2\right\}\\&=\underset{h\in\mathbb{R}^d}{\arg\min}\left\{\infty 1\left(h\notin\frac{C-\beta_0}{\alpha_n}\right)+\sqrt{n}\left(\hat{l}_n^*\left(\bar{\beta}_n\right)-\hat{l}_n\left(\bar{\beta}_n\right)\right)'\left(\frac{\beta_0-\bar{\beta}_n}{\alpha_n}+h\right)+\frac{1}{2}\left\|\frac{\beta_0-\bar{\beta}_n}{\alpha_n}+h\right\|_{\bar{H}_n}^2\right\}\\&=\underset{h\in\mathbb{R}^d}{\arg\min}\left\{\infty 1\left(h\notin\frac{C-\beta_0}{\alpha_n}\right)+h'\sqrt{n}\left(\hat{l}_n^*\left(\bar{\beta}_n\right)-\hat{l}_n\left(\bar{\beta}_n\right)\right)+\frac{1}{2}h'\bar{H}_nh+o_p^*(1)\right\}\end{aligned}$$

Assumption 4 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. Therefore,

$$h'\sqrt{n}\left(\hat{l}_n^*\left(\bar{\beta}_n\right)-\hat{l}_n\left(\bar{\beta}_n\right)\right)+\frac{1}{2}h'\bar{H}_nh\underset{\mathbb{W}}{\overset{\mathbb{P}}{\rightsquigarrow}}h'W_0+\frac{1}{2}h'H_0h$$

A bootstrap in probability version of Theorem 4 of Knight (1999) can then be stated to show that

$$\underbrace{h'\sqrt{n}\left(\hat{l}_n^*\left(\bar{\beta}_n\right)-\hat{l}_n\left(\bar{\beta}_n\right)\right)+\frac{1}{2}h'\bar{H}_nh+\infty 1\left(h\notin\frac{C-\beta_0}{\alpha_n}\right)}_{\hat{\mathbb{G}}_n^*(h)}\overset{p}{\underset{e-d}{\rightarrow}}\underbrace{h'W_0+\frac{1}{2}h'H_0h+\infty 1\left(h\notin T_C\left(\beta_0\right)\right)}_{\mathbb{G}_0(h)}$$

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

A modification of Theorem 1 of Knight (1999) to epi-convergence of conditional laws suggests that

$$\frac{\hat{\beta}_n^*-\beta_0}{\alpha_n}=\arg\min_{h\in\mathbb{R}^d}\hat{\mathbb{G}}_n^*\left(h\right)+o_P^*(1)\underset{\mathbb{W}}{\overset{\mathbb{P}}{\rightsquigarrow}}\arg\min_{h\in\mathbb{R}^d}\mathbb{G}_0\left(h\right)=\mathcal{J}$$

∎

5.1.2 Proof of Theorem 2

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 estimated constraints. We need to constrain $\hat{\beta}_n$ to lie in C and replace his population objective $F(\cdot)$ with the population Lagrangian $\mathcal{L}\left(\beta_0,\lambda_0\right)\equiv Q\left(\beta_0\right)+\sum_{j\in\mathcal{E}\cup\mathcal{I}}\lambda_{0j}f_{0j}\left(\beta_0\right)$. The first order KKT condition $\nabla\mathcal{L}\left(\beta_0,\lambda_0\right)\equiv l\left(\beta_0\right)+\sum_{j\in\mathcal{E}\cup\mathcal{I}}\lambda_{0j}F_{0j}=0$ and positive-definiteness of $\nabla^2\mathcal{L}\left(\beta_0,\lambda_0\right)\equiv H_0+\sum_{j\in\mathcal{E}\cup\mathcal{I}}\lambda_{0j}G_{0j}$ imply the local quadratic expansion $\mathcal{L}\left(\beta,\lambda_0\right)=\mathcal{L}\left(\beta_0,\lambda_0\right)+\frac{1}{2}\|\beta-\beta_0\|^2_{\nabla^2\mathcal{L}(\beta_0,\lambda_0)}+o\left(\|\beta-\beta_0\|^2\right)$ for β in a small neighborhood of β_0. This expansion in combination with the local quadratic approximation of the Lagrangain in Assumption 6 will imply Pollard (1984)'s equation (6), where $F_n(\cdot)$ is replaced by $\hat{\mathcal{L}}_n\left(\cdot\right)$ and the empirical process $E_n\Delta$ is replaced by $\sqrt{n}\left(\hat{l}_n\left(\beta_0\right)-l\left(\beta_0\right)\right)+\sum_{j\in\mathcal{E}\cup\mathcal{I}}\lambda_{0j}\sqrt{n}\left(F_{nj}\left(\beta_0\right)-F_{0j}\right)$, which is still $O_p(1)$ by the assumptions of our Theorem. Note that it is not necessary for β_0 to be in the interior of C_0 to show $\sqrt{n}$-consistency; it would be necessary if we were to show asymptotic normality.

Recall $\hat{\mathcal{L}}_n(\beta) = \hat{Q}_n(\beta) + \sum_{j\in\mathcal{E}\cup\mathcal{I}} \lambda_{nj} f_{nj}(\beta)$ is the sample Lagrangian evaluated at the optimal Lagrange multipliers λ_{nj} for $\hat{\beta}_n$. It is well known that $\hat{\beta}_n = \underset{\beta\in C}{\arg\min}\hat{Q}_n(\beta)$ can be equivalently expressed as $\hat{\beta}_n = \underset{\beta\in C}{\arg\min}\hat{\mathcal{L}}_n(\beta)$ when the first order KKT conditions are satisfied. Shapiro (1990) shows that it is important to use this Lagrangian formulation when deriving the asymptotic distribution of $\hat{\beta}_n$ because it captures the sampling variation in the objective as well as the estimated constraints.

Additionally, LICQ implies that the linearization of the constraint set is sufficient to capture the geometry of the constraints near β_0 (Nocedal and Wright (2006) chapter 12). We can then use this linearized constraint set to derive the asymptotic distribution of $\sqrt{n}\left(\hat{\beta}_n - \beta_0\right)$. Denote the feasible direction set by

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

For some mean value $\tilde{\beta}$ such that $F_{nj}\left(\tilde{\beta}\right) \xrightarrow{p} F_{0j}$, denote the linearized feasible direction set by

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

Minimizing the Lagrangian over $\mathcal{F}_n$ is equivalent to minimizing the Lagrangian over Σ_n:

$$\begin{aligned}
\sqrt{n}\left(\hat{\beta}_n - \beta_0\right) &= \underset{h\in\mathcal{F}_n}{\arg\min}\left\{n\hat{\mathcal{L}}_n\left(\beta_0 + \frac{h}{\sqrt{n}}\right) - n\hat{\mathcal{L}}_n(\beta_0)\right\} \\
&= \underset{h\in\Sigma_n}{\arg\min}\left\{n\hat{\mathcal{L}}_n\left(\beta_0 + \frac{h}{\sqrt{n}}\right) - n\hat{\mathcal{L}}_n(\beta_0)\right\} \\
&= \underset{h\in\Sigma_n}{\arg\min}\left\{n\hat{Q}_n\left(\beta_0 + \frac{h}{\sqrt{n}}\right) - n\hat{Q}_n(\beta_0) + \sum_{j\in\mathcal{E}\cup\mathcal{I}} \lambda_{nj} n\left(f_{nj}\left(\beta_0 + \frac{h}{\sqrt{n}}\right) - f_{nj}(\beta_0)\right)\right\} \\
&\rightsquigarrow \underset{h\in\Sigma}{\arg\min}\left\{h'W_0 + \frac{1}{2}h'H_0 h + \sum_{j\in\mathcal{E}\cup\mathcal{I}_+^*(\lambda_0)} \lambda_{0j}\left(h'V_{0j} + \frac{1}{2}h'G_{0j}h\right)\right\} = \mathcal{J}
\end{aligned}$$

where the last line follows from the following arguments. First note that Assumption 6 implies that for any $\delta_n \to 0$,

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

$$= o_P(1)$$

Therefore, uniformly in h,

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

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 $j \in \mathcal{I}\backslash\mathcal{I}_+^*\left(\lambda_0\right)$. 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,

$$\begin{aligned} &h'\sqrt{n}\left(\hat{l}_n\left(\beta_0\right) - l\left(\beta_0\right)\right) + \frac{1}{2}h'H_0h + \sum_{j\in\mathcal{E}\cup\mathcal{I}} \lambda_{0j}\left(\sqrt{n}\left(F_{nj}\left(\beta_0\right) - F_{0j}\right)' h + \frac{1}{2}h'G_{0j}h\right) + o_P(1) \\ &\rightsquigarrow h'W_0 + \frac{1}{2}h'H_0h + \sum_{j\in\mathcal{E}\cup\mathcal{I}} \lambda_{0j}\left(h'V_{0j} + \frac{1}{2}h'G_{0j}h\right) \\ &= h'W_0 + \frac{1}{2}h'H_0h + \sum_{j\in\mathcal{E}\cup\mathcal{I}_+^*\left(\lambda_0\right)} \lambda_{0j}\left(h'V_{0j} + \frac{1}{2}h'G_{0j}h\right) \end{aligned}$$

