

The Proximal Bootstrap for Constrained Estimators *

Jessie Li

March 14, 2022

We demonstrate how to use the proximal bootstrap to consistently estimate the limiting distribution of \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 fixed 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 estimated or fixed

*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, and UC Davis Econometrics workshop (in particular Bulat Gafarov and Takuya Ura) for helpful comments and suggestions.

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 consistency of 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 fixed 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 fixed 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 $\alpha_n \rightarrow 0$ and $\sqrt{n}\alpha_n \rightarrow \infty$. 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 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 \(2016\)](#), [Gafarov \(2016\)](#), [Hsieh et al. \(2022\)](#), [Kaido et al. \(2019\)](#), [Kaido et al. \(2021\)](#), and [Horowitz and Lee \(2019\)](#). While several of these papers are concerned with conducting inference on the optimal value of the constrained optimization problem, we are instead interested in conducting inference on the optimal solution. Perhaps the closest paper to ours is [Hsieh et al. \(2022\)](#) who also consider inference for the optimal solution, but they focus on linear programming (LP) and convex quadratic programming (QP) problems with linear constraints. In contrast to [Hsieh et al. \(2022\)](#), we allow for nonconvex and nonlinear objective and constraint functions, but we do not 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.

Section 2 briefly reviews the concept of proximal mappings from the optimization literature. Section 3 contains all the main theoretical results for finite-dimensional constrained estimators with either fixed constraints (subsection 3.2) or estimated constraints (subsection 3.3). Section 4 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 5 concludes. The appendix contains proofs of the theorems.

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

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

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

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

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)).

3 Proximal Bootstrap

3.1 Notation

Consider a random sample X_1, X_2, \dots, 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 $\sup_{f \in BL_1} |E^* f(Z_n) - E f(Z)| \rightarrow 0$ 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): $Z_n \overset{\mathbb{P}}{\underset{\mathbb{W}}{\rightsquigarrow}} Z$ in the metric space (\mathbb{D}, d) if and only if $\sup_{f \in BL_1} |E_{\mathbb{W}} f(Z_n) - E f(Z)| \xrightarrow{P} 0$ 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 $\lim_{m \rightarrow \infty} \limsup_{n \rightarrow \infty} P(P_n(M_n^* > m) > \epsilon) \rightarrow 0$, $\forall \epsilon > 0$.

3.2 Constrained Estimators with Fixed 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 consider constrained estimators $\hat{\beta}_n = \arg \min_{\beta \in C} \hat{Q}_n(\beta)$, where $\hat{Q}_n(\beta)$ is possibly non-smooth, nonconvex function that converges uniformly to a function $Q(\beta)$ that is twice differentiable at $\beta_0 = \arg \min_{\beta \in C} Q(\beta)$, and C is a closed constraint set that is Chernoff Regular at β_0 .

We will show that the proximal bootstrap is consistent both when β_0 lies in the interior and on the boundary of C . For some $\alpha_n \rightarrow 0$ and $\alpha_n \sqrt{n} \rightarrow \infty$, define the proximal bootstrap estimator as

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

Here, $\bar{\beta}_n$ is an initial \sqrt{n} -consistent estimator of β_0 . For example, we can use $\bar{\beta}_n = \hat{\beta}_n$. In the case where $\hat{Q}_n(\beta)$ is differentiable, $\hat{l}_n(\beta)$ can simply be the Jacobian of $\hat{Q}_n(\beta)$. More generally, to handle non-differentiable $\hat{Q}_n(\beta)$, $\hat{l}_n(\beta)$ is a subgradient of $\hat{Q}_n(\beta)$. $\hat{l}_n^*(\beta)$ is the bootstrap analog of $\hat{l}_n(\beta)$. \bar{H}_n is a consistent, symmetric, positive definite estimate of the population Hessian $H_0 = \frac{\partial^2 Q(\beta_0)}{\partial \beta \partial \beta'}$.

It is well known (see e.g. [Geyer \(1994\)](#)) that if $Q(\beta)$ achieves its minimum over C at some point β_0 where it has a local quadratic approximation $Q(\beta) = Q(\beta_0) + \frac{1}{2}(\beta - \beta_0)' H_0 (\beta - \beta_0) + o(\|\beta - \beta_0\|^2)$, then $\sqrt{n}(\hat{\beta}_n - \beta_0) \rightsquigarrow \mathcal{J} = \arg \min_{h \in T_C(\beta_0)} \{h' W_0 + \frac{1}{2} h' H_0 h\}$, where $T_C(\beta_0) \equiv \limsup_{\tau \downarrow 0} \frac{C - \beta_0}{\tau}$ is the tangent cone of C at β_0 . For closed sets C that are Chernoff Regular at β_0 , the limit exists and $T_C(\beta_0) = \lim_{\tau \downarrow 0} \frac{C - \beta_0}{\tau}$. In the next theorem, we show that $\frac{\hat{\beta}_n^* - \hat{\beta}_n}{\alpha_n} \xrightarrow[\mathbb{W}]{\mathbb{P}} \mathcal{J}$. 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 . It can be found in Corollary 3.2.3 of [van der Vaart and Wellner \(1996\)](#).

Assumption 1. (i) $\hat{\beta}_n = \arg \min_{\beta \in C} \hat{Q}_n(\beta)$ is uniformly tight and unique. ¹

(ii) $\beta_0 = \arg \min_{\beta \in C} Q(\beta)$ is unique, where $Q(\beta)$ is a lower semicontinuous function that is twice differentiable at β_0 and $\sup_{\beta \in K} |\hat{Q}_n(\beta) - Q(\beta)| = o_P(1)$ for every compact subset K of C .

¹For every $\epsilon > 0$, there exists a compact $K \subset C$ with $P(\hat{\beta}_n \in K) \geq 1 - \epsilon$ for every n .

The next assumption states that $\hat{Q}_n(\beta)$ admits a uniform local quadratic approximation around \sqrt{n} neighborhoods of β_0 . It is needed to derive the asymptotic distribution of $\sqrt{n}(\hat{\beta}_n - \beta_0)$.

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

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

Note that assumption 2 effectively implies that $l(\beta_0) = o(1/\sqrt{n})$ because by a first order Taylor expansion of $n\hat{Q}_n\left(\beta_0 + \frac{h}{\sqrt{n}}\right) - nQ\left(\beta_0 + \frac{h}{\sqrt{n}}\right)$ around $n\hat{Q}_n(\beta_0) - nQ(\beta_0)$, followed by a second order Taylor expansion of $nQ\left(\beta_0 + \frac{h}{\sqrt{n}}\right)$ around $nQ(\beta_0)$,

$$\begin{aligned} n\hat{Q}_n\left(\beta_0 + \frac{h}{\sqrt{n}}\right) - n\hat{Q}_n(\beta_0) &= nQ\left(\beta_0 + \frac{h}{\sqrt{n}}\right) - nQ(\beta_0) + h'\sqrt{n}(\hat{l}_n(\beta_0) - l(\beta_0)) + o_P(1) \\ &= h'\sqrt{n}l(\beta_0) + \frac{1}{2}h'H_0h + h'\sqrt{n}(\hat{l}_n(\beta_0) - l(\beta_0)) + o_P(1) \end{aligned}$$

In order for assumption 2 to hold, it must be that $l(\beta_0) = o(1/\sqrt{n})$. When $l(\beta_0)$ does not depend on n , this implies $l(\beta_0) = 0$. We will relax this assumption in section 3.3 to allow for $l(\beta_0) \neq 0$.

The next assumption is needed to show that $\sqrt{n}(\hat{l}_n(\beta_0) - l(\beta_0))$ and $\sqrt{n}(\hat{l}_n^*(\beta_0) - \hat{l}_n(\beta_0))$ have the same asymptotic distribution.

