Nonparametric inference for a class of functionals in the random coefficients logit model

Ahnaf Rafi[*]

January 16, 2024

Updated frequently: click here for the most recent version

Abstract

The random coefficients logit model is widely used in choice analysis, empirical industrial organization, and transport economics among other fields. Much recent work has gone into relaxing distributional assumptions made about the random coefficients (RCs). Many objects of interest in this model such as welfare measures, choice probabilities and their derivatives, can be represented as functionals of the distribution of RCs, specifically averages. This paper provides a nonparametric estimator of the RC distribution under which implied plug-in estimators of such averages are asymptotically normal. This is the first formal limiting distribution result for a nonparametric plug-in estimator of such functionals in the RC logit model. For the particular functionals considered here, this asymptotic normality occurs at the parametric $n^{-1/2}$ rate. A consistent estimator of the variance of this limiting distribution is also provided. Together, these results make consistent tests of hypotheses and valid confidence intervals possible in the RC logit model when the distribution of RCs is estimated nonparametrically.

[*] Department of Economics, Northwestern University. Email: ahnafrafi@u.northwestern.edu. I am indebted to Joel Horowitz, Ivan Canay, Eric Auerbach and Federico Bugni and would like to thank them for their invaluable advice, mentorship and support. I have also benefited immensely from discussions with Deborah Kim and Gaston Illanes.

1 Introduction

The random coefficients logit model, also called the mixed logit model, is widely used in choice analysis, empirical industrial organization, and transport economics. In this model, consumers choose from finitely many alternatives by maximizing utility functions that are linear in observed covariates. Coefficients on a subset of these covariates are assumed to be *random*, allowing marginal utilities to vary among observably similar individuals. Thus, random coefficients (henceforth RC or RCs) represent unobserved heterogeneity in tastes associated with observed covariates. In applied work, researchers are often interested in features (or *functionals*) of the unknown distribution of RCs. These functionals can be important inputs in answering policy questions, for example the effect on average consumer welfare of a proposed intervention. Most applications of the RC logit model use parametric specifications of the RC distribution to estimate such functionals. This paper shows how to carry out inference on functionals of the RC distribution when this distribution is estimated nonparametrically. The main result here is asymptotic normality of estimates of these functionals. For the class of functionals considered here, asymptotic normality occurs at the parametric $n^{-1/2}$ rate. This asymptotic normality result can be used for hypothesis testing or for deriving confidence intervals while retaining flexibility of the distribution of RCs.

In applied work, researchers are often interested in objects that that are averages against the distribution of RCs or smooth functions of such averages. Examples that are exact averages include choice probabilities conditional on covariates, (mean) willingness to pay for a product feature or welfare measures such as consumer surplus and compensating variation. The price elasticity is an example of a smooth function (ratio) of averages. Researchers may additionally be interested in averaging over covariates — thereby integrating out both observed and unobserved heterogeneity. This motivates the focus in this paper on *averaging functionals* which are quantities that can be represented as averages over the distribution of RCs and possibly also the distribution of observed covariates.

Most applications of the RC logit model are parametric, and normally or log-normally distributed RCs are particularly popular. Parametric families impose *a priori* implicit restrictions on features of interest. For example, for normally distributed RCs (and a number of other parametric families), Daly et al. (2012) show that mean willingness to pay (Example 2.2 here) is undefined. Miravete et al. (2022) show that Gaussian RCs limits the set of possible elasticities (and curvature values) for the demand function. This in turn impacts conclusions drawn about the pass-through of taxes on the consumer. Nonparametric estimation of the distribution of RCs provides greater flexibility in terms of implied features of the model. Using nonparametric techniques for estimating the RC distribution also eliminates a source of misspecification. These observations motivate the focus here on nonparametric estimators for the distribution of RCs.

Relaxing parametric assumptions on RC distributions has been a topic of much interest in the econometrics literature, but existing results are limited to identification, consistency or pointwise asymptotic normality. Thus, they do not allow researchers to conduct inference on objects of interest. For instance, in a binary choice model without the logit assumption, Gautier and Kitamura (2013) construct an estimator of the density of RCs that is (consistent and) pointwise asymptotically normal. They do not provide provide limit distribution results for functionals of this density.[1] For multinomial choice with a logit assumption, Train (2008), Train (2016), Fox et al. (2016) and Heiss et al. (2022) consider a variety of possible nonparametric estimators for the distribution of RCs. The last two papers also give consistency and convergence rate results. None of these four papers derive asymptotic distributions, and hence, inference on functionals of interest is not possible using their estimators. A key problem the proposed estimators in these four papers face is an infinite-dimensional counterpart of the "parameters at the boundary" problem. Their choice of parameter space is defined by a set of inequality constraints (probabilities are non-negative) and an equality constraint (total mass equals one). Inequality constraints present a problem for asymptotic normality since it is not known *a priori* whether they bind at the the true parameter value. Equality constraints are always known to bind, and hence not problematic. Some details and relevant references are provided in the section on related literature.

The main contribution of this paper is an asymptotic normality result for estimates of averaging functionals using a nonparametric estimator of the distribution of RCs. This is provided in the context of discrete choice data at the level of the decision maker — that is, consumer level choice data is assumed. The RC logit model used here allows RCs on some covariates and non-random coefficients on remaining covariates. Such random/non-random coefficient breakdowns are commonplace in applied work. Covariates are assumed to be exogenous so that the nonparametric point identification results of Fox et al. (2012) apply. The distribution of RCs is restricted to be continuous so that it so that it admits a density. This density is not restricted to a finite-dimensional parametric family, but is instead estimated using nonparametric sieve maximum likelihood. The sieves considered here are linear approximation spaces for the square root of the density. The square root transformation is important in establishing asymptotic normality. It avoids the infinite-dimensional counterpart of a "parameter at the boundary problem" since the implied density is automatically non-negative, and the only remaining constraint is the equality constraint that a density should integrate to one. In addition, the square-root transformation facilitates appropriate differentiability properties for both the likelihood and the functionals of interest. The plug-in estimator of averaging functionals based on this sieve estimator is shown to be asymptotically normal at the parametric $n^{-1/2}$ rate under regularity conditions. A consistent data-based estimator of the limit variance is provided. Combining the asymptotic normality and consistent variance estimation results gives tractable and consistent hypothesis tests and confidence intervals for functionals of interest. In the setting of consumer level nonparametric RC logit models, this is the first result on asymptotically valid inference for functionals.

The nonparametric RC density estimator in this paper is subject to the curse of dimensionality; allowing for more RCs slows down its rate of convergence. I show that the curse can be mitigated by assuming that individual RCs are independent. In this case, the estimator enjoys the one-

[1] Pointwise asymptotic normality of an estimated function, such as a density or a conditional mean, does not imply asymptotic normality of implied plug-in estimators of functionals. See Bickel and Ritov (2003) for a counterexample and related discussion.

dimensional rate of convergence, i.e. the rate with a single RC in the model. The intuition is similar to that for dimension reduction in nonparametric additive models. Independence assumptions of this sort are commonplace in applied work using parametric specifications of the distribution of RCs.

The utility of the method in this paper is illustrated through an empirical application on the value of a statistical life (VSL) in Sierra Leone. This empirical application is inspired by León and Miguel (2017) and uses their dataset with the nonparametric approach of the present paper. The key finding is a higher VSL estimate, leading to higher estimates of the benefit due solely to reduced mortality risk of a large public infrastructure project in Sierra Leone. The 95% confidence intervals around this higher nonparametric VSL estimate (and the implied infrastructure investment benefit) do not contain the original VSL estimate in León and Miguel (2017).

1.1 Related literature

The parametric RC logit model can be traced back to Boyd and Mellman (1980) and Cardell and Dunbar (1980); both papers used it to study the U.S. automobile market with aggregate data, i.e. outcome variables were market shares. A number of subsequent papers use the parametric RC logit model with consumer level data (henceforth micro data); for example Train et al. (1987) studied the residential telephone service market. Textbook treatments of the parametric RC logit model can be found in Train (2009, Chapter 6) and Hensher et al. (2015, Chapter 15). Examples of applications and survey papers include Small et al. (2005), León and Miguel (2017), Hensher and Greene (2003) and Keane and Wasi (2013). In terms of parametric statistical inference, under some regularity conditions (see Lee (1992), Hajivassiliou and Ruud (1994) and Lee (1995)) standard tools as can be found in Newey and McFadden (1994) apply. Horowitz and Nesheim (2021) provides an extension to inference with high-dimensional covariates under sparsity using penalized maximum likelihood and the adaptive-LASSO. The present paper contributes to this literature by providing a method for conducting consistent inference on objects of interest while using a nonparametric estimator of the RC distribution.

There is a recent growing econometrics literature on RC discrete choice models with nonparametric treatment of the distribution of RCs. Papers on nonparametric point identification include Fox et al. (2012) and Allen and Rehbeck (2023) for static discrete choice models and Bunting (2022) for a dynamic model. Previously mentioned papers on nonparametric estimation are: Gautier and Kitamura (2013), Train (2008), Train (2016), Fox et al. (2016) and Heiss et al. (2022). In this literature, limit distribution theory for functionals of the RC distribution remains unexplored and this is the primary contribution of the present paper.

The fixed-grid discrete distribution sieve estimator of Bajari et al. (2007) has been used in several applied papers — examples are Nevo et al. (2016), Blundell et al. (2020) and Illanes and Padi (2021). Fox et al. (2016) provides consistency theory and convergence rates for this class of estimators in nonparametric discrete choice RC models including but not limited to the RC logit. The applied papers mentioned use discrete choice RC models, but without the logit assumption. Of these, Nevo et al. (2016) and Blundell et al. (2020) use bootstrap methods to get standard errors for inference. They do not prove bootstrap consistency, and I have not found any proofs of bootstrap consistency using fixed-grid estimators in the broader literature. I also have not found any results on limit distribution theory using fixed-grid estimators. An issue that fixed-grid estimators have to contend with is the "parameters at the boundary" problem — see Fox et al. (2011)[2] for discussion specific to fixed-grid estimators and Geyer (1994) and Andrews (1999) for treatments of general theory around this phenomenon. In addition to creating a hurdle for asymptotic normality, the true parameter being a boundary point of the parameter space typically renders bootstrap procedures inconsistent — see Andrews (2000) and Fang and Santos (2018). The approach used here does not run into the boundary problem and it may be possible to generalize this approach to similar non-logit settings. Exploration of this possibility is left to future research.

An issue commonplace in practice that is not addressed in this paper is endogeneity. In parametric RC logit models, the workhorse approach under endogeneity using aggregate data is that of Berry (1994) and Berry et al. (1995) (henceforth BLP95). There are also extensions of BLP95 to micro data, for example Berry et al. (2004), Goolsbee and Petrin (2004), Grieco et al. (2023). A few recent papers have explored nonparametric generalizations of BLP95 in the aggregate data context: Compiani (2022), Lu et al. (2023) and Wang (2022). Of these, the first two provide inference results. Nonparametric inference on functionals in micro data demand models in the presence of endogeneity is unexplored, to my knowledge. It may be possible to combine the approach in the present paper with approaches from parametric micro data demand models with endogeneity. Examples of such approaches are two-step likelihood methods (as in Goolsbee and Petrin (2004)) or control functions (as in Petrin and Train (2010)). The possibility of combining results here with these alternate approaches involving instruments is left to future research.

1.2 Structure of the rest of the paper

The remainder of the paper is organized as follows. Section 2 presents the RC logit model in Section 2.1, the functionals of interest in Section 2.2 and discussion of the choice of square root transformation of the density in Section 2.3. Section 3 describes the plug-in estimation procedure. Section 4 is divided into four subsections. Section 4.1 treats consistency and convergence rates for the sieve estimator for model primitives. The main asymptotic normality result is provided in Section 4.2. Results on mitigating the curse of dimensionality via independence assumptions on the RCs are in Section 4.3. Throughout all of these, point identification of model primitives is maintained as a high-level assumption. Section 4.4 presents the lower level sufficient conditions of Fox et al. (2012) for point identification. Section 5 illustrates the finite sample performance of the estimator in Monte Carlo simulations. Section 6 presents results from the empirical application on the value of a statistical life in Sierra Leone. Section 7 concludes.

[2] Specifically, see the paragraphs including and following footnotes 7 and 8 in Section 6 of Fox et al. (2011)

2 Setup

2.1 The random coefficients logit model

Consider a population of decision makers (DMs) each making a single choice from a finite set of mutually exclusive and exhaustive alternatives: $\mathcal{Y} = \{0, \dots, J\}$ with $J \in \mathbb{N}$. The observed choice, Y, is determined through (indirect) utility maximization:

$$Y = \operatorname*{argmax}_{y \in \mathcal{Y}} \left\{ W_y' \alpha_0 + X_y' \beta + \varepsilon_y \right\}. \tag{1}$$

For each $y \in \mathcal{Y}$, W_y and X_y are observed covariates of respective dimensions d_W and d_X. They can contain variables specific to the alternative, the DM or both. Variables in W_y are associated with *non-random coefficients* (NRCs) α_0 to be estimated. Variables in X_y are associated with unobserved *random coefficients* (RCs) β whose distribution F_0 is to be estimated. Such non-random/random decompositions of the coefficients are commonplace in practice. Finally, ε_y is an unobserved idiosyncratic additive error.[3] Stack the covariates so that $W' = (W_0', \dots, W_J')$ and $X' = (X_0', \dots, X_J')$. The joint distribution of (W, X), denoted G_0, is also unknown. The researcher has a sample of n i.i.d. observations, $\{Y_i, W_i, X_i\}_{i=1}^n$ from the distribution of (Y, W, X).

The following restrictions produce a random coefficients logit model from (1):

(i) independently across $y \in \mathcal{Y}$, ε_y has a standard Gumbel distribution,

(ii) $(\varepsilon_0, \dots, \varepsilon_J) \perp\!\!\!\perp (W, X, \beta)$,

(iii) $\beta \perp\!\!\!\perp (W, X)$.

These restrictions and (1) imply that conditional choice probability for an alternative given only observable covariates are:

$$\begin{gathered} P_0(y, w, x) := \Pr(Y = y | W = w, X = x) = \int \kappa\left(y, w, x; \alpha_0, b\right) F_0(\mathrm{d}b), \\ \text{where} \quad \kappa(y, w, x; \alpha, b) = \frac{\exp\left(w_y' \alpha + x_y' b\right)}{\sum_{j=0}^J \exp\left(w_j' \alpha + x_j' b\right)}. \end{gathered} \tag{2}$$

The model unknowns are: the true NRCs α_0, the distribution of RCs F_0 and the distribution of observed covariates G_0.

2.2 The averaging functional: definition and examples

The objects of interest in this paper are *averaging functionals*; these are defined as having the form

$$\tau_0 = \int \left\{ \int t\left(w, x; \alpha_0, b\right) F_0(\mathrm{d}b) \right\} G(\mathrm{d}w, \mathrm{d}x), \tag{3}$$

[3] The RCs β and the utility shock ε_y are unobserved by the econometrician. The individual DM knows the values of utilities in (1).

for a fixed, known and possibly vector-valued function $t(\cdot)$. Many quantities of economic interest have the representation (3) or are smooth transformations thereof.

Example 2.1 (Welfare changes)**.** An intervention, for example a tax or an innovation, will change prices and/or the number of available alternatives. The researcher is interested in the resulting compensating variation. Denote values before and after the intervention by $(\cdot)_{\text{pre}}$ and $(\cdot)_{\text{post}}$ respectively. Small and Rosen (1981) show that under the logit assumption, individual compensating variation conditional on (W, X, β) is given by a "log-sum-exp" formula,

$$\begin{aligned} t_{\text{cv}}\left(W, X; \alpha_0, \beta\right) = \frac{1}{\pi_{\text{price}}\left(\alpha_0, \beta\right)} \; & \left\{ \log\left(\textstyle\sum_{y=0}^{J_{\text{post}}} \exp\left(W_{\text{post},y}' \alpha_0 + X_{\text{post},y}' \beta\right)\right) \right. \\ & \left. - \log\left(\textstyle\sum_{y=0}^{J_{\text{pre}}} \exp\left(W_{\text{pre},y}' \alpha_0 + X_{\text{pre},y}' \beta\right)\right) \right\}, \end{aligned}$$

where $\pi_{\text{price}}(\cdot)$ returns the coefficient in (α_0, β) associated with price. Taking the expectation of $t_{\text{CV}}(W, X; \alpha, \beta)$ with respect to (W, X, β) gives average compensating variation,

$$\tau_{\text{cv},0} = \int \left\{ \int t_{\text{cv}}(w, x; \alpha, b) F_0(\mathrm{d}b) \right\} G(\mathrm{d}w, \mathrm{d}x).$$

Remark 2.1 (On the pre/post subscripts in Example 2.1)**.** When data from both before and after an intervention are available, the subscripts $(\cdot)_{\text{pre}}$ and $(\cdot)_{\text{post}}$ can be taken to mean sample splits. If the intervention has not yet happened, the $(\cdot)_{\text{post}}$ values can be predicted values determined by an economic model.