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

Now consider the constraints. Since $\sqrt{n} f_{nj}\left(\beta_0\right) + F_{nj}\left(\tilde{\beta}\right)' h \xrightarrow{p} -\infty$ for $j \in \mathcal{I}\backslash\mathcal{I}^*$, the nonactive inequality constraints do not affect the asymptotic distribution. Since $\sqrt{n} f_{nj}\left(\beta_0\right) \rightsquigarrow U_{0j}$, jointly, for all $j \in \mathcal{E} \cup \mathcal{I}^*$, $F_{nj}\left(\tilde{\beta}\right) = F_0 + o_P(1)$, and finite dimensional convergence in distribution implies epi-convergence in distribution for convex functions,

$$\infty 1\left(h \notin \Sigma_n\right) \to_{e-d} \infty 1\left(h \notin \left\{h : U_{0j} + F_{0j}'h = 0 \text{ for } j \in \mathcal{E}, U_{0j} + F_{0j}'h \leqslant 0 \text{ for } j \in \mathcal{I}^*\right\}\right)$$

Because we have assumed LICQ at β_0, Theorem 2.1 of Shapiro (1988) implies that minimizing over $\left\{h : U_{0j} + F_{0j}'h = 0 \text{ for } j \in \mathcal{E}, U_{0j} + F_{0j}'h \leqslant 0 \text{ for } j \in \mathcal{I}^*\right\}$ will produce the same set of solutions as minimizing over $\Sigma \equiv \left\{h : U_{0j} + F_{0j}'h = 0 \text{ for } j \in \mathcal{E} \cup \mathcal{I}_+^*(\lambda_0), U_{0j} + F_{0j}'h \leqslant 0 \text{ for } j \in \mathcal{I}_0^*(\lambda_0)\right\}$.

Condition (iii) is a second order sufficient condition and guarantees that the argmin in $\mathcal{J}$ is unique. Then by the argmin continuous mapping theorem (Theorem 1 of Knight (1999)), $\arg\min_h \hat{\mathbb{G}}_n(h) \rightarrow_{e-d} \arg\min_h \mathbb{G}_0(h)$, where

$$
\begin{aligned}
\hat{\mathbb{G}}_n(h) &= n\hat{Q}_n\left(\beta_0 + \frac{h}{\sqrt{n}}\right) - n\hat{Q}_n(\beta_0) + \sum_{j\in\mathcal{E}\cup\mathcal{I}} \lambda_{nj} n \left(f_{nj}\left(\beta_0 + \frac{h}{\sqrt{n}}\right) - f_{nj}(\beta_0)\right) + \infty 1(h \notin \Sigma_n) \\
\mathbb{G}_0(h) &= h'W_0 + \frac{1}{2}h'H_0 h + \sum_{j\in\mathcal{E}\cup\mathcal{I}_+^*(\lambda_0)} \lambda_{0j}\left(h'V_{0j} + \frac{1}{2}h'G_{0j}h\right) + \infty 1(h \notin \Sigma)
\end{aligned}
$$

Now we show consistency of the proximal bootstrap $\hat{\beta}_n^* \xrightarrow{p} \beta_0$. $\alpha_n \to 0$ implies $\alpha_n\sqrt{n}\bar{H}_n\left(\hat{l}_n^*(\bar{\beta}_n) - \hat{l}_n(\bar{\beta}_n)\right) = o_p^*(1)$ and $\alpha_n\sqrt{n}\left(\bar{F}_{nj}^* - \bar{F}_{nj}\right) = o_p^*(1)$ for all $j \in \mathcal{E}\cup\mathcal{I}$. Using convexity of the proximal bootstrap objective function and compactness of $C^* - \beta_0$,

$$
\begin{aligned}
\hat{\beta}_n^* - \beta_0 = &\underset{u\in(C^*-\beta_0)}{\arg\min}\left\{\frac{1}{2}\left\|u + \beta_0 - \bar{\beta}_n + \alpha_n\sqrt{n}\bar{H}_n^{-1}\left(\hat{l}_n^*(\bar{\beta}_n) - \hat{l}_n(\bar{\beta}_n)\right)\right\|_{\bar{H}_n}^2\right. \\
&\left. + \sum_{j\in\mathcal{E}\cup\mathcal{I}} \bar{\lambda}_{nj}\left(\alpha_n\sqrt{n}\left(\bar{F}_{nj}^* - \bar{F}_{nj}\right)'\left(u + \beta_0 - \bar{\beta}_n\right) + \frac{1}{2}\left\|u + \beta_0 - \bar{\beta}_n\right\|_{\bar{G}_{nj}}^2\right)\right\} \\
= &\underset{u\in(C^*-\beta_0)}{\arg\min}\left\{\frac{1}{2}u'\left(H_0 + \sum_{j\in\mathcal{E}\cup\mathcal{I}}\lambda_{0j}G_{0j}\right)u + u'\left(H_0 + \sum_{j\in\mathcal{E}\cup\mathcal{I}}\lambda_{0j}G_{0j}\right)\left(\beta_0 - \bar{\beta}_n\right)\right. \\
&\left. + \frac{1}{2}\left\|\beta_0 - \bar{\beta}_n\right\|_{H_0}^2 + \frac{1}{2}\sum_{j\in\mathcal{E}\cup\mathcal{I}}\lambda_{0j}\left\|\beta_0 - \bar{\beta}_n\right\|_{G_{0j}}^2\right\} + o_p(1) \\
= &\bar{\beta}_n - \beta_0 + o_p(1) = o_p(1)
\end{aligned}
$$

Next we derive the asymptotic distribution of the proximal bootstrap. Note that since C^* is already a linearized constraint set, the linearized feasible direction set is simply

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

Note that $\frac{f_{nj}(\bar{\beta}_n)}{\alpha_n} \xrightarrow{p} -\infty$ for $j \in \mathcal{I} \backslash \mathcal{I}^*$ while $\frac{f_{nj}(\bar{\beta}_n)}{\alpha_n} = \frac{\sqrt{n}(f_{nj}(\bar{\beta}_n) - f_{0j}(\beta_0))}{\sqrt{n}\alpha_n} = \frac{\sqrt{n}(f_{nj}(\bar{\beta}_n) - f_{nj}(\beta_0))}{\sqrt{n}\alpha_n} + \frac{\sqrt{n}(f_{nj}(\beta_0) - f_{0j}(\beta_0))}{\sqrt{n}\alpha_n} = o_P(1)$ for $j \in \mathcal{E} \cup \mathcal{I}^*$. Additionally, $\bar{F}_{nj}' \left(\frac{\beta_0 - \bar{\beta}_n}{\alpha_n}\right) = o_P(1)$, $\bar{F}_n = F_0 + o_P(1)$, $\sqrt{n}\left(f_{nj}^*(\beta_0) - f_{nj}(\beta_0)\right) \underset{\mathbb{W}}{\overset{\mathbb{P}}{\rightsquigarrow}} U_{0j}$, jointly, for all $j \in \mathcal{E} \cup \mathcal{I}^*$, and $\sqrt{n}\left(f_{nj}^*(\bar{\beta}_n) - f_{nj}(\bar{\beta}_n)\right) \underset{\mathbb{W}}{\overset{\mathbb{P}}{\rightsquigarrow}} U_{0j}$, jointly, for all $j \in \mathcal{E} \cup \mathcal{I}^*$ because $\sup_{\|\beta - \beta_0\| \leqslant o(1)} \sqrt{n}\left(f_n^*(\beta) - f_n(\beta) - f_n^*(\beta_0) + f_n(\beta_0)\right) = o_P^*(1)$. Therefore,

$$\infty 1\left(h \notin \Sigma_n^*\right) \underset{e-d}{\xrightarrow{p}} \infty 1\left(h \notin \left\{h : U_{0j} + F_{0j}' h = 0 \text{ for } j \in \mathcal{E}, U_{0j} + F_{0j}' h \leqslant 0 \text{ for } j \in \mathcal{I}^*\right\}\right)$$