Assumption 3. *There exists a function $g : \mathcal{X} \mapsto \mathbb{R}$ indexed by a parameter $\beta \in \mathbb{R}^d$ such that for any $\beta \in \mathbb{R}^d$, $\sqrt{n}(\hat{l}_n(\beta) - l(\beta)) = \sqrt{n}(P_n - P)g(\cdot, \beta) + o_P(1)$ and $\sqrt{n}(\hat{l}_n^*(\beta) - \hat{l}_n(\beta)) = \sqrt{n}(P_n^* - P_n)g(\cdot, \beta) + o_P^*(1)$, where $\lim_{n \rightarrow \infty} P\|g(\cdot, \beta_0)\|^2 \mathbf{1}(\|g(\cdot, \beta_0)\| > \epsilon\sqrt{n}) = 0$ for each $\epsilon > 0$.*

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

Assumption 4. *(i) $\mathcal{G}_R \equiv \{g(\cdot, \beta) - g(\cdot, \beta_0) : \|\beta - \beta_0\| \leq R\}$ is a Donsker class for some $R > 0$ and $P(g(\cdot, \beta) - g(\cdot, \beta_0))^2 \rightarrow 0$ for $\beta \rightarrow \beta_0$. (ii) $\lim_{\lambda \rightarrow \infty} \limsup_{n \rightarrow \infty} \sup_{t \geq \lambda} t^2 P \left\{ \sup_{g(\cdot, \beta) \in \mathcal{G}_{\delta_n}} \left\| \frac{g(\cdot, \beta) - g(\cdot, \beta_0)}{1 + \sqrt{n}\|\beta - \beta_0\|} \right\| > t \right\} = 0$ for any $\delta_n \rightarrow 0$.*

(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

$$\sup_{g(\cdot, \beta) \in \mathcal{G}_{\delta_n}} \left\| \frac{g(\cdot, \beta) - g(\cdot, \beta_0)}{1 + \sqrt{n} \|\beta - \beta_0\|} \right\| \leq \kappa \text{ for some constant } \kappa > 0 \text{ and any } \delta_n \rightarrow 0.$$

Our first theorem shows that the proximal bootstrap can consistently estimate the non-standard distribution of constrained estimators with fixed constraint sets. 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 $\beta_0 = \arg \min_{\beta \in C} Q(\beta)$, and $Q(\beta) = Q(\beta_0) + \frac{1}{2}(\beta - \beta_0)' H_0 (\beta - \beta_0) + o(\|\beta - \beta_0\|^2)$. For any $\bar{\beta}_n$ such that $\sqrt{n}(\bar{\beta}_n - \beta_0) = O_P(1)$ and $\bar{H}_n \xrightarrow{P} H_0$, let*

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

and $\hat{\beta}_n^* - \hat{\beta}_n = o_P^*(1)$. Then for any sequence α_n such that $\alpha_n \rightarrow 0$ and $\sqrt{n}\alpha_n \rightarrow \infty$, $\sqrt{n}(\hat{\beta}_n - \beta_0) \rightsquigarrow \mathcal{J}$ and $\frac{\hat{\beta}_n^* - \hat{\beta}_n}{\alpha_n} \xrightarrow[\mathbb{W}]{\mathbb{P}} \mathcal{J}$, where $\mathcal{J} = \arg \min_{h \in T_C(\beta_0)} \{h' W_0 + \frac{1}{2} h' H_0 h\}$, $T_C(\beta_0) = \lim_{\tau \downarrow 0} \frac{C - \beta_0}{\tau}$, and $W_0 \sim N(0, P(g(\cdot, \beta_0) - Pg(\cdot, \beta_0))(g(\cdot, \beta_0) - Pg(\cdot, \beta_0))')$.

Remark 1. We can remove the assumption that C is a closed set by assuming instead that $\mathcal{J} = \arg \min_{h \in T_C(\beta_0)} \{h' W_0 + \frac{1}{2} h' H_0 h\}$ 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 2. A special case is when $\beta_0 = \arg \min_{\beta \in C} Q(\beta)$ lies in the interior of C . Then $T_C(\beta_0) = \mathbb{R}^d$ and \mathcal{J} is multivariate normal. Another special case of C is when there are only equality constraints: $C = \{\beta \in \mathbb{R}^d : f(\beta) = 0\}$ 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 3. Note that the assumption that $Q(\beta)$ has a local quadratic approximation at β_0 of the

form $Q(\beta) = Q(\beta_0) + \frac{1}{2}(\beta - \beta_0)' H_0 (\beta - \beta_0) + o(\|\beta - \beta_0\|^2)$ effectively assumes $l(\beta_0) = 0$ (this is noted on the top of page 2000 of [Geyer \(1994\)](#)). 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:

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

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

3.3 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 continuously differentiable over a parameter space $\mathbb{B} \subset \mathbb{R}^d$ which is assumed to be compact.

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

We will define the population analog of C to be $C_0 \equiv \{\beta \in \mathbb{B} : f_{0j}(\beta) = 0 \text{ for } j \in \mathcal{E}, f_{0j}(\beta) \leq 0 \text{ for } j \in \mathcal{I}\}$, 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 require the population inequality constraints to be fixed, rather than drifting towards zero. The pseudo-true parameters are defined as $\beta_0 \equiv \arg \min_{\beta \in C_0} Q(\beta)$, where $Q(\beta)$ is a lower semicontinuous function that is twice differentiable at β_0 and $\sup_{\beta \in \mathbb{B}} \left| \hat{Q}_n(\beta) - Q(\beta) \right| = o_P(1)$. The pseudo-true parameters differ from the true parameters when the constraints are violated at the unconstrained optimum. If the constraints are satisfied at the unconstrained optimum, then they are the same.

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 sufficient and necessary to ensure that the set of optimal Lagrange multipliers that satisfy the first order KKT conditions is a singleton ([Wachsmuth \(2013\)](#)). It is also possible

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 $F_{0j} \equiv \frac{\partial f_{0j}(\beta)}{\partial \beta} \Big|_{\beta=\beta_0}$ 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 expansion 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: $l(\beta_0) \neq 0$.

Assumption 6. *Suppose $f_{nj} : \mathbb{B} \mapsto \mathbb{R}$ and $f_{0j} : \mathbb{B} \mapsto \mathbb{R}$ are twice continuously differentiable functions. Define λ_n to be the unique optimal Lagrange multipliers for $\hat{\beta}_n$. Define $\hat{\mathcal{L}}_n(\beta) \equiv \hat{Q}_n(\beta) + \sum_{j \in \mathcal{E} \cup \mathcal{I}} \lambda_{nj} f_{nj}(\beta)$, $F_{nj}(\beta_0) \equiv \frac{\partial f_{nj}(\beta)}{\partial \beta} \Big|_{\beta=\beta_0}$, and $G_{0j} \equiv \frac{\partial^2 f_{0j}(\beta)}{\partial \beta \partial \beta'} \Big|_{\beta=\beta_0}$. Suppose $B_0 = H_0 + \sum_{j \in \mathcal{E} \cup \mathcal{I}} \lambda_{0j} G_{0j}$ is symmetric, positive definite. For any $\delta_n \rightarrow 0$,*

$$\sup_{\frac{\|h\|}{\sqrt{n}} \leq \delta_n} \left| \frac{n\hat{\mathcal{L}}_n\left(\beta_0 + \frac{h}{\sqrt{n}}\right) - n\hat{\mathcal{L}}_n(\beta_0) - h' \sqrt{n} \hat{l}_n(\beta_0) - \frac{1}{2} h' H_0 h - \sum_{j \in \mathcal{E} \cup \mathcal{I}} \lambda_{0j} (\sqrt{n} F_{nj}(\beta_0)' h + \frac{1}{2} h' G_{0j} h)}{1 + \|h\|^2} \right|$$

$$= o_P(1)$$

where λ_0 are the unique Lagrange multipliers that satisfy $\lambda_{0j} f_{0j}(\beta_0) = 0$ for all $j \in \mathcal{E} \cup \mathcal{I}$, $0 \leq \lambda_{0j} < \infty$ for all $j \in \mathcal{E} \cup \mathcal{I}$, and $\nabla \mathcal{L}(\beta_0, \lambda_0) \equiv l(\beta_0) + \sum_{j \in \mathcal{E} \cup \mathcal{I}} \lambda_{0j} F_{0j} = 0$.