Example 2.2 (Mean marginal willingness to pay)**.** Let $l \in \{1, \ldots, d_W + d_X\}$ be the index of a non-price covariate and $\pi_l(\alpha, b)$ denote the corresponding coefficient. Let $\pi_{\text{price}}(\alpha, b)$ denote the coefficient on price. Since coefficients measure marginal indirect utilities, the coefficient ratio

$$t_{\text{wtp}}\left(W, X; \alpha_0, \beta\right) = t_{\text{WTP}}\left(\alpha_0, \beta\right) = \frac{\pi_l\left(\alpha_0, \beta\right)}{\pi_{\text{price}}\left(\alpha_0, \beta\right)}$$

is a measure of *marginal willingness to pay* for feature l — see for instance Train and Weeks (2005). Integrating out (W, X, β) amounts to only integrating out β since there are no covariates on the right hand side of the above display. The associated averaging functional is *mean marginal willingness to pay* for feature l

$$\tau_{\text{wtp},0} = \int \frac{\pi_l\left(\alpha_0, b\right)}{\pi_{\text{price}}\left(\alpha_0, b\right)} F_0(\mathrm{d}b).$$

Example 2.3 (Choice probabilities conditional on covariates, their derivatives and elasticities)**.** Let (w_*, x_*) be a fixed point of evaluation in the support of the stacked covariates (W, X) and let $y \in \mathcal{Y}$ be any alternative. Individual choice probabilities conditional on $(W, X) = (w_*, x_*)$ are

$$\tau_{\text{ccp},0} = P_0\left(y, w_*, x_*\right) = \int \kappa\left(y, w_*, x_*; \alpha_0, b\right) F_0(\mathrm{d}b),$$

with $\kappa(\cdot)$ as in (2). Now, stack alternative specific covariates as $z_j' = (w_j', x_j')$. The derivative of the choice probability for alternative y at (w_*, x_*) with respect to a feature $z_{j,l}$ for alternative j is the averaging functional

$$\tau_{\text{dccp},0} = \frac{\partial}{\partial z_{j,l}} P_0\left(y, w_*, x_*\right) = \int \frac{\partial}{\partial z_{j,l}} \kappa\left(y, w_*, x_*; \alpha_0, b\right) F_0(\mathrm{d}b).$$

Derivatives can be with respect to features of the same alternative ($j = y$) or a distinct one ($j \neq y$). The logit assumption provides convenient closed forms for the integrand $\frac{\partial}{\partial z_{j,l}} \kappa$. An elasticity is accommodated as a smooth transformation (ratio) of τ_{ccp} and τ_{dccp}:

$$\tau_{\text{elasticity},0} = z_{j,l,*} \cdot \frac{\frac{\partial}{\partial z_{j,l}} P_0\left(y, w_*, x_*\right)}{P_0\left(y, w_*, x_*\right)} = z_{j,l,*} \cdot \frac{\tau_{\text{dccp},0}}{\tau_{\text{ccp},0}}.$$

Example 2.4 (Average marginal effects)**.** For a fixed alternative $y \in \mathcal{Y}$, and a feature $z_{j,l}$ of (possibly the same) alternative $j \in \mathcal{Y}$, the *average marginal effect* is

$$\tau_{\text{ame},0} = \int \int \frac{\partial}{\partial z_{j,l}} \kappa\left(y, w, x; \alpha_0, b\right) F_0(\mathrm{d}b)\, G(\mathrm{d}w, \mathrm{d}x).$$

$\tau_{\text{ame},0}$ integrates the derivative of $\kappa(\cdot)$ over covariates and RCs. In contrast, $\tau_{\text{dccp},0}$ in Example 2.3 fixes values of the covariates and integrates out only the RCs.

Remark 2.2. The averaging functional (3) is allowed to integrate over two sources of heterogeneity, the RCs β and covariates (W, X), and has an additional unknown: the NRCS α_0. As Examples 2.2 and 2.3 show, one or two of these can be dropped when the integrand $t(w, x, \alpha, b)$ is constant in some of its arguments. Thus averaging functionals cover the NRCs α_0 and moments of β through appropriate choice of $t(\cdot)$.

2.3 Some precursors to estimation

In this paper, the approach to estimation of (3) will be to first estimate α_0 and F_0 from the data and replace them whereever they appear with respective estimates. Similarly, G_0 will be replaced by the empirical distribution of covariates, so that expectations against G_0 become sample averages.

The approach to estimation of F_0 will be nonparametric. This means that the set of restrictions imposed on this distribution will not imply that it is restricted to a finite-dimensional parametric family. Estimation of F_0 requires specifying a support for the RCs β. Assumption 2.1 below specifies a compactness restrictions on this support required by the theory in Section 4.

Assumption 2.1. The set $\mathcal{B} \subseteq \mathbb{R}^{d_X}$ is a Cartesian product of d_X compact intervals, each with non-empty interior. That is, there are known $-\infty < \underline{\beta}_l < \overline{\beta}_l < \infty$ ($l \in \{1, \ldots, d_X\}$) such that $\mathcal{B} = \prod_{l=1}^{d_X} \left[\underline{\beta}_l, \overline{\beta}_l\right]$. The true support of β, support$\,(F_0)$, is contained in $\mathcal{B}$.

The specified set $\mathcal{B}$ must contain the true support, but exact knowledge of this support is not required. Compactness of $\mathcal{B}$ is necessitated by the lower level sufficient conditions for the main identification assumption (Assumption 2.2 below), as well as the asymptotic results in Section 4. See Remark A.1 in the appendix for a discussion. Henceforth, integration with respect to the RCs will always be over $\mathcal{B}$ so that $\int g(b)\mathrm{d}b \equiv \int_{\mathcal{B}} g(b)\mathrm{d}b$.

The nonparametric approach taken in this paper will be to approximate a transformation of F_0 and use the data to estimate the best approximation. This is known as *the method of sieves.* For simplicity, suppose finite dimensional linear approximations are used, so that

$$\text{a transformtion of } F_0 \approx \gamma_1\psi_1 + \cdots + \gamma_K\psi_K,$$

where $\approx$ indicates that this is an approximation and not an exact expression. The ψ_k above are functions that are required to approximate the transformation above to within any level of precision if their number, K, is sufficiently large. For example, one can choose to use the c.d.f. associated with F_0 as the transformation and approximate this c.d.f using a finite linear combination of discrete c.d.f's.[4] Another option is to use the density as the transformation of F_0 (assuming it has a Lebesgue density) and use a finite linear combination of trigonometric functions or polynomials as the approximating functions.[5] For both these choices, a key issue will be a problem known as "parameters at the boundary". The set of c.d.f's and densities are defined by (i) an equality constraint: the mass assigned to $\mathcal{B}$ must equal one and (ii) a set of inequality constraints: at any point in $\mathcal{B}$ both of these objects must take non-negative values. Inequality constraints pose a problem for asymptotic distribution theory since it is not known *a priori* whether they bind at the true value. In the finite-dimensional case, this is reflected in the limit distribution of estimators: a Gaussian limiting distribution occurs when inequality constraints do not bind at the true value, whereas a non-Gaussian limit is obtained when inequality constraints do bind. Equality constraints are always known to bind and hence, do not pose a problem — in the finite-dimensional case, a Gaussian limit distribution is achieved when only equality constraints are present. See for example Geyer (1994) and Andrews (1999) for details. In infinite dimensions, this problem does not disappear.

For the remainder of this paper, F_0 will be assumed to have a Lebesgue density. There are transformations of the density under which non-negativity, the source of the problem, is not an issue. One can take the square-root of the density or the logarithm and estimate approximations to these.[6] Then, only the equality constraint remains, since the implied density must still integrate to one. In the present paper, the transformation used will be the square-root density. Since the square root is not uniquely defined (one can use both $-\sqrt{a}$ and $\sqrt{a}$ for $a \geq 0$), this transformation creates a uniqueness problem. The uniqueness problem however, is not difficult to handle.

[4] Linear combinations of discrete c.d.f's can approximate any c.d.f: this fact is implied by Theorem 15.10 of Aliprantis and Border (2006) for example.

[5] Polynomials and trigonometric functions are universal approximators for a number of function classes. As a result of the Stone-Weierstrass Theorem, one of these function classes is $\mathscr{L}_1$, the set of all Lebesgue integrable functions. Since densities integrate to 1, they are automatically members of $\mathscr{L}_1$.

[6] These are not new ideas in the nonparametric estimation literature. See Gallant and Nychka (1987), Chen et al. (2006) and Bierens (2014) for examples of the square-root as the chosen transformation. For the logarithm of the density, there are a number of papers on "log-spline" estimation — see for example Stone (1990) and references therein.

For a function h such that $\int h(b)^2 \mathrm{d}b = 1$, denote

$$P(y, w, x; \alpha, h) = \int \kappa(y, w, x; \alpha, b) h(b)^2 \mathrm{d}b,$$
$$\text{where } \kappa(y, w, x; \alpha, b) = \frac{\exp\left(w_y'\alpha + x_y'b\right)}{\sum_{j=0}^{J} \exp\left(w_j'\alpha + x_j'b\right)}. \tag{4}$$

The condition $\int h(b)^2 \mathrm{d}b = 1$ means that $h(\cdot)^2$ is a density function since $h(\cdot)^2 \geq 0$. Let h_0 be the true value of h so that h_0^2 is the density of F_0. To ensure h_0 is unique, it will be restricted to be non-negative: $h_0(\cdot) \geq 0$. This restriction is used only to define the true value and will not be imposed during estimation. The true conditional choice probabilities in (2) then satisfy $P_0(y, w, x) \equiv P(y, w, x; \alpha_0, h_0)$. Thus, (4) expresses the model in terms of the key unknowns, α, h. The object of interest, the averaging functional in (3) can also be similarly represented:

$$\tau(\alpha, h, G) = \int \left\{ \int t(w, x; \alpha, b) h(b)^2 \, \mathrm{d}b \right\} G(\mathrm{d}w, \mathrm{d}x), \tag{5}$$

with true value $\tau_0 = \tau(\alpha_0, h_0, G_0)$. Point identification of the pair (α_0, h_0) will be assumed throughout in Assumption 2.2 below.

Assumption 2.2. Let $\mathcal{A} \times \mathcal{H}$ be the set of candidates for (α_0, h_0). The pair (α_0, h_0) is point identified relative to $\mathcal{A} \times \mathcal{H}$: for any $(\alpha, h) \in \mathcal{A} \times \mathcal{H}$,

if $P(y, w, x; \alpha, h) = P_0(y, w, x)$ for every $y \in \mathcal{Y}$ almost surely with respect to G_0,
then $\alpha = \alpha_0$ and $h^2 = h_0^2$ almost everywhere.

Lower level sufficient conditions for Assumption 2.2 exist in the literature, for example those of Fox et al. (2012). The sufficient conditions of Fox et al. (2012) for Assumption 2.2 are presented in Section 4.4. The parameter set in Assumption 2.2 is a product of $\mathcal{A} \subseteq \mathbb{R}^{d_W}$ and a set of functions, $\mathcal{H}$, that produce densities upon squaring. Additional restrictions will be placed on both these sets in Section 4. $\mathcal{H}$ will nonetheless be a nonparametric (i.e. infinite-dimensional) family; it will not be parametrized by a subset of finite-dimensional Euclidean space.

3 The plug-in estimator

A plug-in estimator of τ_0 is formed by substituting estimators $\widehat{\alpha}_n, \widehat{h}_n, \widehat{G}_n$ of α_0, h_0, G_0 into (5). The estimator $\widehat{G}_n$ of the distribution of observed covariates in this paper will be the empirical distribution $\widehat{G}_n(A) := (1/n) \sum_{i=1}^{n} \mathbb{I}\{(W_i, X_i) \in A\}$. Then, given $\widehat{\alpha}_n$ and $\widehat{h}_n$, the plug-in estimator is a sample average

$$\widehat{\tau}_n = \tau\left(\widehat{\alpha}_n, \widehat{h}_n, \widehat{G}_n\right) = \frac{1}{n} \sum_{i=1}^{n} \int t\left(W_i, X_i; \widehat{\alpha}_n, b\right) \widehat{h}_n(b)^2 \, \mathrm{d}b. \tag{6}$$

The main result in this paper can be summarized as follows. When the estimators $\widehat{\alpha}_n$ and $\widehat{h}_n$ are formed using a sieve maximum likelihood procedure to be described, there is a consistent data-based variance estimator $\widehat{V}_{\tau,n}$ for $\widehat{\tau}_n$ such that

$$\frac{\sqrt{n}\,(\widehat{\tau}_n - \tau_0)}{\sqrt{\widehat{V}_{\tau,n}}} \xrightarrow{\mathrm{d}} \mathcal{N}(0,1) \quad \text{as } n \to \infty,$$

under some regularity conditions. The formal statement of this result is Theorem 4.3. The variance estimator $\widehat{V}_{\tau,n}$ can be written as the sample variance of estimated influence functions, see (18). Given estimates $\widehat{\alpha}_n$ and $\widehat{h}_n$, $\widehat{\tau}_n$ can be computed as follows:

1. Compute $\int t\,(W_i, X_i; \widehat{\alpha}_n, b)\,\widehat{h}_n(b)^2\,\mathrm{d}b$ for each observation i using numerical integration.

2. Take the sample average.

The numerical integration method in the first step can be polynomial-based or Monte Carlo.

The estimation procedure for (α_0, h_0) will be sieve maximum likelihood. The sample average log-likelihood function is

$$\ell_n(\alpha, h) := \frac{1}{n}\sum_{i=1}^{n} \log P\,(Y_i, W_i, X_i; \alpha, h) = \frac{1}{n}\sum_{i=1}^{n} \log \int \kappa\,(Y_i, W_i, X_i; \alpha, b)\,h(b)^2 \mathrm{d}b.$$

The model primitives (α_0, h_0) will be estimated by maximizing ℓ_n over a sequence of approximating sets $\mathcal{A} \times \mathcal{H}_n$ (sieves):

$$\left(\widehat{\alpha}_n, \widehat{h}_n\right) = \operatorname*{argmax}_{(\alpha,h)\in\mathcal{A}\times\mathcal{H}_n} \ell_n(\alpha, h). \tag{7}$$

Linear approximations will be used for elements of $\mathcal{H}$:

$$\mathcal{H}_n = \left\{\gamma' \boldsymbol{\psi}_{K_n}(\cdot) : \gamma \in \Gamma_n\right\},$$

where $\boldsymbol{\psi}_{K_n}(\cdot)' = (\psi_{1,K_n}(\cdot), \ldots, \psi_{K_n,K_n}(\cdot))$ is a vector of K_n approximating basis functions defined on $\mathcal{B}$, γ is a vector of coefficients constrained to $\Gamma_n \subseteq \mathbb{R}^{K_n}$. In practice, I propose using splines (piecewise polynomials) for $\boldsymbol{\psi}_{K_n}$, but other approximating bases can also be used — see Section 4 for theoretical restrictions.