Next, we can center and scale the bootstrap estimator to get

$$\begin{aligned}
\frac{\hat{\beta}_n^* - \beta_0}{\alpha_n} &= \underset{h \in \Sigma_n^*}{\arg\min} \Bigg\{ \alpha_n \sqrt{n}\left(\hat{l}_n^*(\bar{\beta}_n) - \hat{l}_n(\bar{\beta}_n)\right)'\left(\beta_0 - \bar{\beta}_n + \alpha_n h\right) + \frac{1}{2}\left\|\beta_0 - \bar{\beta}_n + \alpha_n h\right\|_{\bar{H}_n}^2 \\
&\qquad + \sum_{j \in \mathcal{E} \cup \mathcal{I}} \bar{\lambda}_{nj}\left(\alpha_n \sqrt{n}\left(\bar{F}_{nj}^* - \bar{F}_{nj}\right)'\left(\beta_0 - \bar{\beta}_n + \alpha_n h\right) + \frac{1}{2}\left\|\beta_0 - \bar{\beta}_n + \alpha_n h\right\|_{\bar{G}_{nj}}^2\right)\Bigg\} \\
&= \underset{h \in \Sigma_n^*}{\arg\min} \Bigg\{ \sqrt{n}\left(\hat{l}_n^*(\bar{\beta}_n) - \hat{l}_n(\bar{\beta}_n)\right)'\left(\frac{\beta_0 - \bar{\beta}_n}{\alpha_n} + h\right) + \frac{1}{2}\left\|\frac{\beta_0 - \bar{\beta}_n}{\alpha_n} + h\right\|_{\bar{H}_n}^2 \\
&\qquad + \sum_{j \in \mathcal{E} \cup \mathcal{I}} \bar{\lambda}_{nj}\left(\sqrt{n}\left(\bar{F}_{nj}^* - \bar{F}_{nj}\right)'\left(\frac{\beta_0 - \bar{\beta}_n}{\alpha_n} + h\right) + \frac{1}{2}\left\|\frac{\beta_0 - \bar{\beta}_n}{\alpha_n} + h\right\|_{\bar{G}_{nj}}^2\right)\Bigg\} \\
&= \underset{h \in \Sigma_n^*}{\arg\min} \Bigg\{ h' \sqrt{n}\left(\hat{l}_n^*(\bar{\beta}_n) - \hat{l}_n(\bar{\beta}_n)\right) + \frac{1}{2} h' \bar{H}_n h \\
&\qquad + \sum_{j \in \mathcal{E} \cup \mathcal{I}} \bar{\lambda}_{nj}\left(h' \sqrt{n}\left(\bar{F}_{nj}^* - \bar{F}_{nj}\right) + \frac{1}{2} h' \bar{G}_{nj} h\right) + o_P^*(1)\Bigg\} \\
&\underset{\mathbb{W}}{\overset{\mathbb{P}}{\rightsquigarrow}} \underset{h \in \Sigma}{\arg\min} \left\{ h' W_0 + \frac{1}{2} h' H_0 h + \sum_{j \in \mathcal{E} \cup \mathcal{I}_+^*(\lambda_0)} \lambda_{0j}\left(h' V_{0j} + \frac{1}{2} h' G_{0j} h\right)\right\} = \mathcal{J}
\end{aligned}$$

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

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

Next, note that Assumption 4, $\max_{j\in\mathcal{E}\cup\mathcal{I}}\left|\bar{\lambda}_{nj}-\lambda_{0j}\right| \xrightarrow{p} 0$, and $\sup_{\|\beta-\beta_0\|\leqslant o(1)} \sqrt{n}\left(F_n^*(\beta)-F_n(\beta)-F_n^*(\beta_0)+F_n(\beta_0)\right)=o_P^*(1)$ imply $\sqrt{n}\left(\hat{l}_n^*\left(\bar{\beta}_n\right)-\hat{l}_n\left(\bar{\beta}_n\right)\right)+\sum_{j\in\mathcal{E}\cup\mathcal{I}}\bar{\lambda}_{nj}\sqrt{n}\left(\bar{F}_{nj}^*-\bar{F}_{nj}\right)\underset{\mathbb{W}}{\overset{\mathbb{P}}{\rightsquigarrow}}W_0+\sum_{j\in\mathcal{E}\cup\mathcal{I}}\lambda_{0j}V_{0j}$ 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 $\sqrt{n}\left(\hat{l}_n^*\left(\beta_0\right)-\hat{l}_n\left(\beta_0\right)\right)+\sum_{j\in\mathcal{E}\cup\mathcal{I}}\lambda_{0j}\sqrt{n}\left(F_{nj}^*\left(\beta_0\right)-F_{nj}\left(\beta_0\right)\right)\underset{\mathbb{W}}{\overset{\mathbb{P}}{\rightsquigarrow}}W_0+\sum_{j\in\mathcal{E}\cup\mathcal{I}}\lambda_{0j}V_{0j}$. 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,

$$
\begin{aligned}
&h'\sqrt{n}\left(\hat{l}_n^*\left(\bar{\beta}_n\right)-\hat{l}_n\left(\bar{\beta}_n\right)\right)+\frac{1}{2}h'H_0h+\sum_{j\in\mathcal{E}\cup\mathcal{I}}\bar{\lambda}_{nj}\left(h'\sqrt{n}\left(\bar{F}_{nj}^*-\bar{F}_{nj}\right)+\frac{1}{2}h'G_{0j}h\right)\\
&\underset{\mathbb{W}}{\overset{\mathbb{P}}{\rightsquigarrow}} h'W_0+\frac{1}{2}h'H_0h+\sum_{j\in\mathcal{E}\cup\mathcal{I}}\lambda_{0j}\left(h'V_{0j}+\frac{1}{2}h'G_{0j}h\right)\\
&=h'W_0+\frac{1}{2}h'H_0h+\sum_{j\in\mathcal{E}\cup\mathcal{I}_+^*(\lambda_0)}\lambda_{0j}\left(h'V_{0j}+\frac{1}{2}h'G_{0j}h\right)
\end{aligned}
$$

as a process indexed by h in the space of bounded functions on compact sets $\ell^\infty(K)$ for any compact $K\subset\mathbb{R}^d$. Finally, note that $\hat{\beta}_n^*$ is unique because $\bar{H}_n+\sum_{j\in\mathcal{E}\cup\mathcal{I}}\bar{\lambda}_{nj}\bar{G}_{nj}$ is symmetric and positive definite. 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

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

$$\mathbb{G}_0(h) = h'W_0 + \frac{1}{2}h'H_0h + \sum_{j\in\mathcal{E}\cup\mathcal{I}_+^*(\lambda_0)} \lambda_{0j}\left(h'V_{0j} + \frac{1}{2}h'G_{0j}h\right) + \infty 1(h \notin \Sigma)$$

■

5.1.3 Proof of Theorem 3

We will first show $\Omega^* \subseteq \Omega$ for all drifting sequences and then apply stochastic dominance arguments to show uniform coverage of the confidence set. For active inequality and equality constraints $j \in \mathcal{E} \cup \mathcal{I}^*$ where $f_{0j}(\beta_0) = 0$, $\sqrt{n}f_{nj}(\beta_0) - F_{nj}\left(\tilde{\beta}\right)' h \rightsquigarrow U_{0j} - F_{0j}'h$. For $\sqrt{n}$-drifting inequality constraints $j \in \mathcal{I}_{1/2}^{\#}$ where $f_{0j}(\beta_0) = c/\sqrt{n}$ for $c < 0$, $\sqrt{n}f_{nj}(\beta_0) - F_{nj}\left(\tilde{\beta}\right)' h = \sqrt{n}\left(f_{nj}(\beta_0) - f_{0j}(\beta_0)\right) - F_{nj}\left(\tilde{\beta}\right)' h + c \rightsquigarrow U_{0j} - F_{0j}'h + c$. For slower than $\sqrt{n}$ drifting constraints $j \in \mathcal{I}_{<1/2}^{\#}$ where $f_{0j}(\beta_0) = c/n^{\rho}$ for $c < 0$ and $\rho < 1/2$, $\sqrt{n}f_{nj}(\beta_0) - F_{nj}\left(\tilde{\beta}\right)' h = \sqrt{n}\left(f_{nj}(\beta_0) - f_{0j}(\beta_0)\right) - F_{nj}\left(\tilde{\beta}\right)' h + cn^{1/2-\rho} \xrightarrow{p} -\infty$. For faster than $\sqrt{n}$ drifting constraints $j \in \mathcal{I}_{>1/2}^{\#}$ where $f_{0j}(\beta_0) = c/n^{\rho}$ for $c < 0$ and $\rho > 1/2$, $\sqrt{n}f_{nj}(\beta_0) - F_{nj}\left(\tilde{\beta}\right)' h = \sqrt{n}\left(f_{nj}(\beta_0) - f_{0j}(\beta_0)\right) - F_{nj}\left(\tilde{\beta}\right)' h + cn^{1/2-\rho} \rightsquigarrow U_{0j} - F_{0j}'h$. For the nonactive and nondrifting inequality constraints $j \in \mathcal{I}\backslash\left(\mathcal{I}^* \cup \mathcal{I}_{1/2}^{\#} \cup \mathcal{I}_{<1/2}^{\#} \cup \mathcal{I}_{>1/2}^{\#}\right)$, $\sqrt{n}f_{nj}(\beta_0) - F_{nj}\left(\tilde{\beta}\right)' h \xrightarrow{p} -\infty$. Therefore,

$$\Omega = \left\{h : U_{0j} - F_{0j}'h = 0 \text{ for } j \in \mathcal{E}, U_{0j} - F_{0j}'h \leqslant 0 \text{ for } j \in \mathcal{I}^* \cup \mathcal{I}_{>1/2}^{\#}, U_{0j} - F_{0j}'h \leqslant -c \text{ for } j \in \mathcal{I}_{1/2}^{\#}\right\}$$