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 $\sqrt{n}(\bar{\beta}_n - \beta_0) = O_P(1)$, let $\bar{F}_{nj} \equiv F_{nj}(\bar{\beta}_n)$, $\bar{F}_{nj}^* \equiv F_{nj}^*(\bar{\beta}_n)$, $\bar{G}_{nj} \xrightarrow{p} \frac{\partial^2 f_{0j}(\beta)}{\partial \beta \partial \beta'} \Big|_{\beta=\bar{\beta}_n}$ for all j , and let $\bar{\lambda}_n$ be the unique optimal Lagrange multipliers

for $\bar{\beta}_n$. Define $\hat{\beta}_n^* \equiv \arg \min_{\beta \in C^*} \hat{A}_n^*(\beta)$, for

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

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). Furthermore, the proximal bootstrap is able to consistently estimate the non-standard asymptotic distribution of certain 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 Assumptions 3-4 and 5 - 6 are satisfied in addition to the following:*

(i) *Suppose $\hat{\beta}_n \xrightarrow{P} \beta_0$, where both $\hat{\beta}_n$ and $\beta_0 \equiv \arg \min_{\beta \in C_0} Q(\beta)$ are unique.*

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

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

$$\sqrt{n} (\hat{l}_n^*(\beta_0) - \hat{l}_n(\beta_0)) + \sum_{j \in \mathcal{E} \cup \mathcal{I}} \lambda_{0j} \sqrt{n} (F_n^*(\beta_0) - F_n(\beta_0)) \rightsquigarrow_{\mathbb{W}} W_0 + \sum_{j \in \mathcal{E} \cup \mathcal{I}} \lambda_{0j} V_{0j},$$

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

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

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

Then, for any sequence α_n such that $\alpha_n \rightarrow 0$ and $\sqrt{n}\alpha_n \rightarrow \infty$, $\sqrt{n}(\hat{\beta}_n - \beta_0) \rightsquigarrow \mathcal{J}$ and $\frac{\hat{\beta}_n^* - \hat{\beta}_n}{\alpha_n} \rightsquigarrow_{\mathbb{W}} \mathbb{W}$, where for $\mathcal{I}_+^*(\lambda_0) \equiv \{j \in \mathcal{I}^* : \lambda_{0j} > 0\}$ and $\mathcal{I}_0^*(\lambda_0) \equiv \{j \in \mathcal{I}^* : \lambda_{0j} = 0\}$,

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

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

Remark 4. If $l(\beta_0) = 0$, which is implied by $Q(\beta) = Q(\beta_0) + \frac{1}{2}(\beta - \beta_0)'H_0(\beta - \beta_0) + o(\|\beta - \beta_0\|^2)$, then \mathcal{J} reduces down to

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

$$\Sigma = \{h : U_{0j} + F'_{0j}h = 0 \text{ for } j \in \mathcal{E}, U_{0j} + F'_{0j}h \leq 0 \text{ for } j \in \mathcal{I}_0^*(\lambda_0)\}$$

This is because by the KKT conditions, λ_0 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 $j \in \mathcal{I} \setminus \mathcal{I}^*$ 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 $\sum_{j \in \mathcal{E} \cup \mathcal{I}} \lambda_{0j} (h'V_{0j} + \frac{1}{2}h'G_{0j}h) = 0$, $\mathcal{I}_+^*(\lambda_0) = \emptyset$, 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 $\hat{\beta}_n^* \equiv \arg \min_{\beta \in C^*} \hat{A}_n^*(\beta)$, where

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

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

Remark 5. If there are only equality constraints, then the asymptotic distribution becomes $\mathcal{J} = \arg \min_{h \in \Sigma} \left\{ h'W_0 + \frac{1}{2}h' \left(H_0 + \sum_{j \in \mathcal{E}} \lambda_{0j} G_{0j} \right) h \right\}$ for $\Sigma = \left\{ h : U_{0j} + F'_{0j}h = 0 \text{ for } j \in \mathcal{E} \right\}$. Using standard arguments in Amemiya (1985) section 1.4.1 or Newey and McFadden (1994) section 9.1 (which are repeated in Lemma 6.1 in the appendix), $J = -B_0^{-1} \left(I - F_0 (F'_0 B_0^{-1} F_0)^{-1} F'_0 B_0^{-1} \right) W_0 - B_0^{-1} F_0 (F'_0 B_0^{-1} F_0)^{-1} U_0$. 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 $\mathcal{J} = -H_0^{-1} \left(I - F_0 (F'_0 H_0^{-1} F_0)^{-1} F'_0 H_0^{-1} \right) W_0 - H_0^{-1} F_0 (F'_0 H_0^{-1} F_0)^{-1} U_0$.

Remark 6. 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

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

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 $\mathcal{J} = -B_0^{-1} \left(I - F_0 (F'_0 B_0^{-1} F_0)^{-1} F'_0 B_0^{-1} \right) \left(W_0 + \sum_{j \in \mathcal{E} \cup \mathcal{I}_+^*(\lambda_0)} \lambda_{0j} V_{0j} \right) - B_0^{-1} F_0 (F'_0 B_0^{-1} F_0)^{-1} U_0$, 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 $\sum_{j \in \mathcal{E} \cup \mathcal{I}^*} \lambda_{0j} (h'V_{0j} + \frac{1}{2}h'G_{0j}h) = 0$ and $\mathcal{I}_+^*(\lambda_0) = \emptyset$, so \mathcal{J} reduces down to $\mathcal{J} = \arg \min_{h \in \Sigma} \left\{ h'W_0 + \frac{1}{2}h'H_0h \right\}$, for $\Sigma = \left\{ h : U_{0j} + F'_{0j}h = 0 \text{ for } j \in \mathcal{E} \right\}$.