The set Γ_n is defined by two constraints. First, elements of $\mathcal{H}_n$ must produce densities upon squaring. Second, $\mathcal{H}_n$ must satisfy the same smoothness restrictions as those imposed on $\mathcal{H}$ in Section 4. Taken together, $\gamma \in \Gamma_n$ if and only if

$$\gamma' \left[\int \boldsymbol{\psi}_{K_n}(b)\boldsymbol{\psi}_{K_n}(b)' \mathrm{d}b\right] \gamma = 1, \tag{8}$$

$$\gamma' \left[\sum_{0\le|\mathbf{s}|\le s} \int \left[D^{\mathbf{s}}\boldsymbol{\psi}_{K_n}(b)\right]\left[D^{\mathbf{s}}\boldsymbol{\psi}_{K_n}(b)\right]' \mathrm{d}b\right] \gamma \le C^2. \tag{9}$$

Equation (8) expresses "squaring produces a density" as a quadratic equality constraint. The quadratic form in (9) is a measure of smoothness of the linear combination $\gamma'\boldsymbol{\psi}_{K_n}$. The vector $\mathbf{s}' = (s_1, \ldots, s_{d_X})$ of non-negative integers specifies coordinate-wise partial derivative orders. $D^{\mathbf{s}} = \partial^{|\mathbf{s}|} / \prod_{l=1}^{d_X} \partial b_l^{s_l}$ is the corresponding partial derivative with total order $|\mathbf{s}| := \sum_{l=1}^{d_X} s_l$. Section 4 provides additional restrictions on the order, s, of partial derivatives required and the constant C in the quadratic inequality constraint (9). The program (7) is equivalent to first solving the constrained optimization problem

$$(\widehat{\alpha}_n, \widehat{\gamma}_n) = \underset{(\alpha,\gamma)\in\mathcal{A}\times\Gamma_n}{\operatorname{argmax}} \frac{1}{n}\sum_{i=1}^{n} \log \int \kappa\left(Y_i, W_i, X_i; \alpha, b\right) \left(\gamma'\boldsymbol{\psi}_{K_n}(b)\right)^2 \mathrm{d}b, \tag{10}$$

and then setting $\widehat{h}_n = \widehat{\gamma}_n'\boldsymbol{\psi}_{K_n}$. With $\widehat{\alpha}_n$ and $\widehat{h}_n$ so defined, the plug-in estimator $\widehat{\tau}_n$ in (6) can be implemented. The full set of steps to get $\widehat{\tau}_n$ are described in Algorithm 1 below.

Algorithm 1: Plug-in estimation routine for $\widehat{\tau}_n$ in (6)

```
Input: sample $\{Y_i, W_i, X_i\}_{i=1}^{n}$, support $\mathcal{B}$ for densities, basis functions $\boldsymbol{\psi}_{K_n}$ (and length of basis $K_n$), smoothness level $s > d_X/2$ and constraint upper bound $C > 1$ for (9), a numerical integration rule
Result: $\widehat{\tau}_n$ in (6).
Solve (10) to get $(\widehat{\alpha}_n, \widehat{\gamma}_n)$;
Set $\widehat{h}_n = \widehat{\gamma}_n'\boldsymbol{\psi}_{K_n}$;
if $\tau(\cdot)$ in (5) requires sample averaging then
    Compute $\int t(W_i, X_i; \widehat{\alpha}_n, b)\, \widehat{h}_n(b)^2\, \mathrm{d}b$ for each observation $i$ using numerical integration;
    Set $\widehat{\tau}_n$ equal to the sample average of the above integrals;
else
    Compute $\widehat{\tau}_n = \int t(\widehat{\alpha}_n, b)\, \widehat{h}_n(b)^2\, \mathrm{d}b$ using numerical integration;
```

4 Theoretical results

This section presents theoretical results. Consistency and convergence rates for the sieve MLE for model primitives are in Section 4.1. These are forerunners to the main asymptotic normality result in Section 4.2 as can be seen from Algorithm 1. Results on mitigating the curse of dimensionality are in Section 4.3. Throughout, Assumptions 2.1 and 2.2 are maintained. Sufficient conditions for Assumption 2.2 from Fox et al. (2012) are presented in Section 4.4.

4.1 Consistency and convergence rates for model primitives

The estimator $(\widehat{\alpha}_n, \widehat{h}_n)$ is a maximum likelihood estimator. Lemma 4.1 below shows that identification in Assumption 2.2 is sufficient to ensure that the target, (α_0, h_0), must be the unique maximizer of the expected log-likelihood:

$$\ell_*(\alpha, h) = \mathbb{E}\left[\log P(Y, W, X; \alpha, h)\right] = \int \left[\sum_{y=0}^{J} P_0\left(y, w, x\right) \log P(y, w, x; \alpha, h)\right] G_0(\mathrm{d}w, \mathrm{d}x).$$

Lemma 4.1. *Suppose Assumptions 2.1 and 2.2 hold. Then*

$$\ell_*\left(\alpha_0, h_0\right) \geq \ell_*(\alpha, h) \textit{ with equality only if } \alpha = \alpha_0 \textit{ and } h^2 = h_0^2 \textit{ almost everywhere.}$$

Lemma 4.1 is well known and follows from a conditional variant of the information inequality. A proof is provided for completeness. For estimation, the researcher must specify a domain $\mathcal{B}$ for the basis functions $\boldsymbol{\psi}_{K_n}$. The following two assumptions restrict the parameter set $\mathcal{A}$ for α_0 and the set $\mathcal{B}$ used as the domain for functions in $\mathcal{H}$.

Assumption 4.1. $\mathcal{A} \subseteq \mathbb{R}^{d_W}$ is compact and has a non-empty interior. Furthermore, $\alpha_0 \in \mathrm{int}(\mathcal{A})$.

Assumption 4.2. The distribution, G_0, satisfies the following exponential moment condition

$$\int \exp\left(8\mathcal{M}_{\mathcal{A}} \sum_{j=0}^{J} \left\|w_j\right\|_2 + 8\mathcal{M}_{\mathcal{B}} \sum_{j=0}^{J} \left\|x_j\right\|_2\right) G_0(\mathrm{d}w, \mathrm{d}x) < \infty,$$

where $\mathcal{M}_{\mathcal{A}} = \sup_{\alpha \in \mathcal{A}} \|\alpha\|_2$ and $\mathcal{M}_{\mathcal{A}} = \sup_{b \in \mathcal{B}} \|b\|_2$.

The exponential moment condition (4.2) in Assumption 4.2 is used to ensure that choice probabilities $P(\cdot)$ in (4) do not get too close to zero. In particular, this condition ensures that the logarithm of choice probabilities used to define the log-likelihood has finite moments. This helps with proofs of consistency and convergence rates.

As before, for a multi-index, $\mathbf{s} = (s_1, \ldots, s_{d_X}) \in \mathbb{Z}_+^{d_X}$, let $|\mathbf{s}| = \sum_{l=1}^{d_X} s_l$ and $D^{\mathbf{s}} = \partial^{|\mathbf{s}|} / \prod_{l=1}^{d_X} \partial b_l^{s_l}$. I use the convention $D^{\mathbf{0}} h \equiv h$. For $s \in \mathbb{N}$, and s-times differentiable h, the $(s, 2)$-Sobolev norm is

$$\|h\|_{s,2} = \left[\sum_{0 \leq |\mathbf{s}| \leq s} \int \left|D^{|\mathbf{s}|} h(b)\right|^2 \mathrm{d}b\right]^{1/2}.$$

The $(s, 2)$-Sobolev space is $\mathscr{W}_{s,2}(\mathcal{B}) = \{h : \|h\|_{s,2} < \infty\}$. For $0 < C < \infty$, the $(s, 2)$-Sobolev closed ball of radius C is $\mathscr{W}_{s,2}(\mathcal{B}, C) = \{h : \|h\|_{s,2} \leq C\}$.

Assumption 4.3. For some $1 < C < \infty$ and $s > d_X/2$, $\mathcal{H}$ is the intersection of $\mathscr{W}_{s,2}(\mathcal{B}, C)$ and the unit sphere in $\mathscr{L}_2$, i.e. $\mathcal{H} = \left\{h \in \mathscr{W}_{s,2}(\mathcal{B}, C) : \int h(b)^2 \mathrm{d}b = 1\right\}$. In addition, $\|h_0\|_{s,2} < C$.

An important consequence of Assumption 4.3 is that $\mathcal{H}$ is a subset of a smoothness class for which there are well-established approximation error rates for common basis functions. An additional useful consequence is that $\mathcal{H}$ is $\mathscr{L}_2$-compact — see Theorem 2 of Freyberger and Masten (2019).

Assumption 4.4. For each $K \in \mathbb{N}$, $\boldsymbol{\psi}_K' = (\psi_{1,K}, \ldots, \psi_{K,K})$ is a vector of K functions on $\mathcal{B}$ which satisfy the following.

(i) The components of $\boldsymbol{\psi}_K$ are orthonormal, i.e. for every $k_1, k_2 \in \{1, \ldots, K\}$,

$$\int \psi_{k_1,K}(b) \cdot \psi_{k_2,K}(b) \mathrm{d}b = \begin{cases} 1 & \text{if } k_1 = k_2, \\ 0 & \text{if } k_1 \neq k_2. \end{cases}$$

(ii) For $s > d_X/2$ and any $h \in \mathscr{W}_{s,2}(\mathcal{B})$, there exists $C_h > 0$ such that

$$\limsup_{K \to \infty} \left(K^{s/d_X} \cdot \inf_{\gamma \in \mathbb{R}^K} \sqrt{\int \left(h(b) - \gamma' \boldsymbol{\psi}_K(b)\right)^2 \, \mathrm{d}b} \right) \leq C_h.$$

Assumption 4.5. Let C and s be as in Assumption 4.3, $\boldsymbol{\psi}_K$ be basis functions satisfying Assumption 4.4 and let $\{K_n\}$ be a sequence of basis dimensions. Then $\mathcal{H}_n = \{\gamma' \boldsymbol{\psi}_{K_n} : \gamma \in \Gamma_n\}$, where

$$\Gamma_n = \left\{ \gamma \in \mathbb{R}^{K_n} : \gamma'\gamma = 1, \text{ and } \gamma' \left[\sum_{0 \leq |\mathbf{s}| \leq s} \int \left[D^{\mathbf{s}} \boldsymbol{\psi}_{K_n}(b)\right] \left[D^{\mathbf{s}} \boldsymbol{\psi}_{K_n}(b)\right]' \mathrm{d}b \right] \gamma \leq C^2 \right\}.$$

Assumption 4.4 restricts the basis functions $\boldsymbol{\psi}_K$ to be used in estimation. Assumption 4.4 (i) is an orthonormality restriction that is not stringent; it can be ensured by conducting the Gram-Schmidt procedure. Assumption 4.4 (ii) is the most important and requires that $\boldsymbol{\psi}_{K_n}$ provide approximation rates that are at least as good as K_n^{-s/d_X}. For linear approximation, this is the best known rate and can be guaranteed by polynomial splines of degree $s-1$ with $K+1-s$ knots — see Birman and Solomjak (1967) or Dahmen et al. (1980). Chen (2007) provides conditions under which Assumption 4.4 (ii) can be guaranteed by other classes of basis functions such as Fourier series, orthogonal polynomials and wavelets.

Assumption 4.5 defines the sieve spaces $\mathcal{H}_n$. The quadratic equality constraint in the definition of Γ_n enforces production of a density upon squaring the linear approximation. The subsequent quadratic inequality imposes $\|\gamma' \boldsymbol{\psi}_{K_n}\|_{s,2} \leq C$, mirroring restriction (4.3) on $\mathcal{H}$ in $\mathcal{H}_n$.

Theorem 4.1 (Consistency)**.** *Suppose Assumptions 2.1, 2.2 and 4.1-4.5 hold. Suppose also that $K_n \to \infty$ and $\frac{K_n}{n} \to 0$ as $n \to \infty$. Then, the sieve MLE in (7) satisfies*

$$\|\widehat{\alpha}_n - \alpha_0\|_2^2 + \int \left(|\widehat{h}_n(b)| - h_0(b)\right)^2 \mathrm{d}b = o_{\mathrm{p}}(1). \tag{11}$$

The corresponding density estimator is also consistent in the $\mathscr{L}_1$ norm:

$$\int \left|\widehat{h}_n(b)^2 - h_0(b)^2\right| \mathrm{d}b = o_{\mathrm{p}}(1). \tag{12}$$

Theorem 4.1 establishes consistency of the sieve MLE in (11). Distances in $\mathcal{A} \times \mathcal{H}$ are measured by combining the Euclidean norm for $\mathcal{A}$ and the $\mathscr{L}_2$-norm on $\mathcal{H}$. The absolute value around $\widehat{h}_n$

accounts for the fact that the sign of $\widehat{h}_n$ is inconsequential. $\mathscr{L}_1$-consistency of the corresponding density estimator is established in (12); this follows from (11) and the fact that $\mathscr{L}_1$ distance between densities is bounded above by twice the $\mathscr{L}_2$ distance between their square roots. The proof of Theorem 4.1 verifies the regularity conditions of Proposition 10 of Freyberger and Masten (2019). The key ingredients in applying their result are Lemma 4.1, a compactification argument for $\mathcal{H}$ and a uniform weak law of large numbers of the form $\sup_{\theta\in\mathcal{A}\times\mathcal{H}_n}|\ell_n(\alpha,h)-\ell_*(\alpha,h)|\xrightarrow{\mathrm{p}}0$. The last requirement is established using empirical process arguments. The $\mathscr{L}_1$-consistency of densities, (14), implies consistency of the associated distribution functions in the Lévy-Prohorov metric. This latter metric is used by Fox et al. (2016) for showing consistency of fixed-grid estimators. Hence, the consistency results in Theorem 4.1 are comparable to theirs.

Theorem 4.2 (Convergence Rate)**.** *Suppose Assumptions 2.1, 2.2 and 4.1-4.5 hold. Suppose also that $K_n\to\infty$ and $\frac{K_n}{\sqrt{n}}\to 0$ as $n\to\infty$. Then, the sieve MLE in (7) satisfies*

$$\left\{\|\widehat{\alpha}_n-\alpha_0\|_2^2+\int\left(|\widehat{h}_n(b)|-h_0(b)\right)^2\mathrm{d}b\right\}^{\frac{1}{2}}=O_{\mathrm{p}}\left(\max\left\{K_n^{-s/d_X},\sqrt{\frac{K_n}{n}}\right\}\right). \tag{13}$$

The corresponding density estimator converges in $\mathscr{L}_1$ at the same rate:

$$\int\left|\widehat{h}_n(b)^2-h_0(b)^2\right|\mathrm{d}b=O_{\mathrm{p}}\left(\max\left\{K_n^{-s/d_X},\sqrt{\frac{K_n}{n}}\right\}\right). \tag{14}$$

Theorem 4.2 shows that error in estimation of α_0 and h_0 consists of two components: a deterministic "bias" term with decay rate K_n^{-s/d_X} and a "standard deviation" (or "variance") term with decay rate $\sqrt{K_n/n}$. The "bias" term arises because the linear approximation is never exact; this approximation error is controlled by Assumption 4.4 (ii). The variance term arises from random sampling error in estimating approximating coefficients. The convergence rate (13) is established using Theorem 3.2 in Chen (2007). The key technical step uses empirical process tools to characterizes an upper bound on the modulus of continuity of the centered empirical process indexed by log-likelihoods. Finally, the rate for the density (14) follows from the rate for its square root (13) by the same arguments in the discussion for Theorem 4.1.

The rate result in (13) characterizes what growth rate of K_n relative to sample size n gives an optimal rate of convergence for the sieve MLE. These are:

$$K_n=K_0n^{\frac{d_X}{2s+d_X}}\quad\text{and}\quad\left\{\|\widehat{\alpha}_n-\alpha_0\|_2^2+\int\left(|\widehat{h}_n(b)|-h_0(b)\right)^2\mathrm{d}b\right\}^{\frac{1}{2}}=O_{\mathrm{p}}\left(n^{-\frac{s}{2s+d_X}}\right) \tag{15}$$

where $K_0>0$ is a positive fixed constant. The optimal rate comes from a familiar bias-variance tradeoff: choosing a larger K_n reduces the bias term but increases the variance term. The optimal choice of K_n makes each of these grow at the same rate. The optimal $\mathscr{L}_1$ convergence rate for associated densities is also (15) due to (14). The optimal rate in (15) is the standard nonparametric optimal rate when the underlying primitives are s-smooth in the Sobolev sense.

4.2 Asymptotic normality for the plug-in estimator of averaging functionals

Recall that the true value of the averaging functional (5) and the plugin estimator (6) are:

$$\begin{aligned}\tau_0 &= \int\int t\left(w, x; \alpha_0, b\right) h_0(b)^2 \, \mathrm{d}b \, G_0(\mathrm{d}w, \mathrm{d}x),\\ \widehat{\tau}_n &= \frac{1}{n}\sum_{i=1}^{n}\int t\left(W_i, X_i; \widehat{\alpha}_n, b\right) \widehat{h}_n(b)^2 \, \mathrm{d}b.\end{aligned}$$

This subsection will treat asymptotic normality of $\widehat{\tau}_n - \tau_0$ upon scaling. I first state two additional regularity conditions. Then, I state the asymptotic normality result in studentized form with a data-based variance estimator. This collects two results into one; the separate claims of asymptotic normality with a "population limit variance" and consistent variance estimation are proven in the appendix. Discussion of the result is provided after its statement.