Recall that since $\sqrt{n}\left(f_{nj}^*(\beta_0) - f_{nj}(\beta_0)\right) \underset{\mathbb{W}}{\overset{\mathbb{P}}{\rightsquigarrow}} U_{0j}$ and $\sup_{\|\beta-\beta_0\|\leqslant o(1)} \sqrt{n}\left(f_n^*(\beta) - f_n(\beta) - f_n^*(\beta_0) + f_n(\beta_0)\right) = o_p^*(1)$, $\sqrt{n}\left(f_{nj}^*\left(\bar{\beta}_n\right) - f_{nj}\left(\bar{\beta}_n\right)\right) \underset{\mathbb{W}}{\overset{\mathbb{P}}{\rightsquigarrow}} U_{0j}$ for all $j \in \mathcal{E} \cup \mathcal{I}$. For active inequality and equality constraints $j \in \mathcal{E} \cup \mathcal{I}^*$ where $f_{0j}(\beta_0) = 0$, $\frac{f_{nj}\left(\bar{\beta}_n\right)}{\alpha_n} = \frac{\sqrt{n}\left(f_{nj}\left(\bar{\beta}_n\right) - f_{0j}(\beta_0)\right)}{\sqrt{n}\alpha_n} = o_p(1)$. For $\sqrt{n}$-drifting inequality constraints $j \in \mathcal{I}_{1/2}^{\#}$ where $f_{0j}(\beta_0) = c/\sqrt{n}$ for $c < 0$, $\frac{f_{nj}\left(\bar{\beta}_n\right)}{\alpha_n} = \frac{\sqrt{n}\left(f_{nj}\left(\bar{\beta}_n\right) - f_{0j}(\beta_0)\right)}{\sqrt{n}\alpha_n} + \frac{c}{\sqrt{n}\alpha_n} = \frac{\sqrt{n}\left(f_{nj}\left(\bar{\beta}_n\right) - f_{0j}(\beta_0)\right)}{\sqrt{n}\alpha_n} + o_p(1) = o_p(1)$. For faster than $\sqrt{n}$ drifting constraints $j \in \mathcal{I}_{>1/2}^{\#}$ where $f_{0j}(\beta_0) = c/n^{\rho}$ for $c < 0$ and $\rho > 1/2$, $\frac{f_{nj}\left(\bar{\beta}_n\right)}{\alpha_n} = \frac{\sqrt{n}\left(f_{nj}\left(\bar{\beta}_n\right) - f_{0j}(\beta_0)\right)}{\sqrt{n}\alpha_n} + \frac{c}{n^{\rho}\alpha_n} = \frac{\sqrt{n}\left(f_{nj}\left(\bar{\beta}_n\right) - f_{0j}(\beta_0)\right)}{\sqrt{n}\alpha_n} + o_p(1) = o_p(1)$. For slower than $\sqrt{n}$ drifting constraints where $f_{0j}(\beta_0) = c/n^{\rho}$ for $c < 0$ and $\rho < 1/2$, $\frac{f_{nj}\left(\bar{\beta}_n\right)}{\alpha_n} = \frac{\sqrt{n}\left(f_{nj}\left(\bar{\beta}_n\right) - f_{0j}(\beta_0)\right)}{\sqrt{n}\alpha_n} + \frac{c}{n^{\rho}\alpha_n} \xrightarrow{p} \begin{cases} 0 & \text{if } \alpha_n n^{\rho} \to \infty \\ c/k & \text{if } \alpha_n n^{\rho} \to k \\ -\infty & \text{if } \alpha_n n^{\rho} \to 0 \end{cases}$. We will label these constraints as $\mathcal{I}_{<1/2,\infty}^{\#}$, $\mathcal{I}_{<1/2,k}^{\#}$, and $\mathcal{I}_{<1/2,0}^{\#}$, respectively to reflect the limit of $\alpha_n n^{\rho}$. For the nonactive and nondrifting inequality constraints $j \in \mathcal{I} \backslash \left(\mathcal{I}^* \cup \mathcal{I}^{\#}_{1/2} \cup \mathcal{I}^{\#}_{>1/2} \cup \mathcal{I}^{\#}_{<1/2,\infty} \cup \mathcal{I}^{\#}_{<1/2,k} \cup \mathcal{I}^{\#}_{<1/2,0}\right)$, $\frac{f_{nj}(\bar{\beta}_n)}{\alpha_n} \xrightarrow{p} -\infty$. Therefore,

$$\Omega^* = \Big\{h : U_{0j} - F'_{0j}h = 0 \text{ for } j \in \mathcal{E}, U_{0j} - F'_{0j}h \leqslant 0 \text{ for } j \in \mathcal{I}^* \cup \mathcal{I}^{\#}_{1/2} \cup \mathcal{I}^{\#}_{>1/2} \cup \mathcal{I}^{\#}_{<1/2,\infty},$$
$$U_{0j} - F'_{0j}h \leqslant -c/k \text{ for } j \in \mathcal{I}^{\#}_{<1/2,k}\Big\}$$

Assumption 6 and the continuous mapping theorem imply that

$$\begin{aligned}
& n\left(\hat{\mathcal{L}}_n(\beta_0) - \hat{\mathcal{L}}_n\left(\hat{\beta}_n\right)\right) \\
& = \sqrt{n}\left(\beta_0 - \hat{\beta}_n\right)' \sqrt{n}\left(\hat{l}_n(\beta_0) - l(\beta_0)\right) + \frac{1}{2}\left\|\sqrt{n}\left(\beta_0 - \hat{\beta}_n\right)\right\|^2_{H_0} \\
& + \sum_{j\in\mathcal{E}\cup\mathcal{I}} \lambda_{0j}\left(\sqrt{n}\left(F_{nj}(\beta_0) - F_{0j}\right)' \sqrt{n}\left(\beta_0 - \hat{\beta}_n\right) + \frac{1}{2}\left\|\sqrt{n}\left(\beta_0 - \hat{\beta}_n\right)\right\|^2_{G_{0j}}\right) + o_P(1) \\
& \rightsquigarrow q(\mathcal{J})
\end{aligned}$$

where the $o_P(1)$ term is uniform in P, $q(h) \equiv -h'W_0 + \frac{1}{2}h'H_0h + \sum_{j\in\mathcal{E}\cup\mathcal{I}^*_+(\lambda_0)} \lambda_{0j}\left(-h'V_{0j} + \frac{1}{2}h'G_{0j}h\right)$, and $\mathcal{J} = \underset{h\in\Omega}{\arg\min}\, q(h)$. Note that $q(\mathcal{J}) = \underset{h\in\Omega}{\min}\, q(h)$. Similarly, for $\mathcal{J}^* = \underset{h\in\Omega^*}{\arg\min}\, q(h)$,

$$\begin{aligned}
& \frac{\hat{A}^*_n\left(\hat{\beta}_n\right) - \hat{A}^*_n\left(\hat{\beta}^*_n\right)}{\alpha_n^2} \\
& = \sqrt{n}\left(\hat{l}^*_n(\bar{\beta}_n) - \hat{l}_n(\bar{\beta}_n)\right)' \left(\frac{\beta_0 - \hat{\beta}^*_n}{\alpha_n}\right) + \frac{1}{2}\left\|\frac{\beta_0 - \hat{\beta}^*_n}{\alpha_n}\right\|^2_{\bar{H}_n} \\
& + \sum_{j\in\mathcal{E}\cup\mathcal{I}} \bar{\lambda}_{nj}\left(\sqrt{n}\left(\bar{F}^*_{nj} - \bar{F}_{nj}\right)' \left(\frac{\beta_0 - \hat{\beta}^*_n}{\alpha_n}\right) + \frac{1}{2}\left\|\frac{\beta_0 - \hat{\beta}^*_n}{\alpha_n}\right\|^2_{\bar{G}_{nj}}\right) + o_P(1) \\
& \underset{\mathbb{W}}{\overset{\mathbb{P}}{\rightsquigarrow}} q(\mathcal{J}^*)
\end{aligned}$$

where the $o_P(1)$ term is uniform in P. Since $\Omega^* \subseteq \Omega$ and $q(h)$ is a strictly convex function of h, $\underset{h\in\Omega}{\min}\, q(h) \leqslant \underset{h\in\Omega^*}{\min}\, q(h)$ uniformly over P. This implies that the asymptotic distribution of $\frac{\hat{A}^*_n(\hat{\beta}_n) - \hat{A}^*_n(\hat{\beta}^*_n)}{\alpha_n^2}$ uniformly first order stochastically dominates the asymptotic distribution of $n\left(\hat{\mathcal{L}}_n(\beta_0) - \hat{\mathcal{L}}_n\left(\hat{\beta}_n\right)\right)$. Under the assumptions of this theorem, for all $\epsilon_n > 0$ and n large enough, there exists $\delta_n > 0$ such that $\underset{P\in\mathcal{P}}{\sup} P\left(\underset{x\in\mathbb{R}}{\sup}\left\{J^*_{\alpha_n}(x, P) - J_n(x, P)\right\} > \epsilon_n\right) \leqslant \delta_n$. Let $\hat{c}^*_{1-\alpha}$ be the