Remark 7. If there are only inequality constraints, we can also obtain a closed form expression for \mathcal{J} . Because $\Sigma = \left\{ h : U_{0j} + F'_{0j}h = 0 \text{ for } j \in \mathcal{I}_+^*(\lambda_0), U_{0j} + F'_{0j}h \leq 0 \text{ for } j \in \mathcal{I}_0^*(\lambda_0) \right\}$ in this case, it follows from Lemma 6.2 in the appendix that

$$\mathcal{J} = \max \left\{ -B_0^{-1} \left(I - F_{0+} (F'_{0+} B_0^{-1} F_{0+})^{-1} F'_{0+} B_0^{-1} \right) \left(W_0 + \sum_{j \in \mathcal{I}_+^*(\lambda_0)} \lambda_{0j} V_{0j} \right) \right\}$$

$$\left. -B_0^{-1}F_{0+} (F_{0+}'B_0^{-1}F_{0+})^{-1}U_{0+}, -B_0^{-1} \left(W_0 + \sum_{j \in \mathcal{I}_+^*(\lambda_0)} \lambda_{0j}V_{0j} \right) \right\}$$

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 $\mathcal{I}_+^*(\lambda_0) = \emptyset$, 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 8. In the case of fixed 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

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

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

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

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

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

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

$$\begin{aligned} & n\hat{Q}_n \left(\beta_0 + \frac{h}{\sqrt{n}} \right) - n\hat{Q}_n(\beta_0) + \sum_{j \in \mathcal{E} \cup \mathcal{I}} \lambda_{nj}n \left(f_{0j} \left(\beta_0 + \frac{h}{\sqrt{n}} \right) - f_{0j}(\beta_0) \right) \\ &= h' \sqrt{n} \left(\hat{l}_n(\beta_0) - l(\beta_0) \right) + \frac{1}{2} h' H_0 h + \sum_{j \in \mathcal{E} \cup \mathcal{I}} \lambda_{0j} \left(\sqrt{n} (F_{0j} - F_{0j}') h + \frac{1}{2} h' G_{0j} h \right) + o_P(1) \\ &= h' \sqrt{n} \left(\hat{l}_n(\beta_0) - l(\beta_0) \right) + \frac{1}{2} h' H_0 h + \frac{1}{2} \sum_{j \in \mathcal{E} \cup \mathcal{I}} \lambda_{0j} h' G_{0j} h + o_P(1) \\ &\rightsquigarrow h' W_0 + \frac{1}{2} h' H_0 h + \frac{1}{2} \sum_{j \in \mathcal{E} \cup \mathcal{I}} \lambda_{0j} h' G_{0j} h \end{aligned}$$

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

$$\begin{aligned} \mathcal{J} &= \arg \min_{h \in \Sigma} \left\{ h'W_0 + \frac{1}{2}h'H_0h + \frac{1}{2} \sum_{j \in \mathcal{E} \cup \mathcal{I}} \lambda_{0j} h'G_{0j}h \right\} \\ \Sigma &= \{h : F'_{0j}h = 0 \text{ for } j \in \mathcal{E} \cup \mathcal{I}^*_+(\lambda_0), F'_{0j}h \leq 0 \text{ for } j \in \mathcal{I}^*_0(\lambda_0)\} \end{aligned}$$

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

$$\begin{aligned} \mathcal{J} &= \arg \min_{h \in \Sigma} \left\{ h'W_0 + \frac{1}{2}h'H_0h \right\} \\ \Sigma &= \{h : F'_{0j}h = 0 \text{ for } j \in \mathcal{E}, F'_{0j}h \leq 0 \text{ for } j \in \mathcal{I}^*\} \end{aligned}$$

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 $\mu \geq 0$ 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 9. Alternatively, we can define the proximal bootstrap estimator as $\hat{\beta}_n^* = \arg \min_{\beta \in C^*} \hat{A}_n^*(\beta)$, where

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

The feasible direction set is

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

and the linearized feasible direction set is

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

Note that since $\frac{f_{nj}(\beta_0)}{\alpha_n} + F_{nj}(\beta_0)' h + \sqrt{n} (f_{nj}^*(\bar{\beta}_n) - f_{nj}(\bar{\beta}_n)) \xrightarrow{P} -\infty$ for $j \in \mathcal{I} \setminus \mathcal{I}^*$, the nonactive inequality constraints do not affect the asymptotic distribution under our pointwise asymptotics. Since $\frac{f_{nj}(\beta_0)}{\alpha_n} = \frac{\sqrt{n}(f_{nj}(\beta_0) - f_{0j}(\beta_0))}{\sqrt{n}\alpha_n} = o_P(1)$ for all $j \in \mathcal{E} \cup \mathcal{I}^*$, $\sqrt{n} (f_{nj}^*(\beta_0) - f_{nj}(\beta_0)) \xrightarrow[\mathbb{W}]{\mathbb{P}} U_{0j}$, jointly, for all $j \in \mathcal{E} \cup \mathcal{I}^*$, and $\sqrt{n} (f_{nj}^*(\bar{\beta}_n) - f_{nj}(\bar{\beta}_n)) \xrightarrow[\mathbb{W}]{\mathbb{P}} U_{0j}$, jointly, for all $j \in \mathcal{E} \cup \mathcal{I}^*$, it follows that $\mathbb{1}(h \notin \Sigma_n^*) \xrightarrow[e-d]{P} \mathbb{1}(h \notin \Sigma)$.

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

Remark 10. 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 $b_1 = 1 \dots B_1$. Conditional on each of these bootstrap samples $b_1 = 1 \dots B_1$, draw B_2 bootstrap samples and compute bootstrap estimates $\hat{\beta}_n^{(*,b_1,b_2)}$ for $b_2 = 1 \dots B_2$. Compute the empirical frequency with which equal tailed intervals centered at $\hat{\beta}_n^{(*,b_1)}$ cover $\hat{\beta}_n$. If the current value of α_n achieves coverage at or above the nominal frequency, 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.

4 Monte Carlo Simulations

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

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

We solve the following constrained GMM problem:

$$\hat{\beta}_n = \arg \min_{\beta \geq 0} \hat{Q}_n(\beta) = \hat{\pi}_n(\beta)' \hat{\pi}_n(\beta)$$

We use Matlab's built-in fmincon solver to compute $\bar{\beta}_n = \hat{\beta}_n$ and also

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

$$\hat{l}_n(\bar{\beta}_n) = \hat{G}_n' \hat{\pi}_n(\bar{\beta}_n), \hat{l}_n^*(\bar{\beta}_n) = \hat{G}_n^{*'} \hat{\pi}_n^*(\bar{\beta}_n), \text{ and}$$

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

$$K_h(x) = K(x/h), K(x) = (2\pi)^{-1/2} \exp(-x^2/2), K_h'(x) = K'(x/h) \text{ and } K'(x) = -(2\pi)^{-1/2} x \exp(-x^2/2).$$

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 : $\alpha_n \in \{n^{-1/3}, n^{-1/4}, n^{-1/6}, n^{-1/8}, n^{-1/10}\}$. We use 5000 bootstrap iterations and 2000 Monte Carlo simulations. Empirical coverage frequencies for equal-tailed nominal 95% confidence intervals and average interval lengths are reported in table 1.

Table 1: Proximal Bootstrap Coverage Frequencies and Interval Lengths

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

We now compare the proximal bootstrap with the centered standard bootstrap estimator $\hat{\beta}_n^{**} = \arg \min_{\beta \in \mathcal{C}} \left(\hat{\pi}_n^*(\beta) - \hat{\pi}_n(\hat{\beta}_n) \right)' \left(\hat{\pi}_n^*(\beta) - \hat{\pi}_n(\hat{\beta}_n) \right)$. Empirical coverage frequencies for equal-tailed nominal 95% confidence intervals and average interval lengths are reported in table 2. Interestingly, the coverage frequencies are similar, although the intervals are wider.

Table 2: Standard Bootstrap Coverage Frequencies and Interval Lengths

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)

4.2 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 $i = 1 \dots n$ from picking choice $j = 1 \dots J$ is given by

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

and $\epsilon_{ij} \stackrel{i.i.d.}{\sim}$ 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 $P_{ij} \equiv \frac{\exp(\beta x_{ij})}{\sum_l \exp(\beta x_{il})}$,

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

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

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

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

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

We consider $n \in \{100, 500, 1000\}$, $J = 20$, and $\alpha_n \in \{n^{-1/3}, n^{-1/4}, n^{-1/6}, n^{-1/8}, n^{-1/10}\}$. Empirical coverage frequencies for equal-tailed nominal 95% confidence intervals and average interval lengths are reported in table 3. 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 4. 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

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

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

are reported in table 5. The coverage is less than that for the proximal bootstrap for all values of n and the intervals are sometimes wider.