Assumption 4.6. There is a function of covariates $\overline{T}(w, x)$ with $0 < \int \overline{T}^2 \mathrm{d}G_0 < \infty$, such that for every $\alpha_1, \alpha_2 \in \mathcal{A}$ and $b_1, b_2 \in \mathcal{B}$,

$$\left|t\left(w, x; \alpha_1, b_1\right) - t\left(w, x; \alpha_2, b_2\right)\right| \leq \overline{T}(w, x)\left(\|\alpha_1 - \alpha_2\|_2^2 + \|b_1 - b_2\|_2^2\right)^{1/2}.$$

Furthermore, there is a pair $\left(\overline{\alpha}, \overline{b}\right) \in \mathcal{A} \times \mathcal{B}$ for which $\int t\left(w, x; \overline{\alpha}, \overline{b}\right)^2 G_0(\mathrm{d}w, \mathrm{d}x) < \infty$.

Assumption 4.7. For each (w, x, b), $t(w, x; \alpha, b)$ is twice continuously differentiable in α and its derivatives satisfy the integrability conditions:

$$\begin{aligned}\int \sup_{\alpha\in\mathcal{A}, b\in\mathcal{B}} \left\|\frac{\partial}{\partial\alpha} t(w, x; \alpha, b)\right\|_2 G_0(\mathrm{d}w, \mathrm{d}x) &< \infty,\\ \text{and } \int \sup_{\alpha\in\mathcal{A}, b\in\mathcal{B}} \left\|\frac{\partial^2}{\partial\alpha\partial\alpha'} t(w, x; \alpha, b)\right\|_2 G_0(\mathrm{d}w, \mathrm{d}x) &< \infty.\end{aligned}$$

In addition to the assumptions listed above, asymptotic normality requires controlling the behavior of the Hessian of the log-likelihood in certain neighborhoods of the true value (α_0, h_0). The notion of a gradient and Hessian in infinite dimensions, and the particular neighborhoods of (α_0, h_0) all require additional definitions. Furthermore, the nature of uniform control of the behavior of this infinite dimensional Hessian requires an additional assumption. The requisite definitions are collected in Appendix C.3 and the additional assumption is stated in Assumption C.1.

Theorem 4.3. *Suppose Assumptions 2.1, 2.2 and 4.1-4.7 and C.1 all hold. Assume further as* $n \to \infty$, $K_n \to \infty$, $\frac{K_n}{\sqrt{n}} \to 0$ *and* $\sqrt{n} \cdot K_n^{-s/d_X} \to 0$. *Then, for the variance estimator* $\widehat{V}_{\tau,n}$ *defined in* (18) *below,*

$$\frac{\sqrt{n}\cdot\left(\widehat{\tau}_n - \tau_0\right)}{\sqrt{\widehat{V}_{\tau,n}}} \xrightarrow{\mathrm{d}} \mathcal{N}(0, 1). \tag{16}$$

Furthermore, there is $V_\tau \in (0, \infty)$ *such that* $\widehat{V}_{\tau,n} \xrightarrow{\mathrm{p}} V_\tau$.

Theorem 4.3 shows asymptotic standard normality of the studentized statistic in (16). Furthermore, the subsequent result that $\widehat{V}_{\tau,n} = \widehat{V}_{\tau,n} \xrightarrow{\mathrm{p}} V_\tau \in (0,\infty)$ shows that $\widehat{\tau}_n$ is a $n^{-1/2}$-consistent estimator of τ_0. Thus, averaging functionals can be estimated at the usual parametric rate. The additional condition that $\sqrt{n} \cdot K_n^{-s/d_X} \to 0$ is an undersmoothing condition since it requires K_n to grow faster than the optimal one in (15). The undersmoothing condition prevents approximation error from appearing in the limit distribution as a bias term.

The variance estimator $\widehat{V}_{\tau,n}$ to be described involves accounting for two sources of estimation error in $\widehat{\tau}_n - \tau_0$. One is from estimating the model primitives, the other is from sample averaging to estimate G_0. To see this, first define the following:

$$\begin{aligned} T_1(\alpha, b) &= \int t(w, x; \alpha, b) G_0(\mathrm{d}w, \mathrm{d}x), \\ T_{2,0}(w,x) &= \int t\left(w, x; \alpha_0, b\right) h_0(b)^2 \mathrm{d}b, \\ \widehat{T}_{2,n}(w,x) &= \int t\left(w, x; \widehat{\alpha}_n, b\right) \widehat{h}_n(b)^2 \mathrm{d}b. \end{aligned} \tag{17}$$

Let $\mu_n[g] := (1/n)\sum_{i=1}^n \left\{g\left(Y_i, W_i, X_i\right) - \mathbb{E}\left[g(Y,W,X)\right]\right\}$ denote the centered empirical process. The difference between the plug-in estimator and the true value is

$$\begin{aligned} \widehat{\tau}_n - \tau_0 &= R_{1,n} + R_{2,n} + R_{3,n}, \\ \text{where} \quad R_{1,n} &= \int T_1\left(\widehat{\alpha}_n, b\right) \widehat{h}_n(b)^2 \mathrm{d}b - \int T_1\left(\alpha_0, b\right) h_0(b)^2 \mathrm{d}b, \\ R_{2,n} &= \mu_n\left[T_{2,0}\right] = \frac{1}{n}\sum_{i=1}^n \left\{T_{2,0}\left(W_i, X_i\right) - \mathbb{E}\left[T_{2,0}(W,X)\right]\right\}, \\ R_{3,n} &= \mu_n\left[\widehat{T}_{2,n} - T_{2,0}\right]. \end{aligned}$$

The third term is a negligible remainder — Assumptions 4.1, 2.1 and 4.6 imply a Donsker property sufficient for $\sqrt{n} \cdot R_{3,n} = o_{\mathrm{p}}(1)$. The second term $R_{2,n}$ captures error due to sample averaging to estimate G_0. $\sqrt{n} \cdot R_{2,n}$ is asymptotically normal by the usual central limit theorem. $R_{1,n}$ captures error from estimating (α_0, h_0) if G_0 were known. The key to establishing (16) is showing that $\sqrt{n} \cdot R_{1,n}$ is asymptotically normal.

Shen (1997), Chen et al. (2014) and Chen and Liao (2014) show that under some regularity conditions, an analogy to the usual "CLT + Delta Method" argument for parametric models extends to functionals of sieve estimators. In particular, functional derivatives (or rather their Riesz representers) replace Delta Method gradients. Assumptions 4.6 and 4.7 and the quadratic structure of $\tau(\cdot)$ help to ensure that these functional derivatives are bounded in an appropriate sense. In particular, a sufficient condition for a mapping of the form $h \mapsto \int T(b)h(b)^2 \mathrm{d}b$ to be "smooth" in the sense of having a derivative operator at h_0 that is a continuous (or equivalently a bounded) linear operator is for uniform boundedness of $\int T(b)^2 h(b)^2 \mathrm{d}b$ in a $\mathscr{L}_2$-ball around h_0. For a rigorous statement and proof of this smoothness property of the mean mapping $h \mapsto \int T(b)h(b)^2 \mathrm{d}b$, see for example Proposition A.5.2 in Bickel et al. (1998). Combining the Delta Method analogy with the additional term due to averaging, the variance estimator $\widehat{V}_{\tau,n}$ in (16) is:

$$\widehat{V}_{\tau,n} = \frac{1}{n}\sum_{i=1}^{n}\left(\widetilde{T}_{1,n}\left(Y_i, W_i, X_i\right) + \widetilde{T}_{2,n}\left(W_i, X_i\right)\right)^2. \tag{18}$$

In (18), the term $\widetilde{T}_{1,n}$ captures estimation error due to estimating $\widehat{\alpha}_n$ and $\widehat{h}_n$, whereas $\widetilde{T}_{2,n}$ captures estimation error due to sample averaging. These can be defined as follows: with $\widehat{T}_{2,n}$ in (17),

$$\begin{aligned}\widetilde{T}_{1,n}(y,w,x) &= \widehat{\boldsymbol{\lambda}}_n' \widehat{\boldsymbol{\eta}}_n(y,w,x),\\ \widetilde{T}_{2,n}(w,x) &= \widehat{T}_{2,n}(w,x) - \frac{1}{n}\sum_{i=1}^{n}\widehat{T}_{2,n}\left(W_i, X_i\right),\end{aligned}$$

and

$$\begin{aligned}\widehat{\boldsymbol{\eta}}_n(y,w,x) &= \frac{1}{P\left(y,w,x;\widehat{\theta}_n\right)}\begin{bmatrix}\int \frac{\partial}{\partial\alpha}\kappa\left(y,w,x;\widehat{\alpha}_n,b\right)\widehat{h}_n(b)^2\,\mathrm{d}b\\ 2\int \kappa\left(y,w,x;\widehat{\alpha}_n,b\right)\widehat{h}_n(b)\boldsymbol{\psi}_{K_n}(b)\mathrm{d}b\end{bmatrix},\\ \widehat{\boldsymbol{\lambda}}_n &= \left[\frac{1}{n}\sum_{i=1}^{n}\widehat{\boldsymbol{\eta}}_n\left(Y_i,W_i,X_i\right)\widehat{\boldsymbol{\eta}}_n\left(Y_i,W_i,X_i\right)'\right]^{-1}\\ &\quad\times\frac{1}{n}\sum_{i=1}^{n}\begin{pmatrix}\int \frac{\partial}{\partial\alpha}t\left(W_i,X_i;\widehat{\alpha}_n,b\right)\widehat{h}_n(b)^2\,\mathrm{d}b\\ 2\int t\left(W_i,X_i;\widehat{\alpha}_n,b\right)\widehat{h}_n(b)\boldsymbol{\psi}_{K_n}(b)\,\mathrm{d}b\end{pmatrix}.\end{aligned} \tag{19}$$

Importantly, all of the above elements defining $\widehat{V}_{\tau,n}$ in (18) can be computed from the data.

4.3 Mitigating the curse of dimensionality through independence

In Theorem 4.2, the dimension d_X of the covariates with random coefficients slows down the rate of convergence. Specifically, the bias term in (13) is of order K_n^{-s/d_X}. A way to mitigate this curse will be through an independence assumption on the distribution of random coefficients stated below in Assumption 4.8.

Assumption 4.8. Let $l = 1, \ldots, d_X$ index components of the random coefficients β. The coefficients in β are mutually independent, so that the density h_0^2 is a product, $h_0(b)^2 = \prod_{l=1}^{d_X} h_{0,l}\left(b_l\right)^2$. for each $b' = (b_1, \ldots, b_{d_X})$.

Assumption 4.8 motivates using a product of univariate linear series approximations:

$$\widehat{h}_n(b) = \prod_{l=1}^{d_X}\widehat{\gamma}_{n,l}'\boldsymbol{\psi}_{K_n,l}\left(b_l\right). \tag{20}$$

The same number of terms, K_n, is used for each dimension l; this makes the statement of the result simpler, but is not strictly necessary. Each univariate basis $\boldsymbol{\psi}_{K_n,l}$ must satisfy the univariate counterpart of Assumption 4.4 for every $l \in \{1, \ldots, d_X\}$. The approximation coefficient estimates $\widehat{\gamma}_{n,l}$ are derived from (10).

Theorem 4.4 (Convergence Rate with Independent Random Coefficients)**.** *Suppose Assumptions 2.2, 2.1, 4.1-4.5 and 4.8 all hold. Let $\widehat{h}_n$ now be defined by (20) and $s > 1/2$ in 4.3. Assume further that the sieve dimension K_n is chosen to ensure $K_n \to \infty$ and $\frac{K_n}{n} \to 0$. Then,*

$$\left\{ \|\widehat{\alpha}_n - \alpha_0\|_2^2 + \int \left(|\widehat{h}_n(b)| - h_0(b) \right)^2 \right\}^{\frac{1}{2}} = O_{\mathrm{p}} \left(\max \left\{ K_n^{-s}, \sqrt{\frac{K_n}{n}} \right\} \right). \tag{21}$$

The corresponding density estimator converges in $\mathscr{L}_1$ norm at the same rate:

$$\int \left| \widehat{h}_n(b)^2 - f_0(b) \right| \mathrm{d}b = O_{\mathrm{p}} \left(\max \left\{ K_n^{-s}, \sqrt{\frac{K_n}{n}} \right\} \right).$$

To compare (21) and (13), note that if we use K_n terms in the original sieve, we get K_n^{-s/d_X} for the bias term without the independence restriction, whereas we get K_n^{-s} for the bias term with the independence restriction which is considerably smaller. The variance terms stay the same (up to dimension-dependent positive multiplicative constants). The optimal rate with independence is also different: the optimal choice of K_n and the resulting optimal rate for $\widehat{\theta}_n$ are now

$$K_n = K_0 n^{\frac{1}{2s+1}} \quad \text{and} \quad \left\{ \|\widehat{\alpha}_n - \alpha_0\|_2^2 + \int \left(|\widehat{h}_n(b)| - h_0(b) \right)^2 \mathrm{d}b \right\}^{\frac{1}{2}} = O_{\mathrm{p}} \left(n^{-\frac{s}{2s+1}} \right), \tag{22}$$

where again $K_0 > 0$ is a positive fixed constant. Therefore, we get the one-dimensional rate of convergence in both (21)-(22) regardless of what the actual dimension d_X is. In addition, the one-dimensional rate of convergence also affects the undersmoothing condition for asymptotic normality, as will be shown in Theorem 4.5 below.

Theorem 4.5. *Suppose Assumptions 2.2, 2.1, 4.1-4.8 and C.1 all hold. Assume further that the sieve dimension K_n is chosen to ensure that as $n \to \infty$, $K_n \to \infty$, $\frac{K_n}{\sqrt{n}} \to 0$ and $\sqrt{n} \cdot K_n^{-s} \to 0$. Let $\widehat{V}_{\tau,n}$ be as in (18). Then,*

$$\frac{\sqrt{n} \cdot (\widehat{\tau}_n - \tau_0)}{\sqrt{\widehat{V}_{\tau,n}}} \xrightarrow{\mathrm{d}} \mathcal{N}(0, 1). \tag{23}$$

4.4 Sufficient conditions for point identification of model primitives

Assumption 2.2 maintains nonparametric point identification of (α_0, h_0) as a high-level assumption to be maintained throughout. In this subsection, I present and briefly discuss the nonparametric identification result proven in Fox et al. (2012) providing lower level sufficient conditions for Assumption 2.2.

Assumption 4.9. The support of the covariates with RCs, X, contains a non-empty open set.

Assumption 4.10. Suppose $d_W > 0$. Then there are two distinct alternatives $y_1, y_2 \in \mathcal{Y}$ and a point $c \in \mathbb{R}^{d_X}$ such that $(c', c')'$ is a point of support of $\left(X'_{y_1}, X'_{y_2}\right)'$. That is, c is a support point of X_{y_1} and X_{y_2} jointly. Furthermore, the matrix

$$\mathbb{E}\left[(W_{y_1} - W_{y_2})(W_{y_1} - W_{y_2})' \,|X_{y_1} = c, X_{y_2} = c\right]$$

has finite components and is non-singular.

Lemma 4.2 (Point identification)**.** *Suppose Assumption 2.1 holds. Then Assumptions 4.9 and 4.10 imply Assumption 2.2.*

Proof of Lemma 4.2. This result is a combination of Theorems 4 and 15 of Fox et al. (2012). □

Assumption 4.9 restricts covariates associated with RCs by ruling out purely discrete covariates as well as interactions between continuous covariates.[7]

If α_0 does not enter the estimation problem, i.e. $d_W = 0$, Assumption 4.9 is the only requirement for point identification of h_0. This is the content of Theorem 4 in Fox et al. (2012). Assumption 4.10 provides conditions under which α_0 can be identified separately to h_0 when α_0 needs to be estimated (i.e. $d_W > 0$). Once α_0 is separately identified, it can be treated as known and then Assumption 4.9 identifies h_0.

While Assumptions 4.9 and 4.10 may be strong, the Fox et al. (2012) identification result is the most general one I am aware of for the present setting. Should future identification results arise that incorporate discrete covariates with RCs or interactions between continuous ones with RCs, the results in previous subsections 4.1-4.3 will remain valid. This is because beyond being sufficient conditions for Assumption 2.2, these lower-level assumptions play no role in subsequent asymptotic results.