$1-\alpha$ quantile of $\frac{\hat{A}_n^*(\hat{\beta}_n)-\hat{A}_n^*(\hat{\beta}_n^*)}{\alpha_n^2}$ and let $\hat{c}_{1-\alpha}$ be the $1-\alpha$ quantile of $n\left(\hat{\mathcal{L}}_n\left(\beta_0\right)-\hat{\mathcal{L}}_n\left(\hat{\beta}_n\right)\right)$. Take $\{\epsilon_n\}_{n=1}^{\infty}$ and $\{\delta_n\}_{n=1}^{\infty}$ to be positive sequences such that $\epsilon_n \to 0$ and $\delta_n \to 0$. Then,

$$
\begin{aligned}
&\liminf_{n\to\infty} \inf_{P\in\mathcal{P}} P\left(n\left(\hat{\mathcal{L}}_n\left(\beta_0\right)-\hat{\mathcal{L}}_n\left(\hat{\beta}_n\right)\right) \leqslant \hat{c}_{1-\alpha}^*\right)\\
&\geqslant \liminf_{n\to\infty} \inf_{P\in\mathcal{P}} P\left(n\left(\hat{\mathcal{L}}_n\left(\beta_0\right)-\hat{\mathcal{L}}_n\left(\hat{\beta}_n\right)\right) \leqslant \hat{c}_{1-\alpha}^* \cap \sup_{x\in\mathbb{R}}\left\{J_{\alpha_n}^*\left(x,P\right)-J_n\left(x,P\right)\right\} \leqslant \epsilon_n\right)\\
&\geqslant \liminf_{n\to\infty} \inf_{P\in\mathcal{P}} P\left(n\left(\hat{\mathcal{L}}_n\left(\beta_0\right)-\hat{\mathcal{L}}_n\left(\hat{\beta}_n\right)\right) \leqslant \hat{c}_{1-\alpha-\epsilon_n} \cap \sup_{x\in\mathbb{R}}\left\{J_{\alpha_n}^*\left(x,P\right)-J_n\left(x,P\right)\right\} \leqslant \epsilon_n\right)\\
&\geqslant \liminf_{n\to\infty} \inf_{P\in\mathcal{P}} P\left(n\left(\hat{\mathcal{L}}_n\left(\beta_0\right)-\hat{\mathcal{L}}_n\left(\hat{\beta}_n\right)\right) \leqslant \hat{c}_{1-\alpha-\epsilon_n}\right) - \limsup_{n\to\infty} \sup_{P\in\mathcal{P}} P\left(\sup_{x\in\mathbb{R}}\left\{J_{\alpha_n}^*\left(x,P\right)-J_n\left(x,P\right)\right\} > \epsilon_n\right)\\
&\geqslant 1-\alpha
\end{aligned}
$$

■

5.2 Additional Results

5.2.1 Equality Constrained Quadratic Program

Lemma 5.1. *Suppose $H_0 \in \mathbb{R}^d \times \mathbb{R}^d$ is nonsingular, $R \in \mathbb{R}^d \times \mathbb{R}^m$ has rank m, and $\Delta_n = O_P(1)$. Then*

$$
\begin{aligned}
h^+ = &\ \underset{R'h=\delta}{\arg\min}\ h'\Delta_n + \frac{1}{2}h'H_0h\\
&= -H_0^{-1}\left(I - R\left(R'H_0^{-1}R\right)^{-1}R'H_0^{-1}\right)\Delta_n + H_0^{-1}R\left(R'H_0^{-1}R\right)^{-1}\delta
\end{aligned}
$$

Proof: The Lagrangian and KKT conditions are

$$
\begin{aligned}
&\mathcal{L} = h'\Delta_n + \frac{1}{2}h'H_0h + \lambda \circ \left(R'h-\delta\right)\\
&\Delta_n + H_0h + R\lambda = 0\\
&R'h - \delta = 0
\end{aligned}
$$

The first KKT condition says $h^+ = -H_0^{-1}\left(\Delta_n + R\lambda\right)$. Substituting into the second KKT condition,

$$
-R'H_0^{-1}\left(\Delta_n + R\lambda\right) = \delta \implies \lambda = -\left(R'H_0^{-1}R\right)^{-1}\left(\delta + R'H_0^{-1}\Delta_n\right)
$$

Therefore,

$$
\begin{aligned}
h^{+} &= -H_0^{-1}\Delta_n + H_0^{-1}R\left(R'H_0^{-1}R\right)^{-1}\left(\delta + R'H_0^{-1}\Delta_n\right) \\
&= -H_0^{-1}\left(I - R\left(R'H_0^{-1}R\right)^{-1}R'H_0^{-1}\right)\Delta_n + H_0^{-1}R\left(R'H_0^{-1}R\right)^{-1}\delta
\end{aligned}
$$

5.2.2 Inequality Constrained Quadratic Program

Lemma 5.2. *Suppose $H_0 \in \mathbb{R}^d \times \mathbb{R}^d$ is nonsingular, $R_\Lambda \in \mathbb{R}^d \times \mathbb{R}^{m_\Lambda}$ has rank m_Λ, and $\Delta_n = O_P(1)$, where R_Λ denotes the submatrix of $R \in \mathbb{R}^d \times \mathbb{R}^m$ corresponding to the active constraints. Then*

$$
\begin{aligned}
h^{+} = &\underset{R'h \leqslant \delta}{\arg\min}\; h'\Delta_n + \frac{1}{2}h'H_0h \\
&= \max\left(-H_0^{-1}\left(I - R_\Lambda\left(R_\Lambda'H_0^{-1}R_\Lambda\right)^{-1}R_\Lambda'H_0^{-1}\right)\Delta_n + H_0^{-1}R_\Lambda\left(R_\Lambda'H_0^{-1}R_\Lambda\right)^{-1}\delta_\Lambda, -H_0^{-1}\Delta_n\right)
\end{aligned}
$$

where δ_Λ denotes the subvector of δ corresponding to the active constraints.

Proof: The Lagrangian and KKT Conditions are

$$
\begin{aligned}
&\mathcal{L} = h'\Delta_n + \frac{1}{2}h'H_0h + \sum_{i=1}^{m}\mu_i\left(R_i'h - \delta_i\right) \\
&\Delta_n + H_0h + R\mu = 0 \\
&\mu_i \geqslant 0, \mu_i\left(R_i'h - \delta_i\right) = 0 \forall i = 1...m
\end{aligned}
$$

The first KKT condition says $h^{+} = -H_0^{-1}\left(\Delta_n + R\mu\right)$. The second says that if $\mu_i > 0$, then $R_i'h^{+} - \delta_i = 0$; such an inequality constraint is called strongly active (binding). It can also be the case that $\mu_i = 0$ and $R_i'h^{+} - \delta_i = 0$, in which case the inequality constraint is called weakly active. The assumption that R_Λ has rank m_Λ implies linear independence constraint qualification is satisfied, which means the set of Lagrange multipliers that satisfy the KKT conditions is a singleton (Wachsmuth (2013)). Let the Lagrange multipliers corresponding to active constraints be denoted μ_Λ. The Lagrange multipliers corresponding to nonactive constraints are zero. Therefore $R\mu = R_\Lambda\mu_\Lambda$. Stacking the equations $R_i'h^{+} - \delta_i = 0$ for the active constraints, and accounting for the possibility that $\mu_i = 0$ for the weakly active constraints (since strict complementarity may not hold),

$$R'_\Lambda h^+ - \delta_\Lambda = -R'_\Lambda H_0^{-1}\left(\Delta_n + R_\Lambda \mu_\Lambda\right) - \delta_\Lambda = 0 \implies \mu_\Lambda = \max\left(-\left(R'_\Lambda H_0^{-1} R_\Lambda\right)^{-1}\left(R'_\Lambda H_0^{-1}\Delta_n + \delta_\Lambda\right), 0\right)$$

Therefore,

$$\begin{aligned}
h^+ &= -H_0^{-1}\left(\Delta_n + R_\Lambda \mu_\Lambda\right)\\
&= \max\left(-H_0^{-1}\Delta_n + H_0^{-1}R_\Lambda\left(R'_\Lambda H_0^{-1}R_\Lambda\right)^{-1}\left(R'_\Lambda H_0^{-1}\Delta_n + \delta_\Lambda\right), -H_0^{-1}\Delta_n\right)\\
&= \max\left(-H_0^{-1}\left(I - R_\Lambda\left(R'_\Lambda H_0^{-1}R_\Lambda\right)^{-1}R'_\Lambda H_0^{-1}\right)\Delta_n + H_0^{-1}R_\Lambda\left(R'_\Lambda H_0^{-1}R_\Lambda\right)^{-1}\delta_\Lambda, -H_0^{-1}\Delta_n\right)
\end{aligned}$$

5.2.3 Consistency of Proximal Bootstrap with infinite number of non-drifting constraints when $l(\beta_0) = 0$ (Remark 6)

The limiting distribution of $\sqrt{n}\left(\hat{\beta}_n - \beta_0\right)$ can be difficult to characterize due to the presence of an infinite number of constraints in the limit as $n \to \infty$. To avoid explicitly characterizing the limiting distribution, we will work with the following finite constraint set Σ:

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

Here, $\mathcal{I}_n^* \equiv \{j \in \mathcal{I}_n : f_{0j}(\beta_0) = 0\}$, and Σ_n and Σ_n^* are the same as in the proof of Theorem 2 except allowing for $\mathcal{E}_n$ and $\mathcal{I}_n$ to depend on n. To demonstrate consistency of the proximal bootstrap, we will show that both $\infty 1(h \notin \Sigma_n)$ and $\infty 1(h \notin \Sigma_n^*)$ have the same limit (in the sense of epi-convergence in distribution) without explicitly characterizing this limit. Because $\infty 1(h \notin \Sigma_n)$ and $\infty 1(h \notin \Sigma_n^*)$ are convex functions, to show epi-convergence in distribution, it suffices to show finite dimensional convergence. In particular, we will show that $\infty 1(h \notin \Sigma_n) - \infty 1(h \notin \Sigma)$ and $\infty 1(h \notin \Sigma_n^*) - \infty 1(h \notin \Sigma)$ both converge in finite dimension to zero. To do so, we will assume $\sup_{t \in \mathbb{R}}\left|P\left(\max_{j \in \mathcal{E}_n \cup \mathcal{I}_n^*}\sqrt{n} f_{nj}(\beta_0) \leqslant t\right) - P\left(\max_{j \in \mathcal{E}_n \cup \mathcal{I}_n^*} U_{0j} \leqslant t\right)\right| \to 0$, and

$\sup_{t \in \mathbb{R}}\left|P\left(\max_{j \in \mathcal{E}_n \cup \mathcal{I}_n^*}\sqrt{n}\left(f_{nj}^*\left(\bar{\beta}_n\right) - f_{nj}\left(\bar{\beta}_n\right)\right) \leqslant t \middle| \mathcal{X}_n\right) - P\left(\max_{j \in \mathcal{E}_n \cup \mathcal{I}_n^*} U_{0j} \leqslant t\right)\right| \xrightarrow{p} 0$. These assumptions can be derived using the results in Chernozhukov et al. (2013) and Chernozhukov et al. (2019) for Gaussian approximation of maxima of sums for high dimensional random vectors. We will also need to assume $\max_{j\in\mathcal{E}_n\cup\mathcal{I}_n^*}\left|F_{nj}\left(\beta_0\right)-F_{0j}\right|=o_P(1)$ and $\max_{j\in\mathcal{E}_n\cup\mathcal{I}_n^*}\left|\bar{F}_{nj}-F_{0j}\right|=o_P(1)$.

We now show finite dimensional convergence of $\infty 1\left(h\notin\Sigma_n\right)$ to $\infty 1\left(h\notin\Sigma\right)$. For any $h_1,...,h_k$ where k is fixed,

$$
\begin{aligned}
&P\left(h_1\in\Sigma_n,...,h_k\in\Sigma_n\right)\\
&=P\left(\bigcap_{i=1}^{k}\left\{\sqrt{n}f_{nj}\left(\beta_0\right)+F_{nj}\left(\beta_0\right)'h_i=0\text{ for }j\in\mathcal{E}_n,\sqrt{n}f_{nj}\left(\beta_0\right)+F_{nj}\left(\beta_0\right)'h_i\leqslant 0\text{ for }j\in\mathcal{I}_n\right\}\right)\\
&=P\left(\left\{\max_{j\in\mathcal{E}_n\cup\mathcal{I}_n}\left(\sqrt{n}f_{nj}\left(\beta_0\right)+\max_{1\leqslant i\leqslant k}F_{nj}\left(\beta_0\right)'h_i\right)\leqslant 0\right\}\bigcap\left\{\max_{j\in\mathcal{E}_n}\left(-\sqrt{n}f_{nj}\left(\beta_0\right)-\max_{1\leqslant i\leqslant k}F_{nj}\left(\beta_0\right)'h_i\right)\leqslant 0\right\}\right)\\
&P\left(h_1\in\Sigma,...,h_k\in\Sigma\right)\\
&=P\left(\bigcap_{i=1}^{k}\left\{U_{0j}+F_{0j}'h_i=0\text{ for }j\in\mathcal{E}_n,U_{0j}+F_{0j}'h_i\leqslant 0\text{ for }j\in\mathcal{I}_n^*\right\}\right)\\
&=P\left(\left\{\max_{j\in\mathcal{E}_n\cup\mathcal{I}_n^*}\left(U_{0j}+\max_{1\leqslant i\leqslant k}F_{0j}'h_i\right)\leqslant 0\right\}\bigcap\left\{\max_{j\in\mathcal{E}_n}\left(-U_{0j}-\max_{1\leqslant i\leqslant k}F_{0j}'h_i\right)\leqslant 0\right\}\right)\\
&P\left(h_1\in\Sigma_n,...,h_k\in\Sigma_n\right)-P\left(h_1\in\Sigma,...,h_k\in\Sigma\right)\\
&=P\left(\left\{\max_{j\in\mathcal{E}_n\cup\mathcal{I}_n^*}\left(\sqrt{n}f_{nj}\left(\beta_0\right)+\max_{1\leqslant i\leqslant k}F_{nj}\left(\beta_0\right)'h_i\right)\leqslant 0\right\}\bigcap\left\{\max_{j\in\mathcal{E}_n}\left(-\sqrt{n}f_{nj}\left(\beta_0\right)-\max_{1\leqslant i\leqslant k}F_{nj}\left(\beta_0\right)'h_i\right)\leqslant 0\right\}\right)\\
&-P\left(\left\{\max_{j\in\mathcal{E}_n\cup\mathcal{I}_n^*}\left(U_{0j}+\max_{1\leqslant i\leqslant k}F_{0j}'h_i\right)\leqslant 0\right\}\bigcap\left\{\max_{j\in\mathcal{E}_n}\left(-U_{0j}-\max_{1\leqslant i\leqslant k}F_{0j}'h_i\right)\leqslant 0\right\}\right)+o(1)\\
&=P\left(\left\{\max_{j\in\mathcal{E}_n\cup\mathcal{I}_n^*}\left(\sqrt{n}f_{nj}\left(\beta_0\right)+\max_{1\leqslant i\leqslant k}F_{0j}'h_i\right)\leqslant 0\right\}\bigcap\left\{\max_{j\in\mathcal{E}_n}\left(-\sqrt{n}f_{nj}\left(\beta_0\right)-\max_{1\leqslant i\leqslant k}F_{0j}'h_i\right)\leqslant 0\right\}\right)\\
&-P\left(\left\{\max_{j\in\mathcal{E}_n\cup\mathcal{I}_n^*}\left(U_{0j}+\max_{1\leqslant i\leqslant k}F_{0j}'h_i\right)\leqslant 0\right\}\bigcap\left\{\max_{j\in\mathcal{E}_n}\left(-U_{0j}-\max_{1\leqslant i\leqslant k}F_{0j}'h_i\right)\leqslant 0\right\}\right)+o(1)\\
&=o(1)
\end{aligned}
$$

where we have used $\sqrt{n}f_{nj}\left(\beta_0\right)+\max_{1\leqslant i\leqslant k}F_{nj}\left(\beta_0\right)'h_i\xrightarrow{p}-\infty$ for $j\in\mathcal{I}_n\backslash\mathcal{I}_n^*$, $\max_{j\in\mathcal{E}_n\cup\mathcal{I}_n^*}\left|F_{nj}\left(\beta_0\right)-F_{0j}\right|=o_P(1)$, and $\sup_{t\in\mathbb{R}}\left|P\left(\max_{j\in\mathcal{E}_n\cup\mathcal{I}_n^*}\sqrt{n}f_{nj}\left(\beta_0\right)\leqslant t\right)-P\left(\max_{j\in\mathcal{E}_n\cup\mathcal{I}_n^*}U_{0j}\leqslant t\right)\right|\to 0$. The rest of the arguments are the same as in Theorem 2. It follows that for $c_{1-\alpha}$ the $1-\alpha$ quantile of $\mathcal{J}=\arg\min_{h\in\Sigma}\left\{h'W_0+\frac{1}{2}h'H_0h\right\}$, $P\left(\sqrt{n}\left(\hat{\beta}_n-\beta_0\right)>c_{1-\alpha}\right)\to\alpha$.