4.3 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.

Table 5: 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)

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 $\theta_1 = 2.4569$, 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 $\left((x_t^i, d_t^i)_{t=1}^T \right)_{i=1}^M$, 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 .

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

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 $\hat{\mathbf{x}} = \{\hat{x}_1, \hat{x}_2, \dots, \hat{x}_K\}$:

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

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 $\hat{\mathbf{x}} = \{1, 2, 3, \dots, 175\}$. 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 $p_3(x' = \hat{x}_{1+j} | \hat{x}_k, 1, \theta_3) = \theta_{3j}$. If the engine is not replaced, the transition probabilities are $p_3(x' = \hat{x}_{k+j} | \hat{x}_k, 0, \theta_3) = \theta_{3j}$. The only values of the choice-specific value functions we need to estimate are the ones corresponding to no replacement $EV = [EV(\hat{x}_1, 0), EV(\hat{x}_2, 0), \dots, EV(\hat{x}_K, 0)]$ 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 $\left((x_t^i, d_t^i)_{t=1}^T \right)_{i=1}^M$. Define $\theta \equiv (\theta_1, \theta_3', RC, EV)'$ and $C = \{f_j(\theta) = 0 \text{ for } j \in \mathcal{E}, f_j(\theta) \leq 0 \text{ for } j \in \mathcal{I}\}$, where $f_j(\theta)$ includes the EV fixed point equations (1) as well as the constraints on the transition probabilities satisfying $0 \leq \theta_3 \leq 1$ 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 $\alpha_M \rightarrow 0$ and $\sqrt{M}\alpha_M \rightarrow \infty$, and a \sqrt{M} -consistent estimator $\bar{\theta}_M$, the proximal bootstrap estimator is given by

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

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

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 6-8 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 $\alpha_M \in \{M^{-1/3}, M^{-1/4}, M^{-1/6}\}$. The number of time periods is $T = 120$. 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 6: 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 (0.520)	0.946 (0.373)	0.949 (0.264)	0.942 (0.187)	0.948 (0.167)	0.949 (0.152)
θ_{30}	0.951 (0.005)	0.947 (0.003)	0.945 (0.002)	0.933 (0.002)	0.932 (0.001)	0.935 (0.001)
θ_{31}	0.955 (0.008)	0.944 (0.006)	0.951 (0.004)	0.948 (0.003)	0.94 (0.003)	0.947 (0.002)
θ_{32}	0.949 (0.008)	0.952 (0.006)	0.944 (0.004)	0.942 (0.003)	0.942 (0.003)	0.952 (0.002)
θ_{33}	0.957 (0.002)	0.95 (0.001)	0.949 (0.001)	0.951 (0.001)	0.96 (0.001)	0.957 (0.001)
RC	0.927 (1.683)	0.95 (1.204)	0.949 (0.853)	0.946 (0.604)	0.946 (0.540)	0.947 (0.492)

Table 7: 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 (0.520)	0.949 (0.372)	0.95 (0.264)	0.941 (0.187)	0.949 (0.167)	0.95 (0.153)
θ_{30}	0.952 (0.005)	0.948 (0.003)	0.94 (0.002)	0.935 (0.002)	0.934 (0.001)	0.937 (0.001)
θ_{31}	0.954 (0.008)	0.942 (0.006)	0.95 (0.004)	0.946 (0.003)	0.944 (0.003)	0.948 (0.002)
θ_{32}	0.952 (0.008)	0.95 (0.006)	0.941 (0.004)	0.943 (0.003)	0.939 (0.003)	0.948 (0.002)
θ_{33}	0.959 (0.002)	0.95 (0.001)	0.95 (0.001)	0.949 (0.001)	0.958 (0.001)	0.958 (0.001)
RC	0.927 (1.683)	0.952 (1.204)	0.95 (0.853)	0.945 (0.604)	0.949 (0.540)	0.952 (0.493)

Table 8: 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 (0.520)	0.947 (0.372)	0.949 (0.264)	0.943 (0.187)	0.948 (0.167)	0.95 (0.152)
θ_{30}	0.952 (0.005)	0.95 (0.003)	0.941 (0.002)	0.933 (0.002)	0.933 (0.001)	0.94 (0.001)
θ_{31}	0.955 (0.008)	0.942 (0.006)	0.949 (0.004)	0.949 (0.003)	0.944 (0.003)	0.953 (0.002)
θ_{32}	0.951 (0.008)	0.951 (0.006)	0.942 (0.004)	0.944 (0.003)	0.94 (0.003)	0.946 (0.002)
θ_{33}	0.96 (0.002)	0.951 (0.001)	0.949 (0.001)	0.951 (0.001)	0.959 (0.001)	0.954 (0.001)
RC	0.925 (1.666)	0.951 (1.201)	0.95 (0.852)	0.947 (0.603)	0.948 (0.540)	0.944 (0.493)

5 Conclusion

We have demonstrated how to use a computationally efficient bootstrap procedure to consistently estimate the limiting distribution of \sqrt{n} -consistent constrained optimization estimators with non-standard 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.

6 Appendix

6.1 Proofs of Theorems

6.1.1 Proof of Theorem 2

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, $Q(\beta) = Q(\beta_0) + \frac{1}{2}(\beta - \beta_0)' H_0(\beta - \beta_0) + o(\|\beta - \beta_0\|^2)$, and $\hat{\beta}_n \xrightarrow{P} \beta_0$ imply that the conditions of Lemma 4.3 in [Geyer \(1994\)](#) are satisfied, and therefore $\sqrt{n}(\hat{\beta}_n - \beta_0) = O_P(1)$.

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

$$\begin{aligned}\sqrt{n}(\hat{\beta}_n - \beta_0) &= \arg \min_{h \in \sqrt{n}(C - \beta_0)} \left\{ n\hat{Q}_n\left(\beta_0 + \frac{h}{\sqrt{n}}\right) - n\hat{Q}_n(\beta_0) \right\} \\ &= \arg \min_{h \in \sqrt{n}(C - \beta_0)} \left\{ h' \sqrt{n}(\hat{l}_n(\beta_0) - l(\beta_0)) + \frac{1}{2}h'H_0h + o_P(1) \right\}\end{aligned}$$

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

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

$$n\hat{Q}_n\left(\beta_0 + \frac{h}{\sqrt{n}}\right) - n\hat{Q}_n(\beta_0) \rightsquigarrow h'W_0 + \frac{1}{2}h'H_0h$$

as a process indexed by h in the space of bounded functions $\ell^\infty(K)$ for any compact $K \subset \mathbb{R}^d$. Since $h'W_0 + \frac{1}{2}h'H_0h$ has a continuous sample path, according to page 5 of Knight (1999),

$$n\hat{Q}_n\left(\beta_0 + \frac{h}{\sqrt{n}}\right) - n\hat{Q}_n(\beta_0) \rightarrow_{u-d} h'W_0 + \frac{1}{2}h'H_0h$$

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

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

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

$$n\hat{Q}_n\left(\beta_0 + \frac{h}{\sqrt{n}}\right) - n\hat{Q}_n(\beta_0) + \infty 1(h \notin \sqrt{n}(C - \beta_0)) \rightarrow_{e-d} h'W_0 + \frac{1}{2}h'H_0h + \infty 1(h \notin T_C(\beta_0))$$

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 $h'W_0 + \frac{1}{2}h'H_0h$ almost surely has a unique minimizer over $T_C(\beta_0)$ due to C being a closed set (see Proposition 4.2 and Theorem