5 Simulations

5.1 Setup

In this section, the finite sample performance of the plugin estimator $\widehat{\tau}_n$ and its variance estimator $\widehat{V}_{\tau,n}$ is inspected in Monte Carlo simulations. There are $J = 4$ available alternatives. Sample sizes considered are $n = 500, 1000, 2000, 4000$. The covariates W_i and X_i are independent, with $d_W = 5$ and $d_X = 2$. The vector of covariates with NRCs, W_i, has mutually independent components, each generated from a uniform discrete distribution with mass points at $\{0.2, 0.4, 0.6, 0.8, 1\}$. This is done to match the empirical application in Section 6. The vector of covariates with RCs, X_i also has mutually independent components, each generated from a truncated standard normal distribution, truncated to have support in the interval $[-2, 2]$. The NRCs are set to $\alpha_0 = (2, 1, 2, 1, 2)$. Two candidates are considered for F_0, the distribution of RCs:

[7] If at least one component of X is an interaction between two or more of its other components, there is a surface of dimension at most $J \cdot d_X - 1$ to which the distribution of X assigns probability 1. A lower dimensional surface of a given Euclidean space always has empty interior.

1. F_0 is a product of two cosine distributions. Letting $\beta = (\beta_1, \beta_2)$, $\beta_1 \perp\!\!\!\perp \beta_2$, $\beta_1 \sim \text{Cosine}(5, 5)$ and $\beta_2 \sim \text{Cosine}(-6, 5)$.[8]

2. F_0 is a truncated mixture of bivariate normal distributions. In particular,

$$F_0 = 0.5 \cdot \mathcal{N}\left(\begin{bmatrix} 7.5 \\ -2.5 \end{bmatrix}, \Sigma_\beta\right) + 0.5 \cdot \mathcal{N}\left(\begin{bmatrix} 2.5 \\ -7.5 \end{bmatrix}, \Sigma_\beta\right)$$
$$\text{truncated to } [0, 10] \times [-11, -1]$$

   where

$$\Sigma_\beta = \begin{bmatrix} 0.50 & -0.15 \\ -0.15 & 0.50 \end{bmatrix}$$

For an observation i in the sample, a vector of RCs β_i is drawn from F_0 and a choice Y_i from $\{1, \ldots, 4\}$ is generated according to (1). Observations are i.i.d. across $i = 1, \ldots, n$. Each dataset is replicated 10,000 times.

Four examples of τ_0 in (5), the functionals of interest, are considered. These are:

1. The first component of the NRCs, $\alpha_{0,1} = 2$.

2. The mean of the first component of the RCs, $\mathbb{E}[\beta_1]$

3. The mean willingness to pay (Example 2.2), $\mathbb{E}[\text{WTP}] = \mathbb{E}[\beta_1/\beta_2]$.

4. The mean maximum utility, averaged over both RCs and covariates (W, X). This is:

$$\tau_{0,\text{max.util.}} = \mathbb{E}[\text{Max Util.}] \equiv \mathbb{E}\left[\log\left(\sum_{j=1}^{J} \exp\left(W_{ij}'\alpha_0 + X_{ij}'\beta_i\right)\right)\right]. \tag{24}$$

During estimation one has to choose the support $\mathcal{B}$. I inspect two cases: (i) the support is correctly specified, and (ii) $\mathcal{B}$ is strictly larger than the true support. In the first case, the support is $\mathcal{B} = [0, 10] \times [-11, -1]$. In the second, I set $\mathcal{B} = [0, 13] \times [-13.5, -0.5]$ which is 69% larger than the true support in terms of area.

5.2 Computational details

For the sieve maximum likelihood problem (10), I use polynomial splines with evenly spaced knots. These are piecewise polynomials that are of fixed order on an evenly spaced partition of $\mathcal{B}$. The boundaries of these partition pieces are called “knots”. I use cubic splines, so that on any given piece of $\mathcal{B}$, the estimate $\widehat{h}_n$ is a cubic polynomial. Splines are not orthonormal (they do not satisfy Assumption 4.4 (i) by default), and so I orthonormalize them before estimation. Since the function being estimated is bivariate, I use a two-fold tensor product of univariate splines to create bivariate

[8] A Cosine(μ, s) distribution is supported on $[\mu - s, \mu + s]$ and has its mean (and mode) at μ. The density of this distribution is the cosine function raised to be non-negative.

orthonormal splines. For details on tensor product spaces, see Chen (2007, p. 5573). Cubic splines correspond to $s = 3$. The undersmoothing condition then reads $\sqrt{n} \cdot K_n^{-3/d_X} \to 0$ as $n \to \infty$. I use a deterministic rule of $K_n \approx n^{0.45}$ to satisfy this condition – in particular, I choose

$$K_n = (\text{smallest integer greater than } n^{0.225})^2. \tag{25}$$

This choice satisfies the undersmoothing condition.[9] All integrals against random coefficients are computed using Gauss-Legendre quadrature (a deterministic numerical integration method).

The objective function in (10) is highly non-convex and hence optimization requires some care. In particular, most local gradient based algorithms are not guaranteed to find the optimum. I use the `nlopt` library for non-linear constrained optimization. This library implements both local and global optimization algorithms. For (10), I first use a global branch-and-bound optimization routine that conducts a global search by splitting the domain of α, γ into a number of pieces and doing derivative-based search within each piece. Upon finding a global maximum, I "polish" the optimum by doing a local search using output from the previous step as a starting point.[10] This "polishing" is a recommended step in the `nlopt` documentation for global routines.

5.3 Results

Tables 1–4 show the results of 10000 Monte Carlo simulations. Tables 1 and 2 show results when the distribution of RCs is triangular. These tables report coverage probabilities of 95% confidence intervals constructed using the nonparametric procedure. In addition, bias, standard deviation and root mean squared error are reported, each $\sqrt{n}$-scaled. When the true support is correctly specified, across the various functionals of interest, as the sample size grows, we find that bias decreases after scaling by $\sqrt{n}$, and the sampling variance stays stable after $\sqrt{n}$-scaling. Coverage probabilities are also quite close to the nominal 95% level. When $\mathcal{B}$ is larger than the true support, the bias is considerably larger. This is because over the regions of the support where the true density is zero, the estimated density can be positive, which in turn biases estimates of functionals. The performance of the plug-in estimator therefore suffers relative to the case of correct support specification. As the sample size grows, coverage does improve, but does so quite slowly due to the larger bias.

Tables 3 and 4 show results when the distribution of RCs is a mixture of normals. As before, with a correct support specification, bias decays quickly and coverage is quite good. With an incorrect support specification, the bias is considerably higher than the case with correct specification, and coverage similarly suffers. The mixture of normals is a more difficult distribution to approximate for the splines than the cosine distribution since it is multimodal. This explains the slower improvements in bias, which in turn explains the slower improvements in coverage rates. Both tables 2 and 4 highlight the sensitivity of the method to the specification of the support. This is not particularly surprising since even deterministic approximation error can be shown to be significantly higher when

[9] With $s = 3$ and $d_X = 2$ and $K_n \approx n^{0.45}$, we have $K_n^{-3/2} = n^{-0.675}$ and so, since $0.675 > 0.5$, it follows that $\sqrt{n} K_n^{-s/d_X} \to 0$ for $K_n \approx n^{0.45}$.

[10] The specific global and local algorithms used are `StoGo` and `L-BFGS` respectively.

$\mathcal{B}$ is larger than the true support in both cases. It may be possible to make the support itself an object of estimation by introducing additional (finite-dimensional) parameters through a location and scale transformation. The asymptotic theory will be quite different and is left to future research.

Param.	n	K_n	$\sqrt{n}\cdot$ Bias	$\sqrt{n}\cdot$ Std. Dev.	$\sqrt{n}\cdot$ RMSE	Cover. Prob.
$\alpha_{0,1}$	500	36	0.0247	0.2953	0.2964	0.9312
	1000	49	0.0211	0.2853	0.2861	0.9385
	2000	81	0.0122	0.2806	0.2807	0.9436
	4000	100	0.0077	0.2781	0.2782	0.9476
$\mathbb{E}[\beta_1]$	500	36	0.0536	0.4593	0.4624	0.9275
	1000	49	0.0669	0.4491	0.4540	0.9321
	2000	81	0.0406	0.4442	0.4461	0.9414
	4000	100	0.0376	0.4413	0.4429	0.9459
$\mathbb{E}[\text{WTP}]$	500	36	0.0031	0.0463	0.0464	0.9375
	1000	49	0.0024	0.0403	0.0404	0.9404
	2000	81	0.0027	0.0373	0.0374	0.9407
	4000	100	0.0012	0.0358	0.0358	0.9485
$\mathbb{E}[\text{Max Util.}]$	500	36	0.1091	0.6729	0.6817	0.9304
	1000	49	0.1197	0.6625	0.6733	0.9286
	2000	81	0.0832	0.6587	0.6639	0.9459
	4000	100	0.0791	0.6555	0.6603	0.9471

Table 1: Monte Carlo results from 10000 simulations with F_0 equal to a product of cosine distributions and correct support specification. Coverage probabilities are for 95% confidence intervals.

Param.	n	K_n	$\sqrt{n}\cdot$ Bias	$\sqrt{n}\cdot$ Std. Dev.	$\sqrt{n}\cdot$ RMSE	Cover. Prob.
$\alpha_{0,1}$	500	36	0.0710	0.3533	0.3604	0.9003
	1000	49	0.0371	0.3436	0.3456	0.9212
	2000	81	0.0359	0.3385	0.3404	0.9252
	4000	100	0.0236	0.3359	0.3367	0.9378
$\mathbb{E}[\beta_1]$	500	36	0.1254	0.6084	0.6212	0.9126
	1000	49	0.1491	0.5985	0.6168	0.9102
	2000	81	0.1246	0.5938	0.6067	0.9133
	4000	100	0.0934	0.5918	0.5991	0.9258
$\mathbb{E}[\text{WTP}]$	500	36	0.0110	0.0485	0.0497	0.8788
	1000	49	0.0048	0.0424	0.0426	0.9258
	2000	81	0.0046	0.0394	0.0397	0.9287
	4000	100	0.0040	0.0379	0.0381	0.9351
$\mathbb{E}[\text{Max Util.}]$	500	36	0.2188	0.8203	0.8490	0.9056
	1000	49	0.1489	0.8105	0.8240	0.9231
	2000	81	0.1141	0.8063	0.8143	0.9340
	4000	100	0.0822	0.8039	0.8081	0.9436

Table 2: Monte Carlo results from 10000 simulations with F_0 equal to a product of cosine distributions and $\mathcal{B}$ larger than true support. Coverage probabilities are for 95% confidence intervals.

Param.	n	K_n	$\sqrt{n}\cdot$ Bias	$\sqrt{n}\cdot$ Std. Dev.	$\sqrt{n}\cdot$ RMSE	Cover. Prob.
$\alpha_{0,1}$	500	36	0.0272	0.3019	0.3032	0.9322
	1000	49	0.0381	0.2923	0.2947	0.9295
	2000	81	0.0276	0.2878	0.2891	0.9344
	4000	100	0.0098	0.2847	0.2848	0.9478
$\mathbb{E}[\beta_1]$	500	36	0.0973	0.5053	0.5146	0.9211
	1000	49	0.0864	0.4964	0.5038	0.9255
	2000	81	0.0594	0.4899	0.4935	0.9381
	4000	100	0.0446	0.4879	0.4899	0.9477
$\mathbb{E}[\text{WTP}]$	500	36	0.0045	0.0488	0.0490	0.9372
	1000	49	0.0046	0.0449	0.0450	0.9340
	2000	81	0.0034	0.0429	0.0429	0.9434
	4000	100	0.0024	0.0418	0.0418	0.9478
$\mathbb{E}[\text{Max Util.}]$	500	36	0.1581	0.7584	0.7746	0.9282
	1000	49	0.1424	0.7464	0.7598	0.9303
	2000	81	0.1116	0.7428	0.7511	0.9379
	4000	100	0.0668	0.7401	0.7431	0.9481

Table 3: Monte Carlo results from 10000 simulations with F_0 equal to a mixture of normal distributions and correct support specification. Coverage probabilities are for 95% confidence intervals.

Param.	n	K_n	$\sqrt{n}\cdot$ Bias	$\sqrt{n}\cdot$ Std. Dev.	$\sqrt{n}\cdot$ RMSE	Cover. Prob.
$\alpha_{0,1}$	500	36	0.1084	0.3637	0.3796	0.8241
	1000	49	0.1067	0.3538	0.3695	0.8614
	2000	81	0.0672	0.3485	0.3549	0.9021
	4000	100	0.0391	0.3469	0.3491	0.9263
$\mathbb{E}\left[\beta_1\right]$	500	36	-0.3394	0.7068	0.7841	0.8045
	1000	49	-0.1816	0.6974	0.7207	0.8886
	2000	81	-0.1349	0.6915	0.6924	0.9133
	4000	100	-0.0641	0.6893	0.6901	0.9388
$\mathbb{E}\left[\text{WTP}\right]$	500	36	0.0148	0.0508	0.0529	0.8869
	1000	49	0.0162	0.0467	0.0495	0.8751
	2000	81	0.0118	0.0448	0.0463	0.9094
	4000	100	0.0080	0.0438	0.0445	0.9155
$\mathbb{E}\left[\text{Max Util.}\right]$	500	36	0.2822	1.0622	1.0991	0.8567
	1000	49	0.2809	1.0532	1.0900	0.8592
	2000	81	0.1840	1.0470	1.0631	0.9055
	4000	100	0.1336	1.0444	1.0529	0.9355

Table 4: Monte Carlo results from 10000 simulations with F_0 equal to a product of mixture of normal distributions and $\mathcal{B}$ larger than true support. Coverage probabilities are for 95% confidence intervals.

6 Empirical Application

In this section, an illustration of the use of the nonparametric estimation routine and the associated plugin procedure is provided. In public policy decisions on investment in infrastructure investments to improve safety, cost-benefit analyses are often conducted to see whether the cost of implementation is justified by society's willingness to pay for reductions in mortality risk. This exercise requires estimates of this willingness to pay for mortality risk reduction, which termed *the value of a statistical life* (or VSL). For example, in assessing road safety investments, the California Department of Transport uses a VSL of US$2.7 million. In developing nations, demand for public infrastructure investments is high, but reliable VSL estimates are scarce.

León and Miguel (2017) exploit a particular transportation scenario to estimate VSL among middle and upper class travellers travelling to and from the international aiport to the capital city in Sierra Leone. Travellers going to and from Sierra Leone by air have to cross an estuary that is 16 kilometers across at its widest point and have four alternatives available to make this journey.[11] The four alternatives are: ferry, helicopter, hovercraft or watertaxi. These alternatives vary in terms of historical accident risk, trip duration and monetary cost. Every alternative has non-zero

[11] León and Miguel (2017) report that the best available ground transport going around this estuary involves a six hour journey on potentially dangerous roads, and that they have no reports of travelers ever choosing that option.

fatal accident risk and these risks are widely reported and well known to travellers. In terms of mortality risk, these rank, in increasing order of the probability of fatal accident: water taxi (safest), hovercraft, ferry, helicopter (most dangerous). Travellers' choices among these alternatives reveal the tradeoffs they are willing to make between mortality risk and the cost of travel. The marginal rate of substitution in this trade off is the VSL.

León and Miguel (2017) use a survey of 561 travellers from 2012 to estimate mean VSL through a parametric RC logit model (1) and (2). Two covariates have RCs in their analysis: the probability of trip completion (or mortality risk) and the opportunity cost of transport (monetary cost and travel time). The negative of the ratio of the mortality risk coefficient to the cost coefficient is the VSL, since this ratio is the aforementioned marginal rate of substitution. In addition, there are five covariates with NRCs, which are quality rankings for five attributes of each alternative: comfort of seats, noise level, crowdedness, location convenience and quality of clientele. In their specification, the RCs follow independent triangular distributions with some sign restrictions. The support of the RC on probability of trip completion is restricted to be non-negative, and the support of the RC on travel cost is restricted to be non-positive. Together, the parametric assumptions and these additional restrictions give them 7 parameters to estimate: the means of the two RCs and the five NRCs.