Similarly, to show finite dimensional convergence in probability of $\infty 1\left(h\notin\Sigma_n^*\right)$ to $\infty 1\left(h\notin\Sigma\right)$, for any $h_1,...,h_k$ where k is fixed,

$$P\left(h_1\in\Sigma_n^*,...,h_k\in\Sigma_n^*\middle|\,\mathcal{X}_n\right)$$

$$
\begin{aligned}
&= P\left(\bigcap_{i=1}^{k}\left\{\frac{f_{nj}\left(\bar{\beta}_n\right)}{\alpha_n} + \bar{F}'_{nj}h_i + \sqrt{n}\left(f^*_{nj}\left(\bar{\beta}_n\right) - f_{nj}\left(\bar{\beta}_n\right)\right) + \bar{F}'_{nj}\left(\frac{\beta_0 - \bar{\beta}_n}{\alpha_n}\right) = 0 \text{ for } j \in \mathcal{E}_n,\right.\right.\\
&\qquad \left.\left.\frac{f_{nj}\left(\bar{\beta}_n\right)}{\alpha_n} + \bar{F}'_{nj}h_i + \sqrt{n}\left(f^*_{nj}\left(\bar{\beta}_n\right) - f_{nj}\left(\bar{\beta}_n\right)\right) + \bar{F}'_{nj}\left(\frac{\beta_0 - \bar{\beta}_n}{\alpha_n}\right) \leqslant 0 \text{ for } j \in \mathcal{I}_n\right\}\middle|\, \mathcal{X}_n\right)\\
&= P\left(\left\{\max_{j\in\mathcal{E}_n\cup\mathcal{I}_n}\left(\sqrt{n}\left(f^*_{nj}\left(\bar{\beta}_n\right) - f_{nj}\left(\bar{\beta}_n\right)\right) + \max_{1\leqslant i\leqslant k}\bar{F}'_{nj}h_i + \frac{f_{nj}\left(\bar{\beta}_n\right)}{\alpha_n} + \bar{F}'_{nj}\left(\frac{\beta_0 - \bar{\beta}_n}{\alpha_n}\right)\right) \leqslant 0\right\}\right.\\
&\qquad \left.\bigcap\left\{\max_{j\in\mathcal{E}_n}\left(-\sqrt{n}\left(f^*_{nj}\left(\bar{\beta}_n\right) - f_{nj}\left(\bar{\beta}_n\right)\right) - \max_{1\leqslant i\leqslant k}\bar{F}'_{nj}h_i - \frac{f_{nj}\left(\bar{\beta}_n\right)}{\alpha_n} - \bar{F}'_{nj}\left(\frac{\beta_0 - \bar{\beta}_n}{\alpha_n}\right)\right) \leqslant 0\right\}\middle|\, \mathcal{X}_n\right)
\end{aligned}
$$

$$
\begin{aligned}
&P\left(h_1 \in \Sigma^*_n, ..., h_k \in \Sigma^*_n \middle| \mathcal{X}_n\right) - P\left(h_1 \in \Sigma, ..., h_k \in \Sigma\right)\\
&= P\left(\left\{\max_{j\in\mathcal{E}_n\cup\mathcal{I}^*_n}\left(\sqrt{n}\left(f^*_{nj}\left(\bar{\beta}_n\right) - f_{nj}\left(\bar{\beta}_n\right)\right) + \max_{1\leqslant i\leqslant k}F'_{0j}h_i\right) \leqslant 0\right\}\right.\\
&\qquad \left.\bigcap\left\{\max_{j\in\mathcal{E}_n}\left(-\sqrt{n}\left(f^*_{nj}\left(\bar{\beta}_n\right) - f_{nj}\left(\bar{\beta}_n\right)\right) - \max_{1\leqslant i\leqslant k}F'_{0j}h_i\right) \leqslant 0\right\}\middle|\, \mathcal{X}_n\right)\\
&- P\left(\left\{\max_{j\in\mathcal{E}_n\cup\mathcal{I}^*_n}\left(U_{0j} + \max_{1\leqslant i\leqslant k}F'_{0j}h_i\right) \leqslant 0\right\}\bigcap\left\{\max_{j\in\mathcal{E}_n}\left(-U_{0j} - \max_{1\leqslant i\leqslant k}F'_{0j}h_i\right) \leqslant 0\right\}\right) + o_P(1)\\
&= o_P(1)
\end{aligned}
$$

where we have used $\bar{F}'_{nj}\left(\frac{\beta_0 - \bar{\beta}_n}{\alpha_n}\right) = o_P(1)$, $\frac{f_{nj}\left(\bar{\beta}_n\right)}{\alpha_n} + \max_{1\leqslant i\leqslant k} F_{nj}\left(\beta_0\right)' h_i \xrightarrow{p} -\infty$ for $j \in \mathcal{I}_n \backslash \mathcal{I}^*_n$, $\frac{f_{nj}\left(\bar{\beta}_n\right)}{\alpha_n} = o_P(1)$ for all $j \in \mathcal{E}_n \cup \mathcal{I}^*_n$, $\sup_{t\in\mathbb{R}}\left|P\left(\max_{j\in\mathcal{E}_n\cup\mathcal{I}^*_n}\sqrt{n}\left(f^*_{nj}\left(\bar{\beta}_n\right) - f_{nj}\left(\bar{\beta}_n\right)\right) \leqslant t\middle|\, \mathcal{X}_n\right) - P\left(\max_{j\in\mathcal{E}_n\cup\mathcal{I}^*_n} U_{0j} \leqslant t\right)\right| \xrightarrow{p} 0$, and $\max_{j\in\mathcal{E}_n\cup\mathcal{I}^*_n}\left|\bar{F}_{nj} - F_{0j}\right| = o_P(1)$. The rest of the arguments are the same as in Theorem 2. It follows that for $c_{1-\alpha}$ the $1-\alpha$ quantile of $\mathcal{J} = \arg\min_{h\in\Sigma}\left\{h'W_0 + \frac{1}{2}h'H_0h\right\}$, $P\left(\alpha_n^{-1}\left(\hat{\beta}^*_n - \hat{\beta}_n\right) > c_{1-\alpha}\middle|\, \mathcal{X}_n\right) \xrightarrow{p} \alpha$. Since $P\left(\sqrt{n}\left(\hat{\beta}_n - \beta_0\right) > c_{1-\alpha}\right) \to \alpha$, it follows that $P\left(\sqrt{n}\left(\hat{\beta}_n - \beta_0\right) > c^B_{1-\alpha}\right) \to \alpha$, where $c^B_{1-\alpha}$ is the $1-\alpha$ empirical quantile of $\alpha_n^{-1}\left(\hat{\beta}^*_n - \hat{\beta}_n\right)$.

We can also show that the uniformity arguments in Theorem 3 extend to the case of an infinite number of constraints when $l\left(\beta_0\right) = 0$. Define

$$
\begin{aligned}
\Omega &= \left\{h : U_{0j} - F'_{0j}h = 0 \text{ for } j \in \mathcal{E}_n, U_{0j} - F'_{0j}h \leqslant 0 \text{ for } j \in \mathcal{I}^*_n \cup \mathcal{I}^{\#}_{n,>1/2}, U_{0j} - F'_{0j}h \leqslant -c \text{ for } j \in \mathcal{I}^{\#}_{n,1/2}\right\}\\
\Omega^* &= \left\{h : U_{0j} - F'_{0j}h = 0 \text{ for } j \in \mathcal{E}_n, U_{0j} - F'_{0j}h \leqslant 0 \text{ for } j \in \mathcal{I}^*_n \cup \mathcal{I}^{\#}_{n,1/2} \cup \mathcal{I}^{\#}_{n,>1/2} \cup \mathcal{I}^{\#}_{n,<1/2,\infty},\right.\\
&\qquad\qquad \left.U_{0j} - F'_{0j}h \leqslant -c/k \text{ for } j \in \mathcal{I}^{\#}_{n,<1/2,k}\right\}
\end{aligned}
$$

We can show that for any $h_1, ..., h_k$ where k is fixed, $P\left(h_1 \in \Sigma_n, ..., h_k \in \Sigma_n\right) - P\left(h_1 \in \Omega, ..., h_k \in \Omega\right) = o_P(1)$ and $P\left(h_1 \in \Sigma^*_n, ..., h_k \in \Sigma^*_n \middle| \mathcal{X}_n\right) - P\left(h_1 \in \Omega^*, ..., h_k \in \Omega^*\right) = o_P(1)$. Let $\hat{c}^*_{1-\alpha}$ be the $1-\alpha$

quantile of $\frac{\hat{A}_n^*(\hat{\beta}_n)-\hat{A}_n^*(\hat{\beta}_n^*)}{\alpha_n^2}$, and let $c_{1-\alpha}^*$ be the $1-\alpha$ quantile of $\min_{h\in\Omega^*} q(h)$, where $\hat{A}_n^*(\beta) = \alpha_n\sqrt{n}\left(\hat{l}_n^*(\bar{\beta}_n) - \hat{l}_n(\bar{\beta}_n)\right)'(\beta-\bar{\beta}_n) + \frac{1}{2}\left\|\beta-\bar{\beta}_n\right\|_{\bar{H}_n}^2$ and $q(h) = \left\{-h'W_0 + \frac{1}{2}h'H_0h\right\}$. Let $J_{\alpha_n}^*(\cdot,P)$ denote the conditional CDF of $\frac{\hat{A}_n^*(\hat{\beta}_n)-\hat{A}_n^*(\hat{\beta}_n^*)}{\alpha_n^2}$ under P, and assume for all $\epsilon > 0$, $\limsup_{n\to\infty}\sup_{P\in\mathcal{P}} P\left(\sup_{x\in\mathbb{R}}\left|J_{\alpha_n}^*(x,P) - J^*(x,P)\right| > \epsilon\right) = 0$, where the limiting distributions $\{J^*(\cdot,P) : P\in\mathcal{P}\}$ are equicontinuous at their $1-\alpha$ quantiles. For positive sequences $\{\epsilon_n\}_{n=1}^\infty$ and $\{\delta_n\}_{n=1}^\infty$ such that $\epsilon_n \to 0$ and $\delta_n \to 0$,