4.4 of [Geyer \(1994\)](#)),

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

For the proximal bootstrap, since $\sqrt{n}\alpha_n \rightarrow \infty$ and $\sqrt{n}(\hat{\beta}_n - \beta_0) = O_P(1)$, $\frac{\hat{\beta}_n^* - \hat{\beta}_n}{\alpha_n} = \frac{\hat{\beta}_n^* - \beta_0}{\alpha_n} + o_P^*(1)$,

where

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

Assumption 4 implies $\sqrt{n}(\hat{l}_n^*(\bar{\beta}_n) - \hat{l}_n(\bar{\beta}_n))$ and $\sqrt{n}(\hat{l}_n^*(\beta_0) - \hat{l}_n(\beta_0))$ have the same asymptotic distribution. Therefore,

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

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

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

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

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

■

6.1.2 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 the population objective $F(\cdot)$ with the population Lagrangian $\mathcal{L}(\beta_0, \lambda_0) \equiv Q(\beta_0) + \sum_{j \in \mathcal{E} \cup \mathcal{I}} \lambda_{0j} f_{0j}(\beta_0)$ and the empirical process $E_n \Delta$ with $\sqrt{n} \left(\hat{l}_n(\beta_0) - l(\beta_0) \right) + \sum_{j \in \mathcal{E} \cup \mathcal{I}} \lambda_{0j} \sqrt{n} (F_{nj}(\beta_0) - F_{0j}(\beta_0))$. The first order KKT condition $\nabla \mathcal{L}(\beta_0, \lambda_0) \equiv l(\beta_0) + \sum_{j \in \mathcal{E} \cup \mathcal{I}} \lambda_{0j} F_{0j}(\beta_0) = 0$ and positive-definiteness of $\nabla^2 \mathcal{L}(\beta_0, \lambda_0) \equiv H_0 + \sum_{j \in \mathcal{E} \cup \mathcal{I}} \lambda_{0j} G_{0j}$ imply the local quadratic expansion $\mathcal{L}(\beta, \lambda_0) = \mathcal{L}(\beta_0, \lambda_0) + \frac{1}{2} \|\beta - \beta_0\|_{\nabla^2 \mathcal{L}(\beta_0, \lambda_0)}^2 + o(\|\beta - \beta_0\|^2)$ for β in a small neighborhood of β_0 . The rest of the arguments are the same as [Pollard \(1984\)](#).

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_n for $\hat{\beta}_n$. It is well known that $\hat{\beta}_n = \arg \min_{\beta \in C} \hat{Q}_n(\beta)$ can be equivalently expressed as $\hat{\beta}_n = \arg \min_{\beta \in C} \hat{\mathcal{L}}_n(\beta)$ 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 (and also the weaker Mangasarian-Fromovitz and Abadie Constraint Qualifications) 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); in particular, the tangent cone of C at β_0 is equal to the linearized feasible direction set at β_0 . 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

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

Denote the linearized feasible direction set by

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

Under LICQ, minimizing the Lagrangian over \mathcal{F}_n is asymptotically equivalent to minimizing the Lagrangian over Σ_n :

$$\begin{aligned} \sqrt{n}(\hat{\beta}_n - \beta_0) &= \arg \min_{h \in \mathcal{F}_n} \left\{ n\hat{\mathcal{L}}_n \left(\beta_0 + \frac{h}{\sqrt{n}} \right) - n\hat{\mathcal{L}}_n(\beta_0) \right\} \\ &= \arg \min_{h \in \Sigma_n} \left\{ n\hat{\mathcal{L}}_n \left(\beta_0 + \frac{h}{\sqrt{n}} \right) - n\hat{\mathcal{L}}_n(\beta_0) \right\} + o_P(1) \\ &= \arg \min_{h \in \Sigma_n} \left\{ n\hat{Q}_n \left(\beta_0 + \frac{h}{\sqrt{n}} \right) - n\hat{Q}_n(\beta_0) + \sum_{j \in \mathcal{E} \cup \mathcal{I}} \lambda_{nj}n \left(f_{nj} \left(\beta_0 + \frac{h}{\sqrt{n}} \right) - f_{nj}(\beta_0) \right) \right\} + o_P(1) \\ &\rightsquigarrow \arg \min_{h \in \Sigma} \left\{ h'W_0 + \frac{1}{2}h'H_0h + \sum_{j \in \mathcal{E} \cup \mathcal{I}_+^*(\lambda_0)} \lambda_{0j} \left(h'V_{0j} + \frac{1}{2}h'G_{0j}h \right) \right\} = \mathcal{J} \end{aligned}$$

where the last line follows from the following arguments. First note that assumption 6 in combination with $\nabla \mathcal{L}(\beta_0, \lambda_0) \equiv l(\beta_0) + \sum_{j \in \mathcal{E} \cup \mathcal{I}} \lambda_{0j}F_{0j}(\beta_0) = 0$ implies that for any $\delta_n \rightarrow 0$,

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

Therefore, uniformly in h ,

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

Recall $\sqrt{n}(\hat{l}_n(\beta_0) - l(\beta_0)) + \sum_{j \in \mathcal{E} \cup \mathcal{I}} \lambda_{0j}\sqrt{n}(F_{nj}(\beta_0) - F_{0j}(\beta_0)) \rightsquigarrow W_0 + \sum_{j \in \mathcal{E} \cup \mathcal{I}} \lambda_{0j}V_{0j}$, and $\lambda_{0j} = 0$ for all $j \in \mathcal{I} \setminus \mathcal{I}_+^*(\lambda_0)$. 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,

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

$$\begin{aligned}
& \rightsquigarrow h'W_0 + \frac{1}{2}h'H_0h + \sum_{j \in \mathcal{E} \cup \mathcal{I}} \lambda_{0j} \left(h'V_{0j} + \frac{1}{2}h'G_{0j}h \right) \\
& = h'W_0 + \frac{1}{2}h'H_0h + \sum_{j \in \mathcal{E} \cup \mathcal{I}_+^*(\lambda_0)} \lambda_{0j} \left(h'V_{0j} + \frac{1}{2}h'G_{0j}h \right)
\end{aligned}$$

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

Note that since $\sqrt{n}f_{nj}(\beta_0) + F_{nj}(\beta_0)'h \xrightarrow{P} -\infty$ for $j \in \mathcal{I} \setminus \mathcal{I}^*$, the nonactive inequality constraints do not affect the asymptotic distribution under our pointwise asymptotics. Since $\sqrt{n}f_{nj}(\beta_0) \rightsquigarrow U_{0j}$, jointly, for all $j \in \mathcal{E} \cup \mathcal{I}^*$, $F_n(\beta_0) = F_0 + o_P(1)$, and finite dimensional convergence in distribution implies epi-convergence in distribution for convex functions,

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

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

$$\begin{aligned}
\hat{\mathbb{G}}_n(h) &= n\hat{Q}_n\left(\beta_0 + \frac{h}{\sqrt{n}}\right) - n\hat{Q}_n(\beta_0) + \sum_{j \in \mathcal{E} \cup \mathcal{I}} \lambda_{nj}n \left(f_{nj}\left(\beta_0 + \frac{h}{\sqrt{n}}\right) - f_{nj}(\beta_0) \right) + \infty 1(h \notin \Sigma_n) \\
\mathbb{G}_0(h) &= h'W_0 + \frac{1}{2}h'H_0h + \sum_{j \in \mathcal{E} \cup \mathcal{I}_+^*(\lambda_0)} \lambda_{0j} \left(h'V_{0j} + \frac{1}{2}h'G_{0j}h \right) + \infty 1(h \notin \Sigma)
\end{aligned}$$