I use the nonparametric procedure described in the present paper to estimate mean VSL and compare results to estimates from a parametric estimation procedure. For each individual in the dataset used by León and Miguel (2017), there is information on multiple (up to five) trips. I extract survey results on the most recent choice situation faced by a given individual in the dataset to create a cross-section of the overall dataset. The resulting dataset contains 561 individuals like the original one. For parametric estimation, I use simulated maximum likelihood estimation (SMLE) with the independent triangular distribution parametrization used by León and Miguel (2017). For parametric estimation, 1024 simulation draws per individual are used to compute the log-likelihood and its derivatives. For nonparametric estimation, I use (orthonormalized) bivariate cubic splines as basis functions as described in Section 5. The support is specified by enlargening the support of the triangular distribution estimated in the parametric case. In particular, for the coefficient on mortality risk, the support is roughly $[0, 20]$ and I extend this to $[0, 25]$. For the coefficient on cost, the support is roughly $[-0.032, 0.0]$ and I extend this to $[-0.05, 0]$. For integration, Gauss-Legendre (product) quadrature is used.

Table 5 shows the results from both estimation routines. Comparing the parametric and nonparametric estimates of mean RCs and NRCs, we see that these parameters are quite close. Comparing the main object of interest for policy questions however, the mean VSL, we see that we get very different estimates and confidence intervals. The parametric mean VSL of US$832,229 is substantially smaller than the nonparametric estimate of US$1,192,271. The associated confidence intervals are [US$513,482, US$1,150,975] for the parametric case and [US$838,801, US$1,545,741] for the nonparametric case. Now, while there is one-sided overlap, neither confidence interval contains the mean VSL estimate from the other routine. That is, the nonparametric confidence interval

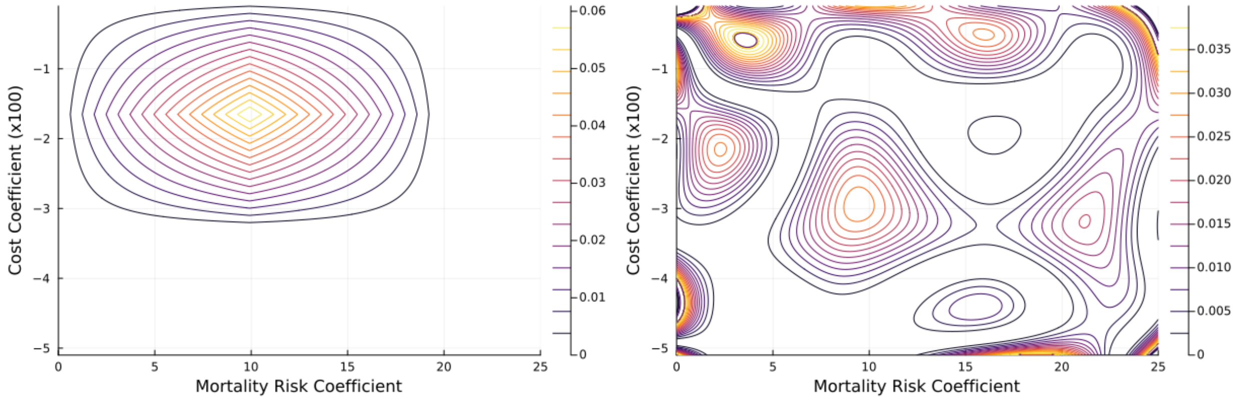

Figure 1: Density contour plots. The left panel shows the estimated triangular distribution from parametric SMLE and the right panel shows the estimated nonparametric density.

does not contain the parametric estimate, and vice versa. The difference in these two estimates arise because the parametric routine ignores dependence structures between the two random coefficients. Figure 1 plots the contours of the two estimated densities. Visual inspection of the nonparametric density plot shows a multimodal density with concentrations of high mortality risk avoidance preferences coupled with lower marginal utilities for money. In contrast, by construction, the product of independent triangular distributions imply no changes in the distribution of preferences for one attribute given preferences for the other. Here, the nonparametric density is unrestricted, and so, it was entirely possible for much lower VSL estimates than the parametric case to have occured.

León and Miguel (2017) use their parametric VSL estimate of US$597,749 to characterize the benefit due solely to improved traveller safety of an new international airport located 40km outside of Freetown. This was a real infrastructure project under public consideration and would have allowed travellers to drive to the capital city from the new airport (and vice versa), thereby removing the need for water or air transport over the estuary. The cost of this project was initially estimated at US$312 million, and was criticized under the claims that the economic benefits of the airport do not justify the cost. Under conservative assumptions[12] about the reduction in mortality risk generated by eliminating the trip across the estuary, the net present value estimated by León and Miguel (2017) due solely to mortality risk reduction was approximately US$60 million, roughly one fifth of the cost. The parametric estimate reported in the present paper brings this up to US$83.5 million. The nonparametric estimate reported in the present paper produces a higher estimate of US$119.7 million. The parametric estimate in the present paper shows that mortality risk reduction accounts for about a quarter of the cost of building the airport whereas the nonparametric estimate brings this closer to four tenths. These do not account for the other benefits of the airport such as

[12] The two conservative assumptions are that (i) ground transport will be only as safe as the safest existing transport mode across the estuary, the water taxi, and (ii) the flow business travel to and from Sierra Leone will remain constant. For (i), the road is likely safer and hence this provides a conservative starting point. For (ii), there was documented rapid increase in business travel to Sierra Leone over the years before the publication of León and Miguel (2017), and hence, flow of travel remaining constant is a conservative assumption. The reader is referred to the policy exercise in León and Miguel (2017, pp. 225–226) for exact details on the remaining calculations.

predicted growth in international trade and economic growth.

Remark 6.1 (Higher VSL estimates)**.** The parametric estimate of mean VSL reported here is higher than the original US$597,749 estimate of León and Miguel (2017). This original estimate is computed by simulating from the estimated triangular distributions. With two independent restricted triangular distributions (each with 0 at the boundary of support), the mean of the ratio is available in closed form: it is the ratio of the means multiplied by a factor of $2 \cdot \log 2$. Using this closed form expression with the mean coefficient estimates in León and Miguel (2017), an alternative estimate would instead be $1000 \times (10.155/0.019) \times 2 \times \log 2$, which gives US$740,938. Thus, the simulated estimate in León and Miguel (2017), while unbiased, under-reports mean VSL due to simulation error. The parametric estimate in this paper is higher still, but the parametric confidence interval reported in the previous paragraph contains both the simulated US$597,749 estimate and the closed form US$740,938 estimate.

Parameter	Covariate	Parametric SMLE	Nonparametric
Mean RC	Pr(trip completion)	9.9193	11.9453
		(1.3318)	(1.5188)
	Transport Cost	-0.0165	-0.02127
		(0.0013)	(0.0009)
NRC	Ranking: Comfort of seats	0.1569	0.1211
		(0.2462)	(0.2533)
	Ranking: Noise level	0.1461	0.1525
		(0.2695)	(0.2729)
	Ranking: Crowdedness	-0.8669	-0.7142
		(0.2459)	(0.2517)
	Ranking: Convenient location	-0.2768	-0.1613
		(0.2101)	(0.2091)
	Ranking: Quality of clientele	-0.3414	-0.3459
		(0.2788)	(0.2706)
Mean VSL	—	832.2292	1192.2713
		(162.6257)	(180.3421)

Table 5: Estimation results based on the León and Miguel (2017) dataset. The column "Parametric SMLE" reports parametric estimates from simulated maximum likelihood using 1024 simulation draws. The column "Nonparametric" reports nonparametric estimates using the method in this paper. The number of basis functions used was set to $K_n = 36$. For parametric mean VSL, standard errors are calculated using the parametric Delta Method.

7 Conclusion

This paper studied the question of inference in the random coefficients logit model when the distribution of RCs is estimated nonparametrically. The focus was conducting inference on functionals of the RC distribution that are averages against the distribution of RCs. Many objects of economic interest in the RC logit model can be represented as such functionals. The paper provides a nonparametric estimator of the distribution of RCs under which plug-in estimators of these functionals is asymptotically normal. Under regularity conditions, this asymptotic normality occurs at the parametric $n^{-\frac{1}{2}}$-rate. A consistent estimator of the asymptotic variance of this limiting distribution is provided. Together, these results researchers to provide consistent confidence intervals and tests of hypotheses for the functionals of interest while using a flexible nonparametric procedure to estimate the distribution of RCs. Theoretical results are confirmed through simulations. An empirical example about the value of a statistical life in Sierra Leone is used to illustrate the relevance of this method.

References

Aliprantis, Charalambos D., and Kim C. Border. 2006. *Infinite Dimensional Analysis: A Hitchhiker's Guide.* Springer-Verlag Berlin Heidelberg.

Allen, Roy, and John Rehbeck. 2023. "Identification of Random Coefficient Latent Utility Models." *Working Paper.*

Andrews, Donald W. K. 1999. "Estimation When a Parameter is on a Boundary." *Econometrica* 67 (6): 1341–1383.

———. 2000. "Inconsistency of the Bootstrap when a Parameter is on the Boundary of the Parameter Space." *Econometrica* 68 (2): 399–405.

Bajari, Patrick, Jeremy T. Fox, and Stephen P. Ryan. 2007. "Linear Regression Estimation of Discrete Choice Models with Nonparametric Distributions of Random Coefficients." *American Economic Review* 97 (2): 459–463.

Berry, Steven T. 1994. "Estimating Discrete-Choice Models of Product Differentiation." *The RAND Journal of Economics* 25 (2): 242–262.

Berry, Steven T., James Levinsohn, and Ariel Pakes. 1995. "Automobile Prices in Market Equilibrium." *Econometrica* 63 (4): 841–890.

———. 2004. "Differentiated Products Demand Systems from a Combination of Micro and Macro Data: The New Car Market." *Journal of Political Economy* 112 (1): 68–105.

Bickel, Peter J., Chris A.J Klassen, Ya'acov Ritov, and Jon A. Wellner. 1998. *Efficient and Adaptive Estimation for Semiparametric Models.* Springer-Verlag.

Bickel, Peter J., and Ya'acov Ritov. 2003. "Nonparametric Estimators Which Can Be "Plugged In"." *The Annals of Statistics* 31 (4): 1033–1053.

Bierens, Herman J. 2014. "Consistency and asymptotic normality of sieve ml estimators under low-level conditions." *Econometric Theory* 30 (5): 1021–1076.

Birman, M., and M. Solomjak. 1967. "Piecewise-polynomial approximations of functions of the classes W_p^α." *Mathematics of the USSR-Sbornik* 2 (3): 295–317.

Blundell, Wesley, Gautam Gowrisankaran, and Ashley Langer. 2020. "Escalation of Scrutiny: The Gains from Dynamic Enforcement of Environmental Regulations." *American Economic Review* 110 (8): 2558–85.

Boyd, J. Hayden, and Robert E. Mellman. 1980. "The effect of fuel economy standards on the U.S. automotive market: An hedonic demand analysis." *Transportation Research Part A: General* 14 (5): 367–378.

Bunting, Jackson. 2022. "Continuous permanent unobserved heterogeneity in dynamic discrete choice models." *arXiv pre-print,* arXiv: 2202.03960.

Cardell, N. Scott, and Frederick C. Dunbar. 1980. "Measuring the societal impacts of automobile downsizing." *Transportation Research Part A: General* 14 (5): 423–434.

Chen, Xiaohong. 2007. "Chapter 76 Large Sample Sieve Estimation of Semi-Nonparametric Models," edited by James J. Heckman and Edward E. Leamer, 6:5549–5632. Handbook of Econometrics. Elsevier.

Chen, Xiaohong, Yanqin Fan, and Viktor Tsyrennikov. 2006. "Efficient Estimation of Semiparametric Multivariate Copula Models." *Journal of the American Statistical Association* 101 (475): 1228–1240.

Chen, Xiaohong, and Zhipeng Liao. 2014. "Sieve M inference on irregular parameters." *Journal of Econometrics* 182 (1): 70–86.

Chen, Xiaohong, Zhipeng Liao, and Yixiao Sun. 2014. "Sieve inference on possibly misspecified semi-nonparametric time series models." *Journal of Econometrics* 178:639–658.

Chen, Xiaohong, Oliver Linton, and Ingrid Van Keilegom. 2003. "Estimation of Semiparametric Models when the Criterion Function Is Not Smooth." *Econometrica* 71 (5): 1591–1608. https://doi.org/https://doi.org/10.1111/1468-0262.00461.

Chen, Xiaohong, and Demian Pouzo. 2015. "Sieve Wald and QLR Inferences on Semi/Nonparametric Conditional Moment Models." *Econometrica* 83 (3): 1013–1079.

Chen, Xiaohong, and Xiaotong Shen. 1998. "Sieve Extremum Estimates for Weakly Dependent Data." *Econometrica* 66 (2): 289–314.

Compiani, Giovanni. 2022. "Market counterfactuals and the specification of multiproduct demand: A nonparametric approach." *Quantitative Economics* 13 (2): 545–591.

Dahmen, W., R. de Vore, and K. Scherer. 1980. "Multidimensional Spline Approximation." *SIAM Journal on Numerical Analysis* 17 (3): 380–402.

Daly, Andrew, Stephane Hess, and Kenneth Train. 2012. "Assuring finite moments for willingness to pay in random coefficient models." *Transportation* 39 (1): 19–31.

Durrett, Rick. 2019. *Probability: Theory and Examples.* Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press.

Fang, Zheng, and Andres Santos. 2018. "Inference on Directionally Differentiable Functions." *The Review of Economic Studies* 86 (1): 377–412.

Fox, Jeremy T., Kyoo il Kim, Stephen P. Ryan, and Patrick Bajari. 2011. "A simple estimator for the distribution of random coefficients." *Quantitative Economics* 2 (3): 381–418.

———. 2012. "The random coefficients logit model is identified." *Journal of Econometrics* 166 (2): 204–212.

Fox, Jeremy T., Kyoo il Kim, and Chenyu Yang. 2016. "A simple nonparametric approach to estimating the distribution of random coefficients in structural models." *Journal of Econometrics* 195 (2): 236–254.

Freyberger, Joachim, and Matthew A. Masten. 2019. "A practical guide to compact infinite dimensional parameter spaces." *Econometric Reviews* 38 (9): 979–1006.

Gallant, A. Ronald, and Douglas W. Nychka. 1987. "Semi-Nonparametric Maximum Likelihood Estimation." *Econometrica* 55 (2): 363–390.

Gallant, A. Ronald, and Geraldo Souza. 1991. "On the asymptotic normality of Fourier flexible form estimates." *Journal of Econometrics* 50 (3): 329–353.

Gautier, Eric, and Yuichi Kitamura. 2013. "Nonparametric Estimation in Random Coefficients Binary Choice Models." *Econometrica* 81 (2): 581–607.

Geyer, Charles J. 1994. "On the Asymptotics of Constrained M-Estimation." *The Annals of Statistics* 22 (4): 1993–2010.

Gibbs, Alison L., and Francis Edward Su. 2002. "On Choosing and Bounding Probability Metrics." *International Statistical Review / Revue Internationale de Statistique* 70 (3): 419–435.

Goolsbee, Austan, and Amil Petrin. 2004. "The Consumer Gains from Direct Broadcast Satellites and the Competition with Cable TV." *Econometrica* 72 (2): 351–381.

Grieco, Paul L. E., Charles Murry, Joris Pinkse, and Stephan Sagl. 2023. “Conformant and Efficient Estimation of Discrete Choice Demand Models.” *Working Paper.*

Hajivassiliou, Vassilis A., and Paul A. Ruud. 1994. “Chapter 40 Classical estimation methods for LDV models using simulation,” 4:2383–2441. Handbook of Econometrics. Elsevier.

Heiss, Florian, Stephan Hetzenecker, and Maximilian Osterhaus. 2022. “Nonparametric estimation of the random coefficients model: An elastic net approach.” *Journal of Econometrics* 229 (2): 299–321.

Hensher, David A., and William H. Greene. 2003. “The Mixed Logit model: The state of practice.” *Transportation* 30 (2): 133–176.

Hensher, David A., John M. Rose, and William H. Greene. 2015. *Applied Choice Analysis.* 2nd ed. Cambridge University Press.

Horowitz, Joel L., and Lars Nesheim. 2021. “Using Penalized Likelihood to Select Parameters in a Random Coefficients Multinomial Logit Model.” *Annals Issue: Structural Econometrics Honoring Daniel McFadden* 222 (1, Part A): 44–55.

Huber, Peter J., and Elvezio M. Ronchetti. 2009. *Robust Statistics.* Wiley Series in Probability and Statistics. Wiley.