$$
\begin{aligned}
&\liminf_{n\to\infty}\inf_{P\in\mathcal{P}} P\left(n\left(\hat{Q}_n(\beta_0) - \hat{Q}_n\left(\hat{\beta}_n\right)\right) \leqslant \hat{c}_{1-\alpha}^*\right)\\
&\geqslant \liminf_{n\to\infty}\inf_{P\in\mathcal{P}} P\left(\min_{h\in\Omega} q(h) \leqslant \hat{c}_{1-\alpha}^* \cap \sup_{x\in\mathbb{R}}\left|J_{\alpha_n}^*(x,P) - J^*(x,P)\right| \leqslant \epsilon_n/2\right)\\
&\geqslant \liminf_{n\to\infty}\inf_{P\in\mathcal{P}} P\left(\min_{h\in\Omega} q(h) \leqslant c_{1-\alpha-\epsilon_n}^* \cap \sup_{x\in\mathbb{R}}\left|J_{\alpha_n}^*(x,P) - J^*(x,P)\right| \leqslant \epsilon_n/2\right)\\
&\geqslant \liminf_{n\to\infty}\inf_{P\in\mathcal{P}} P\left(\min_{h\in\Omega^*} q(h) \leqslant c_{1-\alpha-\epsilon_n}^* \cap \sup_{x\in\mathbb{R}}\left|J_{\alpha_n}^*(x,P) - J^*(x,P)\right| \leqslant \epsilon_n/2\right)\\
&\geqslant \liminf_{n\to\infty}\inf_{P\in\mathcal{P}} P\left(\min_{h\in\Omega^*} q(h) \leqslant c_{1-\alpha-\epsilon_n}^*\right) - \limsup_{n\to\infty}\sup_{P\in\mathcal{P}} P\left(\sup_{x\in\mathbb{R}}\left|J_{\alpha_n}^*(x,P) - J^*(x,P)\right| > \epsilon_n/2\right)\\
&\geqslant 1-\alpha-\epsilon_n-\delta_n \geqslant 1-\alpha
\end{aligned}
$$

References

Amemiya, T. (1985): *Advanced Econometrics*, Harvard University Press. 9, 15

Andrews, D. W. (1999): "Estimation when a parameter is on a boundary," *Econometrica*, 67, 1341–1383. 2, 5, 9

——— (2000): "Inconsistency of the bootstrap when a parameter is on the boundary of the parameter space," *Econometrica*, 68, 399–405. 2, 5, 17

——— (2002): "Generalized method of moments estimation when a parameter is on a boundary," *Journal of Business & Economic Statistics*, 20, 530–544. 2, 9

Beck, A. (2017): *First-order methods in optimization*, vol. 25, SIAM. 5

Byrd, R. H., J. Nocedal, and F. Oztoprak (2016): "An inexact successive quadratic approximation method for L-1 regularized optimization," *Mathematical Programming*, 157, 375–396. 5

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. 19

Chen, X., T. M. Christensen, and E. Tamer (2018): "Monte Carlo confidence sets for identified sets," *Econometrica*, 86, 1965–2018. 3, 9, 19

Chernozhukov, V., D. Chetverikov, K. Kato, and Y. Koike (2019): "Improved Central Limit Theorem and bootstrap approximations in high dimensions," *arXiv preprint arXiv:1912.10529*. 45

Chernozhukov, V., D. Chetverikov, K. Kato, et al. (2013): "Gaussian approximations and multiplier bootstrap for maxima of sums of high-dimensional random vectors," *The Annals of Statistics*, 41, 2786–2819. 45

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. 26

Fang, Z. and J. Seo (2021): "A projection framework for testing shape restrictions that form convex cones," *Econometrica*, 89, 2439–2458. 3

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

Gafarov, B. (2016): "Inference on scalar parameters in set–identified affine models job market paper," Tech. rep., Mimeo: UC Davis. 3

Gallant, A. R., H. Hong, M. P. Leung, and J. Li (2022): "Constrained estimation using penalization and MCMC," *Journal of Econometrics*, 228, 85–106. 8

Geyer, C. J. (1994): "On the asymptotics of constrained M-estimation," *The Annals of Statistics*, 1993–2010. 2, 5, 6, 9, 32, 33, 34

Ghanbari, H. and K. Scheinberg (2016): "Proximal quasi-Newton methods for convex optimization," Tech. rep., Technical Report. 5

Hong, H. and J. Li (2020): "The numerical bootstrap," *The Annals of Statistics*, 48, 397–412. 2, 40

Horowitz, J. L. and S. Lee (2019): "Non-asymptotic inference in a class of optimization problems," . 3

Hsieh, Y.-W., X. Shi, and M. Shum (2022): "Inference on estimators defined by mathematical programming," *Journal of Econometrics*, 226, 248–268. 3

Kaido, H. (2016): "A dual approach to inference for partially identified econometric models," *Journal of Econometrics*, 192, 269–290. 3

Kaido, H., F. Molinari, and J. Stoye (2019): "Confidence intervals for projections of partially identified parameters," *Econometrica*, 87, 1397–1432. 3

——— (2021): "Constraint qualifications in partial identification," *Econometric Theory*, 1–24. 3, 10

Kaido, H. and A. Santos (2014): "Asymptotically efficient estimation of models defined by convex moment inequalities," *Econometrica*, 82, 387–413. 3

Knight, K. (1999): "Epi-convergence in distribution and stochastic equi-semicontinuity," *Unpublished manuscript*, 37. 33, 35, 38

——— (2001): "Limiting distributions of linear programming estimators," *Extremes*, 4, 87–103. 3

——— (2006): "Asymptotic theory for M-estimators of boundaries," in *The Art of Semiparametrics*, Springer, 1–21. 3

——— (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. 3

Kosorok, M. R. (2007): *Introduction to empirical processes and semiparametric inference*, Springer. 4

Lee, J. D., Y. Sun, and M. Saunders (2012): "Proximal Newton-type methods for convex optimization," in *Advances in Neural Information Processing Systems*, 827–835. 5

Lee, J. D., Y. Sun, and M. A. Saunders (2014): "Proximal Newton-type methods for minimizing composite functions," *SIAM Journal on Optimization*, 24, 1420–1443. 5

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. 3

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. 9, 15

Nocedal, J. and S. Wright (2006): *Numerical optimization*, Springer Science & Business Media. 10, 36

Parikh, N., S. Boyd, et al. (2014): "Proximal algorithms," *Foundations and trends® in Optimization*, 1, 127–239. 5

Pollard, D. (1984): "Convergence of stochastic processes," . 35

——— (1985): "New ways to prove central limit theorems," *Econometric Theory*, 1, 295–313. 7

——— (1991): "Asymptotics for least absolute deviation regression estimators," *Econometric Theory*, 7, 186–199. 37

Rockafellar, R. T., R. J.-B. Wets, and M. Wets (1998): *Variational analysis*, vol. 317, Springer. 9

Rodomanov, A. and D. Kropotov (2016): "A superlinearly-convergent proximal Newton-type method for the optimization of finite sums," in *International Conference on Machine Learning*, 2597–2605. 5

Rust, J. (1987): "Optimal replacement of GMC bus engines: An empirical model of Harold Zurcher," *Econometrica: Journal of the Econometric Society*, 999–1033. 1, 4, 12, 28, 32

Shapiro, A. (1988): "Sensitivity analysis of nonlinear programs and differentiability properties of metric projections," *SIAM Journal on Control and Optimization*, 26, 628–645. 3, 37

——— (1989): "Asymptotic properties of statistical estimators in stochastic programming," *The Annals of Statistics*, 841–858. 3

——— (1990): "On differential stability in stochastic programming," *Mathematical Programming*, 47, 107–116. 3, 36

——— (1991): "Asymptotic analysis of stochastic programs," *Annals of Operations Research*, 30, 169–186. 3

——— (1993): "Asymptotic behavior of optimal solutions in stochastic programming," *Mathematics of Operations Research*, 18, 829–845. 3

——— (2000): "Statistical inference of stochastic optimization problems," in *Probabilistic Constrained Optimization*, Springer, 282–307. 3

Su, C.-L. and K. L. Judd (2012): "Constrained optimization approaches to estimation of structural models," *Econometrica*, 80, 2213–2230. 1, 4, 28, 29, 30

Tran-Dinh, Q., A. Kyrillidis, and V. Cevher (2015): "Composite Self-Concordant Minimization," *Journal of Machine Learning Research*, 16, 371–416. 5

van der Vaart, A. and J. A. Wellner (1996): *Weak Convergence and Empirical Processes*, Springer. 32

Wachsmuth, G. (2013): "On LICQ and the uniqueness of Lagrange multipliers," *Operations Research Letters*, 41, 78–80. 10, 44