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

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

$$\left. \frac{f_{nj}(\bar{\beta}_n)}{\alpha_n} + \bar{F}'_{nj}h + \sqrt{n}(f_{nj}^*(\bar{\beta}_n) - f_{nj}(\bar{\beta}_n)) + \bar{F}'_{nj} \left(\frac{\beta_0 - \bar{\beta}_n}{\alpha_n} \right) \leq 0 \text{ for } j \in \mathcal{I} \right\}$$

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

$$\infty 1(h \notin \Sigma_n^*) \xrightarrow{P} \infty 1(h \notin \{h : U_{0j} + F_{0j}'h = 0 \text{ for } j \in \mathcal{E}, U_{0j} + F_{0j}'h \leq 0 \text{ for } j \in \mathcal{I}^*\})$$

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

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

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

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

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

Next, note that assumption 4, $\max_{j \in \mathcal{E} \cup \mathcal{I}} |\bar{\lambda}_{nj} - \lambda_{0j}| \xrightarrow{p} 0$, and $\sup_{\|\beta - \beta_0\| \leq o(1)} \sqrt{n} (F_n^*(\beta) - F_n(\beta) - F_n^*(\beta_0) + F_n(\beta_0)) = o_P^*(1)$ imply $\sqrt{n} \left(\hat{\iota}_n^* (\bar{\beta}_n) - \hat{\iota}_n (\bar{\beta}_n) \right) + \sum_{j \in \mathcal{E} \cup \mathcal{I}} \bar{\lambda}_{nj} \sqrt{n} (\bar{F}_{nj}^* - \bar{F}_{nj}) \xrightarrow[\mathbb{W}]{\mathbb{P}} W_0 + \sum_{j \in \mathcal{E} \cup \mathcal{I}} \lambda_{0j} V_{0j}$ because

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

and we assumed $\sqrt{n} \left(\hat{\iota}_n^* (\beta_0) - \hat{\iota}_n (\beta_0) \right) + \sum_{j \in \mathcal{E} \cup \mathcal{I}} \lambda_{0j} \sqrt{n} (F_n^*(\beta_0) - F_n(\beta_0)) \xrightarrow[\mathbb{W}]{\mathbb{P}} W_0 + \sum_{j \in \mathcal{E} \cup \mathcal{I}} \lambda_{0j} V_{0j}$. Additionally, $\max_{j \in \mathcal{E} \cup \mathcal{I}} |\bar{G}_{nj} - G_{0j}| \xrightarrow{p} 0$ and $\max_{j \in \mathcal{E} \cup \mathcal{I}} |\bar{\lambda}_{nj} - \lambda_{0j}| \xrightarrow{p} 0$ 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,

$$\begin{aligned} & h' \sqrt{n} \left(\hat{\iota}_n^* (\bar{\beta}_n) - \hat{\iota}_n (\bar{\beta}_n) \right) + \frac{1}{2} h' H_0 h + \sum_{j \in \mathcal{E} \cup \mathcal{I}} \bar{\lambda}_{nj} \left(h' \sqrt{n} (\bar{F}_{nj}^* - \bar{F}_{nj}) + \frac{1}{2} h' G_{0j} h \right) \\ & \xrightarrow[\mathbb{W}]{\mathbb{P}} h' W_0 + \frac{1}{2} h' H_0 h + \sum_{j \in \mathcal{E} \cup \mathcal{I}} \lambda_{0j} \left(h' V_{0j} + \frac{1}{2} h' G_{0j} h \right) \\ &= h' W_0 + \frac{1}{2} h' H_0 h + \sum_{j \in \mathcal{E} \cup \mathcal{I}_+^*(\lambda_0)} \lambda_{0j} \left(h' V_{0j} + \frac{1}{2} h' G_{0j} h \right) \end{aligned}$$

as a process indexed by h in the space of bounded functions $\ell^\infty(K)$ 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, $\arg \min_h \hat{\mathbb{G}}_n^*(h) \xrightarrow[e-d]{p} \arg \min_h \mathbb{G}_0(h)$ for

$$\hat{\mathbb{G}}_n^*(h) = h' \sqrt{n} \left(\hat{\iota}_n^* (\bar{\beta}_n) - \hat{\iota}_n (\bar{\beta}_n) \right) + \frac{1}{2} h' \bar{H}_n h$$

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

■

6.2 Additional Results

6.2.1 Equality Constrained Quadratic Program

Lemma 6.1. *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

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

Proof: The Lagrangian and KKT conditions are

$$\begin{aligned}
\mathcal{L} & = h' \Delta_n + \frac{1}{2} h' H_0 h + \lambda \circ (R'h - \delta) \\
\Delta_n + H_0 h + R\lambda & = 0 \\
R'h - \delta & = 0
\end{aligned}$$

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

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

Therefore,

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

6.2.2 Inequality Constrained Quadratic Program

Lemma 6.2. *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*

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

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

Proof: The Lagrangian and KKT Conditions are

$$\begin{aligned} \mathcal{L} &= h' \Delta_n + \frac{1}{2} h' H_0 h + \sum_{i=1}^m \mu_i (R'_i h - \delta_i) \\ \Delta_n + H_0 h + R\mu &= 0 \\ \mu_i \geq 0, \mu_i (R'_i h - \delta_i) &= 0 \forall i = 1 \dots m \end{aligned}$$

The first KKT condition says $h^+ = -H_0^{-1} (\Delta_n + R\mu)$. The second says that if $\mu_i > 0$, then $R'_i h^+ - \delta_i = 0$, meaning the inequality constraint is strongly active (binding). 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 active constraints (since strict complementarity may not hold),

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

Therefore,

$$h^+ = -H_0^{-1} (\Delta_n + R_\Lambda \mu_\Lambda)$$

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

6.2.3 Consistency of Proximal Bootstrap with growing number of constraints in Remark 4

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

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

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 weakly to zero. To do so, we will assume

$$\sup_{t \in \mathbb{R}} \left| P \left(\max_{j \in \mathcal{E}_n \cup \mathcal{I}_n^*} \sqrt{n} f_{nj}(\beta_0) \leq t \right) - P \left(\max_{j \in \mathcal{E}_n \cup \mathcal{I}_n^*} U_{0j} \leq t \right) \right| \rightarrow 0, \text{ and}$$

$$\sup_{t \in \mathbb{R}} \left| P \left(\max_{j \in \mathcal{E}_n \cup \mathcal{I}_n^*} \sqrt{n} (f_{nj}^*(\bar{\beta}_n) - f_{nj}(\bar{\beta}_n)) \leq t \mid \mathcal{X}_n \right) - P \left(\max_{j \in \mathcal{E}_n \cup \mathcal{I}_n^*} U_{0j} \leq t \right) \right| \xrightarrow{P} 0. \text{ These assumptions can be derived using the results in } \text{Chernozhukov et al. (2013)} \text{ and } \text{Chernozhukov et al. (2019)}$$

for Gaussian approximation of maxima of sums for high dimensional random vectors. We will also need to assume $\max_{j \in \mathcal{E}_n \cup \mathcal{I}_n^*} |F_{nj}(\beta_0) - F_{0j}| = o_P(1)$ and $\max_{j \in \mathcal{E}_n \cup \mathcal{I}_n^*} |\bar{F}_{nj} - F_{0j}| = o_P(1)$.