Illanes, Gaston, and Manisha Padi. 2021. “Retirement Policy and Annuity Market Equilibria: Evidence from Chile.” *Working Paper.*

Keane, Michael, and Nada Wasi. 2013. “Comparing alternative models of heterogeneity in consumer choice behavior.” *Journal of Applied Econometrics* 28 (6): 1018–1045.

Lee, Lung-Fei. 1992. “On Efficiency of Methods of Simulated Moments and Maximum Simulated Likelihood Estimation of Discrete Response Models.” *Econometric Theory* 8 (4): 518–552.

———. 1995. “Asymptotic Bias in Simulated Maximum Likelihood Estimation of Discrete Choice Models.” *Econometric Theory* 11 (3): 437–483.

León, Gianmarco, and Edward Miguel. 2017. “Risky Transportation Choices and the Value of a Statistical Life.” *American Economic Journal: Applied Economics* 9 (1): 202–228.

Lu, Zhentong, Xiaoxia Shi, and Jing Tao. 2023. “Semi-nonparametric estimation of random coefficients logit model for aggregate demand.” *Journal of Econometrics* 235 (2): 2245–2265.

Miravete, Eugenio J., Katja Seim, and Jeff Thurk. 2022. “Robust Pass-Through Estimation in Discrete Choice Models.” *Working Paper.*

Nevo, Aviv, John L. Turner, and Jonathan W. Williams. 2016. “Usage-Based Pricing and Demand for Residential Broadband.” *Econometrica* 84 (2): 411–443.

Newey, Whitney K., and Daniel McFadden. 1994. "Chapter 36 Large sample estimation and hypothesis testing," 4:2111–2245. Handbook of Econometrics. Elsevier.

Petrin, Amil, and Kenneth E. Train. 2010. "A Control Function Approach to Endogeneity in Consumer Choice Models." *Journal of Marketing Research* 47 (1): 3–13.

Reiss, Rolf-Dieter. 1989. *Approximate Distributions of Order Statistics: With Applications to Nonparametric Statistics.* Lecture Notes in Statistics. Springer New York.

Shen, Xiaotong. 1997. "On Methods of Sieves and Penalization." *The Annals of Statistics* 25 (6): 2555–2591.

Small, Kenneth A., and Harvey S. Rosen. 1981. "Applied Welfare Economics with Discrete Choice Models." *Econometrica* 49 (1): 105–130.

Small, Kenneth A., Clifford Winston, and Jia Yan. 2005. "Uncovering the Distribution of Motorists' Preferences for Travel Time and Reliability." *Econometrica* 73 (4): 1367–1382.

Stone, Charles J. 1990. "Large-Sample Inference for Log-Spline Models." *The Annals of Statistics* 18 (2): 717–741.

Train, Kenneth, and Melvyn Weeks. 2005. "Discrete Choice Models in Preference Space and Willingness to Pay Space." In *Applications of Simulation Methods in Environmental and Resource Economics,* edited by Riccardo Scarpa and Anna Alberini, 1–16. Dordrecht: Springer Netherlands.

Train, Kenneth E. 2008. "EM Algorithms for nonparametric estimation of mixing distributions." *Journal of Choice Modelling* 1 (1): 40–69.

———. 2009. *Discrete Choice Methods with Simulation.* Cambridge university press.

———. 2016. "Mixed logit with a flexible mixing distribution." *Journal of Choice Modelling* 19:40–53.

Train, Kenneth E., Daniel L. McFadden, and Moshe Ben-Akiva. 1987. "The Demand for Local Telephone Service: A Fully Discrete Model of Residential Calling Patterns and Service Choices." *The RAND Journal of Economics* 18 (1): 109–123.

van der Vaart, Aad, and Jon Wellner. 1996. *Weak Convergence and Empirical Processes: With Applications to Statistics.* 1st ed. Springer Series in Statistics. Springer New York.

———. 2023. *Weak Convergence and Empirical Processes: With Applications to Statistics.* 2nd ed. Springer Series in Statistics. Springer New York.

van der Vaart, Aad W. 1998. *Asymptotic Statistics.* Asymptotic Statistics. Cambridge University Press.

Wang, Ao. 2022. “Sieve BLP: A semi-nonparametric model of demand for differentiated products.” *Journal of Econometrics.*

Appendix for "Nonparametric inference for a class of functionals in the random coefficients logit model"

This appendix presents proofs of the results in the main text. Throughout, all probabilities are defined on a probability space $(\Omega, \mathscr{F}, \Pr)$ assumed to be rich enough to support all random variables defined in the main text and subsequently.

A Additional remarks on assumptions

Remark A.1 (On compactness of $\mathcal{B}$)**.** A compact underlying support, $\mathcal{B}$, for distributions defined by the family $\mathcal{H}$ of root densities is required by the identification results of Fox et al. (2012). In addition, compactness of $\mathcal{B}$ allows the use of integration against unweighted Lebesgue measure to define distances in the consistency and convergence rate results in Section 4.1. The use of unweighted Lebesgue measure makes compactification of the space of root densities $\mathcal{H}$ simpler. With unbounded $\mathcal{B}$, $\mathcal{H}$ has to be defined as a subset of weighted Sobolev space. A similar compactification can be provided — see Section 4 of Freyberger and Masten (2019). This means that consistency with non-compact $\mathcal{B}$ is still possible, since at least one of the conditions of Proposition 10 in Freyberger and Masten (2019) will remain true. It is less clear what primitive conditions will be required for the remaining conditions of Proposition 10 in Freyberger and Masten (2019). The proofs of the convergence rate and asymptotic normality results directly use compactness of $\mathcal{B}$. The asymptotic normality result in particular requires showing a Donsker property for a certain collection of functions. A sufficient condition for this Donsker property utilizes compactness of $\mathcal{B}$. With unbounded $\mathcal{B}$, different sufficient conditions for this Donsker property have to be provided.

B Proof of theorems on Consistency and Convergence Rates

B.1 Proof of Lemma 4.1

Proof of Lemma 4.1. For any (w, x), by Jensen's inequality,

$$\begin{aligned}\sum_{y=0}^{J} P\left(y, w, x; \alpha_0, h_0\right) \log\left[\frac{P(y, w, x; \alpha, h)}{P\left(y, w, x; \alpha_0, h_0\right)}\right] &\leq \log \sum_{y=0}^{J} P\left(y, w, x; \alpha_0, h_0\right)\left[\frac{P(y, w, x; \alpha, h)}{P\left(y, w, x; \alpha_0, h_0\right)}\right] \\ &= \log 1 = 0.\end{aligned}$$

Hence,

$$\sum_{y=0}^{J} P_0\left(y, w, x; \alpha_0, h_0\right) \log P(y, w, x; \alpha, h) \leq \sum_{y=0}^{J} P\left(y, w, x; \alpha_0, h_0\right) \log P\left(y, w, x; \alpha_0, h_0\right) \leq 0, \quad (26)$$

where the rightmost inequality is because $P(\cdot; \alpha, h) \leq 1$. Thus, $-\infty \leq \ell_*(\alpha, h) \leq 0$. If $\ell_*(\alpha, h) = -\infty$, then $\ell_*\left(\alpha_0, h_0\right) > \ell_*(\alpha, h)$ is immediate, since $\ell_*\left(\alpha_0, h_0\right) > -\infty$ by hypothesis. So, assume that $\ell(\alpha, h) > -\infty$. Define the event

$$\mathcal{G}(\alpha, h) = \{(w, x) : P(y, w, x; \alpha, h) \neq P_0(y, w, x) \text{ for some } y \in \mathcal{Y}\} .$$

By Assumption 2.2, $G_0\left(\mathcal{G}(\alpha, h)\right) > 0$. Furthermore, on $\mathcal{G}(\alpha, h)$, since $b \mapsto \log b$ is strictly concave, Jensen's inequality further implies that (26) is strict, i.e.

$$\sum_{y=0}^{J} P\left(y, w, x; \alpha_0, h_0\right) \log P(y, w, x; \alpha, h) < \sum_{y=0}^{J} P\left(y, w, x; \alpha_0, h_0\right) \log P\left(y, w, x; \alpha_0, h_0\right) .$$

The conclusion of Lemma 4.1 follows from applying Lemma B.1 below with the measure G_0. □

Lemma B.1. *On a measure space $(\mathcal{X}, \mathfrak{A}, \nu)$, let $f, g \in \mathscr{L}_1(\nu)$. If $f \geq g$ ν-a.e. and $f > g$ on a set of positive ν-measure, then $\int f \mathrm{d}\nu > \int g \mathrm{d}\nu$.*

Proof of Lemma B.1. $f \geq g$ ν-a.e. implies $\int_A (f-g)\mathrm{d}\nu \geq 0$ for any $A \in \mathfrak{A}$. Define

$$B_m = \left\{x \in \mathcal{X} : f(x) \geq g(x) + \frac{1}{m}\right\} \quad \text{and} \quad B = \{x \in \mathcal{X} : f(x) > g(x)\} .$$

so that $B = \cup_{m=1}^{\infty} B_m$. If $\nu\left(B_m\right) = 0$ for all $m \in \mathbb{N}$, then $0 \leq \nu(B) \leq \sum_{m=1}^{\infty} \nu\left(B_m\right) = 0$, i.e. $\nu(B) = 0$. By contrapositive, the hypothesis $\nu(B) > 0$ implies $\nu\left(B_n\right) > 0$ for some $n \in \mathbb{N}$. On B_n, $f - g > 1/n$. Then $\int f \mathrm{d}\nu > \int g \mathrm{d}\nu$ since

$$\int (f-g)\mathrm{d}\nu = \int_{\mathcal{X} \setminus B_n} (f-g)\mathrm{d}\nu + \int_{B_n} (f-g)\mathrm{d}\nu \geq 0 + \frac{1}{n}\nu\left(B_n\right) > 0.$$

□

B.2 Proof of Theorem 4.1

The proof of Theorem 4.1 follows from verifying the regularity conditions of Proposition 10 of Freyberger and Masten (2019) which is stated as Lemma B.2 in Appendix B.2.1. One of the conditions is the content of Lemma 4.1. Verification of the remaining conditions is broken down into three additional lemmas stated in Appendix B.2.1 immediately after the statement of Proposition 10 of Freyberger and Masten (2019). The actual proof of Theorem 4.1 is given afterwards in Appendix B.2.2. Some additional lemmas containing useful envelope functions for the proof of Theorem 4.1 and other results in this appendix are given in Appendix B.2.3. The remaining parts of this subsection prove the additional lemmas.

B.2.1 Additional definitions and lemmas required for the proof of Theorem 4.1

Lemma B.2 (Proposition 10 in Freyberger and Masten (2019))**.** *Let (Θ, ρ) be a pseudo-metric space and suppose the following hold.*

(i) *Θ is compact in the topology induced by ρ.*

(ii) *$\ell_* : \Theta \to \mathbb{R}$ is continuous on Θ in the relative topology induced by ρ on Θ.*

(iii) *$\ell_*(\theta) \leq \ell_*(\theta_0)$ for all $\theta \in \Theta$. Furthermore, $\ell_*(\theta) = \ell_*(\theta_0)$ implies $\rho(\theta, \theta_0) = 0$.*

(iv) *$\{\Theta_K : K \in \mathbb{N}\}$ is a sequence of subsets of Θ and for each $K \in \mathbb{N}$, there is an element $\pi_K\theta_0 \in \Theta_K$ such that $\lim_{K\to\infty} \rho(\pi_K\theta_0, \theta_0) = 0$.*

(v) *The sequence of natural numbers $\{K_n : n \in \mathbb{N}\}$ and the sequence of functions $\{\ell_n : \Theta_{K_n} \to \mathbb{R}\}$ are chosen to satisfy both $K_n \to \infty$ and $\sup_{\theta\in\Theta_{K_n}} |\ell_n(\theta) - \ell_*(\theta)| \xrightarrow{\text{p}} 0$ as $n \to \infty$.*

Let the sieve estimator be defined by

$$\widehat{\theta}_n \in \operatorname*{argmax}_{\theta\in\Theta_{K_n}} \ell_n(\theta). \tag{27}$$

Then $\rho\left(\widehat{\theta}_n, \theta_0\right) \xrightarrow{\text{p}} 0$ as $n \to \infty$.

Before defining Θ, Θ_K and the pseudometric ρ in Lemma B.2, we provide some definitions of some preliminary objects. Let $\rho_{\mathscr{L}_2}$ denote the $\mathscr{L}_2$ metric:

$$\rho_{\mathscr{L}_2}(h_1, h_2)^2 = \int (h_1(b) - h_2(b))^2 \,\mathrm{d}b. \tag{28}$$

For each $K \in \mathbb{N}$, let $\widetilde{\Gamma}_K$ and $\widetilde{\mathcal{H}}_K$ denote the sets defined in Assumption 4.5 but indexed by the basis dimension rather than sample size. In particular, let Let C and s be as in Assumption 4.3 and $\boldsymbol{\psi}_K$ be basis functions satisfying Assumption 4.4. Then,

$$\begin{aligned}
\widetilde{\Gamma}_K &= \left\{\gamma \in \mathbb{R}^K : \gamma'\gamma = 1, \text{ and } \gamma' \left[\sum_{0\leq|\mathbf{s}|\leq s} \int [D^{\mathbf{s}}\boldsymbol{\psi}_K(b)] [D^{\mathbf{s}}\boldsymbol{\psi}_K(b)]' \,\mathrm{d}b\right] \gamma \leq C^2\right\}, \\
\widetilde{H}_K &= \{\gamma'\boldsymbol{\psi}_K : \gamma \in \Gamma_n\}.
\end{aligned} \tag{29}$$

Thus, the sieve space in the main body of the paper is $\mathcal{H}_n = \widetilde{\mathcal{H}}_{K_n}$. The sets Θ, Θ_K and the pseudometric ρ in Lemma B.2 for the purposes of proving Theorem 4.1 are defined in Definition B.1 below. With these so defined, it is clear that the extremum estimator $\widehat{\theta}_n$ in (27) is $\widehat{\theta}_n = \left(\widehat{\alpha}_n, \widehat{h}_n\right)$ from (7).

Definition B.1. Let $\Theta = \mathcal{A} \times \mathcal{H}$ and $\Theta_K = \mathcal{A} \times \widetilde{\mathcal{H}}_K$, where $\mathcal{H}$ is defined in Assumption 4.3 and $\widetilde{\mathcal{H}}_K$ is defined in (29). Define the pseudo-metric $\rho : \Theta \times \Theta \to \mathbb{R}$ by

$$\rho(\theta_1, \theta_2)^2 = \|\alpha_1 - \alpha_2\|_2^2 + \rho_{\mathscr{L}_2}(|h_1|, |h_2|)^2 = \|\alpha_1 - \alpha_2\|_2^2 + \int (|h_1(b)| - |h_2(b)|)^2 \,\mathrm{d}b. \tag{30}$$

where $\theta_j = (\alpha_j, h_j)$ for $j \in \{1, 2\}$ and $\|\alpha_1 - \alpha_2\|_2^2 = (\alpha_1 - \alpha_2)'(\alpha_1 - \alpha_2)$.

Theorem 4.1 follows from Lemma B.2 combined with Lemma 4.1 and Lemmas B.3 to B.5 below.

Lemma B.3. *Let Θ, ρ be as in Definition B.1. Θ is compact in the topology induced by ρ.*

Proof of Lemma B.3. See Appendix B.2.4. □

Lemma B.4. *Let Θ, ρ be as in Definition B.1. ℓ_* is continuous on Θ in the topology induced by ρ.*

Proof of Lemma B.4. See Appendix B.2.5. □

Lemma B.5. *If $K_n/n \to 0$ as $n \to \infty$, then $\sup_{\theta \in \Theta_{K_n}} |\ell_n(\theta) - \ell_*(\theta)| \xrightarrow{\mathrm{p}} 0$.*

Proof of Lemma B.5. See Appendix B.2.6. □

B.2.2 The actual proof of Theorem 4.1

Proof of Theorem 4.1. Condition (i) of Lemma B.2 is the content of Lemma B.3. Condition (ii) of Lemma B.2 is the content of Lemma B.4. Condition (iii) was shown in Lemma 4.1. The approximation requirement in condition (iv) holds by Assumption 4.4 (ii). For condition (v), $K_n \to \infty$ is assumed and the uniform convergence condition is the content of Lemma B.5.