For any h_1, \dots, h_k where k is fixed,

$$\begin{aligned}
&P(h_1 \in \Sigma_n, \dots, h_k \in \Sigma_n) \\
&= P \left(\bigcap_{i=1}^k \left\{ \sqrt{n} f_{nj}(\beta_0) + F_{nj}(\beta_0)' h_i = 0 \text{ for } j \in \mathcal{E}_n, \sqrt{n} f_{nj}(\beta_0) + F_{nj}(\beta_0)' h_i \leq 0 \text{ for } j \in \mathcal{I}_n \right\} \right)
\end{aligned}$$

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

where we have used $\sqrt{n} f_{nj}(\beta_0) + \max_{1 \leq i \leq k} F_{nj}(\beta_0)' h_i \xrightarrow{P} -\infty$ for $j \in \mathcal{I}_n \setminus \mathcal{I}_n^*$, $\max_{j \in \mathcal{E}_n \cup \mathcal{I}_n^*} |F_{nj}(\beta_0) - F_{0j}| = o_P(1)$, and $\sup_{t \in \mathbb{R}} \left| P \left(\max_{j \in \mathcal{E}_n \cup \mathcal{I}_n^*} \sqrt{n} f_{nj}(\beta_0) \leq t \right) - P \left(\max_{j \in \mathcal{E}_n \cup \mathcal{I}_n^*} U_{0j} \leq t \right) \right| \rightarrow 0$. The rest of the arguments are the same as in Theorem 2. It follows that for $c_{1-\alpha}$ the $1-\alpha$ quantile of $\mathcal{J} = \arg \min_{h \in \Sigma} \{h' W_0 + \frac{1}{2} h' H_0 h\}$, $P(\sqrt{n}(\hat{\beta}_n - \beta_0) > c_{1-\alpha}) \rightarrow \alpha$.

Similarly, to show finite dimensional convergence in probability of $\infty 1(h \notin \Sigma_n^*)$ to $\infty 1(h \notin \Sigma)$,

$$\begin{aligned}
&P(h_1 \in \Sigma_n^*, \dots, h_k \in \Sigma_n^* | \mathcal{X}_n) \\
&= P \left(\bigcap_{i=1}^k \left\{ \frac{f_{nj}(\bar{\beta}_n)}{\alpha_n} + \bar{F}'_{nj} h_i + \sqrt{n} (f_{nj}^*(\bar{\beta}_n) - f_{nj}(\bar{\beta}_n)) + \bar{F}'_{nj} \left(\frac{\beta_0 - \bar{\beta}_n}{\alpha_n} \right) = 0 \text{ for } j \in \mathcal{E}_n, \right. \right. \\
&\quad \left. \left. \frac{f_{nj}(\bar{\beta}_n)}{\alpha_n} + \bar{F}'_{nj} h_i + \sqrt{n} (f_{nj}^*(\bar{\beta}_n) - f_{nj}(\bar{\beta}_n)) + \bar{F}'_{nj} \left(\frac{\beta_0 - \bar{\beta}_n}{\alpha_n} \right) \leq 0 \text{ for } j \in \mathcal{I}_n \right\} \middle| \mathcal{X}_n \right) \\
&= P \left(\left\{ \max_{j \in \mathcal{E}_n \cup \mathcal{I}_n} \left(\sqrt{n} (f_{nj}^*(\bar{\beta}_n) - f_{nj}(\bar{\beta}_n)) + \max_{1 \leq i \leq k} \bar{F}'_{nj} h_i + \frac{f_{nj}(\bar{\beta}_n)}{\alpha_n} + \bar{F}'_{nj} \left(\frac{\beta_0 - \bar{\beta}_n}{\alpha_n} \right) \right) \leq 0 \right\} \right. \\
&\quad \left. \cap \left\{ \max_{j \in \mathcal{E}_n} \left(-\sqrt{n} (f_{nj}^*(\bar{\beta}_n) - f_{nj}(\bar{\beta}_n)) - \max_{1 \leq i \leq k} \bar{F}'_{nj} h_i - \frac{f_{nj}(\bar{\beta}_n)}{\alpha_n} - \bar{F}'_{nj} \left(\frac{\beta_0 - \bar{\beta}_n}{\alpha_n} \right) \right) \leq 0 \right\} \middle| \mathcal{X}_n \right) \\
&P(h_1 \in \Sigma_n^*, \dots, h_k \in \Sigma_n^* | \mathcal{X}_n) - P(h_1 \in \Sigma, \dots, h_k \in \Sigma)
\end{aligned}$$

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

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

References

- AMEMIYA, T. (1985): *Advanced Econometrics*, Harvard University Press. [7](#), [12](#)
- ANDREWS, D. W. (1999): “Estimation when a parameter is on a boundary,” *Econometrica*, 67, 1341–1383. [2](#), [4](#)
- (2000): “Inconsistency of the bootstrap when a parameter is on the boundary of the parameter space,” *Econometrica*, 68, 399–405. [2](#), [4](#), [14](#)
- (2002): “Generalized method of moments estimation when a parameter is on a boundary,” *Journal of Business & Economic Statistics*, 20, 530–544. [2](#)
- BECK, A. (2017): *First-order methods in optimization*, vol. 25, SIAM. [4](#)
- 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. [15](#)
- 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*. [33](#)

- 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. [33](#)
- 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. [17](#)
- 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](#)
- GEYER, C. J. (1994): “On the asymptotics of constrained M-estimation,” *The Annals of Statistics*, 1993–2010. [2](#), [5](#), [7](#), [8](#), [23](#), [24](#), [25](#)
- HONG, H. AND J. LI (2020): “The numerical bootstrap,” *The Annals of Statistics*, 48, 397–412. [2](#), [30](#)
- 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](#)
- KNIGHT, K. (1999): “Epi-convergence in distribution and stochastic equi-semicontinuity,” *Unpublished manuscript*, 37. [24](#), [25](#), [28](#)

- (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. A. SAUNDERS (2014): “Proximal Newton-type methods for minimizing composite functions,” *SIAM Journal on Optimization*, 24, 1420–1443. [4](#)
- 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](#)
- NEWBY, 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. [7](#), [12](#)
- NOCEDAL, J. AND S. WRIGHT (2006): *Numerical optimization*, Springer Science & Business Media. [26](#)
- POLLARD, D. (1984): “Convergence of stochastic processes,” . [26](#)
- (1991): “Asymptotics for least absolute deviation regression estimators,” *Econometric Theory*, 7, 186–199. [27](#)
- ROCKAFELLAR, R. T., R. J.-B. WETS, AND M. WETS (1998): *Variational analysis*, vol. 317, Springer. [7](#)
- RUST, J. (1987): “Optimal replacement of GMC bus engines: An empirical model of Harold Zurcher,” *Econometrica: Journal of the Econometric Society*, 999–1033. [1](#), [3](#), [10](#), [20](#), [23](#)

- SHAPIRO, A. (1988): “Sensitivity analysis of nonlinear programs and differentiability properties of metric projections,” *SIAM Journal on Control and Optimization*, 26, 628–645. 2, 28
- (1989): “Asymptotic properties of statistical estimators in stochastic programming,” *The Annals of Statistics*, 841–858. 2
- (1990): “On differential stability in stochastic programming,” *Mathematical Programming*, 47, 107–116. 2, 26
- (1991): “Asymptotic analysis of stochastic programs,” *Annals of Operations Research*, 30, 169–186. 2
- (1993): “Asymptotic behavior of optimal solutions in stochastic programming,” *Mathematics of Operations Research*, 18, 829–845. 2
- (2000): “Statistical inference of stochastic optimization problems,” in *Probabilistic Constrained Optimization*, Springer, 282–307. 2
- SU, C.-L. AND K. L. JUDD (2012): “Constrained optimization approaches to estimation of structural models,” *Econometrica*, 80, 2213–2230. 1, 3, 20, 21
- VAN DER VAART, A. AND J. A. WELLNER (1996): *Weak Convergence and Empirical Processes*, Springer. 5, 23
- WACHSMUTH, G. (2013): “On LICQ and the uniqueness of Lagrange multipliers,” *Operations Research Letters*, 41, 78–80. 8, 32