Together, these prove (11) in Theorem 4.1. Then, (12) in Theorem 4.1 follows from (11) since

$$\begin{aligned}\int \left|\widehat{h}_n(b)^2 - h_0(b)^2\right| \mathrm{d}b &= \int \left(|\widehat{h}_n(b)| + h_0(b)\right) \left|\widehat{h}_n(b) - h_0(b)\right| \mathrm{d}b \\ &\leq \left(\int \left(|\widehat{h}_n(b)| + h_0(b)\right)^2 \mathrm{d}b\right)^{1/2} \cdot \left(\int \left(|\widehat{h}_n(b)| - h_0(b)\right)^2 \mathrm{d}b\right)^{1/2} \\ &\leq 2 \left(\int \left(|\widehat{h}_n(b)| - h_0(b)\right)^2 \mathrm{d}b\right)^{1/2}.\end{aligned}$$

The left hand side of the last inequality above is $o_{\mathrm{p}}(1)$ by (11). The last inequality above follows from Lemma B.6 below. □

Lemma B.6. *If $\int g_j^2 = 1$ for $j = 1, 2$, then $\int (g_1 + g_2)^2 \leq 4$.*

Proof of Lemma B.6. By Jensen's inequality,

$$(g_1 + g_2)^2 = 4 \left(\frac{g_1}{2} + \frac{g_2}{2}\right)^2 \leq 4 \left(\frac{g_1^2}{2} + \frac{g_2^2}{2}\right) = 2 \left(g_1^2 + g_2^2\right).$$

Integrating both sides, $\int (g_1 + g_2)^2 \leq 2 \left(\int g_1^2 + \int g_2^2\right) = 4$. □

B.2.3 Useful envelope functions for log-likelihoods and choice probabilities

Here we state some envelope function results that are useful in proving various dominance and "stochastic Lipschitz" properties. The proofs of lemmas here are given in Appendix F.1.

Lemma B.7. *Let Assumptions 2.1 and 4.1 hold. Define*

$$\overline{\ell}^*(y,x,w) = \log(J+1) + 2\left(\sum_{j=0}^{J}\|w_j\|_2\right)\cdot\mathcal{M}_{\mathcal{A}} + 2\left(\sum_{j=0}^{J}\|x_j\|_2\right)\cdot\mathcal{M}_{\mathcal{B}}, \qquad \textit{where} \quad \mathcal{M}_{\mathcal{A}} = \sup_{\alpha\in\mathcal{A}}\|\alpha\|_2, \quad \textit{and} \quad \mathcal{M}_{\mathcal{B}} = \sup_{b\in\mathcal{B}}\|b\|_2. \tag{31}$$

Given any function $h:\mathcal{B}\to\mathbb{R}$ such that $\int h(b)^2\,\mathrm{d}b = 1$, for any y,w,x,

$$|\log P(y,w,x;\alpha,h)| \le \overline{\ell}^*(y,x,w). \tag{32}$$

Proof of Lemma B.7. See Appendix F.1.1. □

Lemma B.8. *Let Θ,ρ be as defined in Definition B.1. For any y,w,x and $\theta_1,\theta_2\in\Theta$, denoting $\theta_j=(\alpha_j,h_j)$ for $j=1,2$,*

$$|P(y,w,x;\theta_1) - P(y,w,x;\theta_2)| \le U_P(y,w,x)\cdot\rho(\theta_1,\theta_2), \tag{33}$$

$$\textit{where} \quad U_P(y,w,x) = 2\sqrt{2}\max\left\{1,\sum_{j=0}^{J}\|w_j\|_2\right\}. \tag{34}$$

Proof of Lemma B.8. See Appendix F.1.2. □

Lemma B.9. *For any y,w,x and $\theta_1,\theta_2\in\Theta$, denoting $\theta_j=(\alpha_j,h_j)$ for $j=1,2$,*

$$|\log P(y,w,x;\theta_1) - \log P(y,w,x;\theta_2)| \le U(y,w,x)\rho(\theta_1,\theta_2), \tag{35}$$

$$\textit{where} \quad U(y,w,x) = \exp\left(\overline{\ell}^*(y,w,x)\right)\cdot U_P(y,w,x) \tag{36}$$

Proof of Lemma B.9. See Appendix F.1.3. □

B.2.4 Proof of Lemma B.3

Proof of Lemma B.3. Consider the relation on Θ, $\theta_1 \leftrightarrow \theta_2$ if and only if $\rho(\theta_1,\theta_2)=0$ for every $\theta_1,\theta_2\in\Theta$. It is straightforward to show that $\leftrightarrow$ is an equivalence relation. The topology induced by ρ on Θ is equivalent to the topology it induces on the equivalence classes under $\leftrightarrow$. Let $\mathcal{H}_+ = \{h\in\mathcal{H} : h\ge 0 \text{ Leb-a.e.}\}$ be the set of (equivalence classes of almost everywhere equal) non-negative elements of $\mathcal{H}$. The set of equivalence classes under $\leftrightarrow$ is isomorphic to $\mathcal{A}\times\mathcal{H}_+$. The pseudo metric ρ in Definition B.1 is now a metric on $\Theta_+ = \mathcal{A}\times\mathcal{H}_+$. By the previous equivalence and isomorphism arguments, proving compactness of Θ under ρ is equivalent to proving compactness of Θ_+ under ρ.

ρ metrizes the product topology on Θ_+ where we endow $\mathcal{A}$ with the Euclidean topology and $\mathcal{H}_+$ with the $\mathscr{L}_2$ topology. A product of compact sets is compact in the product topology, and hence it is sufficient to show compactness of $\mathcal{A}$ and $\mathcal{H}_+$ separately in their respective topologies. $\mathcal{A}$ is compact in $\mathbb{R}^{d_W}$ by Assumption 4.1. The Sobolev ball $\mathscr{W}_{s,2}(\mathcal{B},C)$ is $\mathscr{L}_2$-compact by Theorem 1 of

Freyberger and Masten (2019). The set $\mathcal{H}$ is a closed subset of $\mathscr{W}_{s,2}(\mathcal{B},C)$ (it is the intersection of $\mathscr{W}_{s,2}(\mathcal{B},C)$ and the surface of the $\mathscr{L}_2$ unit sphere). Hence, $\mathcal{H}$ is $\mathscr{L}_2$-compact. $\mathcal{H}_+$ is a closed subset of $\mathcal{H}$ (it is the set of non-negative elements of $\mathcal{H}$). Thus, $\mathcal{H}_+$ is also a compact subset of $\mathscr{L}_2$. □

B.2.5 Proof of Lemma B.4

Proof of Lemma B.4. Let θ_n be a sequence in Θ with $\lim_{n\to\infty}\rho(\theta_n,\theta_*)=0$ for some $\theta_*\in\Theta$. Write

$$\ell_*(\theta_n)-\ell_*(\theta_*)=\mathbb{E}\left[\log P(Y,W,X;\theta_n)-\log P(Y,W,X;\theta_*)\right].$$

Showing the above difference tends to zero as $n\to\infty$ proves continuity of ℓ_*. To that end, we show that $|\log P(Y,W,X;\theta)|$ is bounded above over all $\theta\in\Theta$ by an integrable function of (Y,W,X) and that $P(y,w,x;\theta_n)$ converges to $P(y,w,x;\theta_*)>0$ for each y,w,x. The above displayed difference then limits to zero by the dominated convergence theorem.

For dominance, Lemma B.7 shows in (32) that

$$|\log P(y,w,x;\alpha,h)|\le\overline{\ell}^*(y,x,w)$$

where $\overline{\ell}^*(y,x,w)$ is defined in (31) as

$$\begin{aligned}\overline{\ell}^*(y,x,w)=&\ \log(J+1)+2\left(\sum_{j=0}^{J}\|w_j\|_2\right)\cdot\mathcal{M}_{\mathcal{A}}+2\left(\sum_{j=0}^{J}\|x_j\|_2\right)\cdot\mathcal{M}_{\mathcal{B}},\\ \text{where}\quad \mathcal{M}_{\mathcal{A}}=&\ \sup_{\alpha\in\mathcal{A}}\|\alpha\|_2,\quad\text{and}\quad\mathcal{M}_{\mathcal{B}}=\sup_{b\in\mathcal{B}}\|b\|_2.\end{aligned}$$

$\overline{\ell}^*(y,x,w)$ is integrable by Assumption 4.2.

Then, we need to show convergence of the choice probabilities. To that end, (33) in Lemma B.8 shows that

$$|P(y,w,x;\theta_n)-P(y,w,x;\theta_*)|\le2\sqrt{2}\max\left\{1,\sum_{j=0}^{J}\|w_j\|_2\right\}\rho(\theta_n,\theta_*).$$

Hence, $P(y,w,x;\theta_n)\to P(y,w,x;\theta_*)>0$ as $n\to\infty$ for any y,w,x, where positivity follows by $\kappa(\cdot)\in(0,1)$. Thus, $\log P(y,w,x;\theta_n)\to\log P(y,w,x;\theta_*)$ as $n\to\infty$ for any y,w,x. Together with the dominance result in (32), we get continuity of the expected log-likelihood. □

B.2.6 Proof of Lemma B.5

Proof of Lemma B.5. Let ν_n denote the empirical process with respect to $\{Y_i, W_i, X_i\}_{i=1}^n$ and ν_0 denote the true distribution of (Y, W, X) by ν_0 so that

$$\nu_n[g] = \frac{1}{n}\sum_{i=1}^{n} g\left(Y_i, W_i, X_i\right), \tag{37}$$

$$\nu_0[g] = \mathbb{E}\left[g(Y, W, X)\right]. \tag{38}$$

Let $\mathcal{L}_n = \{\log P(\cdot;\theta) : \theta \in \Theta_{K_n}\}$ denote the set of log-likelihood values indexed by the sieve space, and set

$$\|\nu_n - \nu_0\|_{\mathcal{L}_n} = \sup_{\ell \in \mathcal{L}_n} \left|\left[\nu_n - \nu_0\right](\ell)\right| = \sup_{\theta \in \Theta_{K_n}} \left|\ell_n(\theta) - \ell_*(\theta)\right|.$$

The task is then to prove that $\|\nu_n - \nu_0\|_{\mathcal{L}_n} = o_{\mathrm{p}}(1)$. To that end, we apply Theorem 2.4.3 in van der Vaart and Wellner (1996). The conditions of Theorem 2.4.3 in van der Vaart and Wellner (1996) are first are shown to hold through the use of Theorem 2.7.11 in van der Vaart and Wellner (1996).

Let $\theta_1, \theta_2 \in \Theta_K$ be given with $\theta_j = (\alpha_j, h_j)$ for $j = 1, 2$. Then, (35) and (36) in Lemma B.9 show that

$$\left|\log P\left(y, w, x; \theta_1\right) - \log P\left(y, w, x; \theta_2\right)\right| \leq U(y, w, x)\rho\left(\theta_1, \theta_2\right),$$

for a non-negative function $U(\cdot)$. By Assumption 4.2, U is ν_0-integrable, i.e. $\nu_0[U] < \infty$.

Let the ε-covering number of Θ_{K_n} in the metric ρ be $N\left(\varepsilon, \Theta_{K_n}, \rho\right)$. Then, by Theorem 2.7.11 in van der Vaart and Wellner (1996), for any norm $\|\cdot\|$ on $\mathcal{L}_n$,

$$\log N_{[]}\left(\varepsilon, \mathcal{L}_n, \|\cdot\|\right) \leq \log N\left(\varepsilon / \left\|U\right\|, \Theta_{K_n}, \rho\right),$$

where $N_{[]}\left(\varepsilon, \mathcal{L}_n, \|\cdot\|\right)$ is the ε-bracketing number of $\mathcal{L}_n$ with respect to the norm $\|\cdot\|$. By Lemma B.10 below, we can bound the right hand side of the above display as

$$\begin{aligned}\log N_{[]}\left(\varepsilon, \mathcal{L}_n, \|\cdot\|\right) \leq &\log N\left(\varepsilon / \left\|U\right\|, \Theta_{K_n}, \rho\right)\\ &\leq d_W \cdot \log\left(3\sqrt{2}\mathcal{M}_{\mathcal{A}}\left\|U\right\|/\varepsilon\right) + (K_n - 1)\cdot \log\left(3\sqrt{2}\left\|U\right\|/\varepsilon\right) + \log\left(2K_n\right).\end{aligned}$$

Next, we apply Theorem 2.4.3 of van der Vaart and Wellner (1996). Working with the normed space $\mathscr{L}_1\left(\nu_n\right)$, we have $\log N\left(\varepsilon, \mathcal{L}_n, \mathscr{L}_1\left(\nu_n\right)\right) \leq \log N_{[]}\left(\varepsilon, \mathcal{L}_n, \mathscr{L}_1\left(\nu_n\right)\right)$. That is, the covering number is dominated by the bracketing number. Since U is a non-negative function, its $\mathscr{L}_1\left(\nu_n\right)$ norm is $\nu_n[U]$. Therefore,

$$\log N\left(\varepsilon, \mathcal{L}_n, \mathscr{L}_1\left(\nu_n\right)\right) \leq d_W \cdot \log\left(3\sqrt{2}\mathcal{M}_{\mathcal{A}}\nu_n\left[U\right]/\varepsilon\right) + (K_n - 1)\cdot\log\left(3\sqrt{2}\nu_n\left[U\right]/\varepsilon\right) + \log\left(2K_n\right).$$

By Assumption 4.2, $0 < \nu_0[U] < \infty$ and so, $\nu_n\left[U\right] \overset{\text{a.s.}}{\to} \nu_0\left[U\right]$ by the Strong Law of Large Numbers. Since $K_n/n \to 0$ as $n \to \infty$, it then follows that $\log N\left(\varepsilon, \mathcal{L}_n, \mathscr{L}_1\left(\nu_n\right)\right)/n = o_{\mathrm{p}}(1)$. By Theorem

2.4.3 of van der Vaart and Wellner (1996),

$$\mathbb{E}\left[\sup_{\theta\in\Theta_{K_n}} |\ell_n(\theta)-\ell_*(\theta)|\right] = \mathbb{E}\left[\|\nu_n-\nu_0\|_{\mathcal{L}_n}\right] \to 0,$$

and by Markov's inequality, $\sup_{\theta\in\Theta_{K_n}} |\ell_n(\theta)-\ell_*(\theta)| = o_{\mathrm{p}}(1)$. □

Lemma B.10. *Let Θ_K, ρ be as defined in Definition B.1. The ε-covering number of Θ_K with respect to ρ satisfies the following inequality*

$$\log N\left(\varepsilon, \Theta_K, \rho\right) \le d_W \cdot \log\left(3\sqrt{2}\mathcal{M}_{\mathcal{A}}/\varepsilon\right) + (K-1)\cdot\log(3\sqrt{2}/\varepsilon) + \log(2K), \tag{39}$$

where $\mathcal{M}_{\mathcal{A}} = \sup_{\alpha\in\mathcal{A}} \|\alpha\|_2$.

Proof of Lemma B.10. Recall that $\Theta_K = \mathcal{A}\times\widetilde{\mathcal{H}}_K$. The $(\varepsilon/\sqrt{2})$-covering number of $\mathcal{A}$ in the Euclidean norm $\|\cdot\|_2$ is bounded by $\left(3\sqrt{2}\mathcal{M}_{\mathcal{A}}/\varepsilon\right)^{d_W}$ — see for example Exercise 2.1.6 on page 94 of van der Vaart and Wellner (1996). The $\left(\varepsilon/\sqrt{2}\right)$-covering number of $\widetilde{\mathcal{H}}_K$ with respect to $\rho_{\mathscr{L}_2}$ is $2K(3\sqrt{2}/\varepsilon)^{K-1}$. To see this, take any $h_1, h_2 \in \widetilde{\mathcal{H}}_K$ and write $h_j = \boldsymbol{\psi}_K'\gamma_j$ for $j = 1, 2$. By orthonormality of $\boldsymbol{\psi}_K$ in Assumption 4.4 (i), $\|\gamma_j\|_2 = 1$ and

$$\begin{aligned}\rho_{\mathscr{L}_2}\left(h_1,h_2\right) = \int \left(h_1(b)-h_2(b)\right)^2 \mathrm{d}b &= \int \left(\left[\gamma_1-\gamma_2\right]'\boldsymbol{\psi}_K(b)\right)^2 \mathrm{d}b \\ &= (\gamma_1-\gamma_2)'\left[\int \boldsymbol{\psi}_K(b)\boldsymbol{\psi}_K(b)'\mathrm{d}b\right](\gamma_1-\gamma_2) \\ &= \|\gamma_1-\gamma_2\|_2^2 .\end{aligned}$$

The $\left(\varepsilon/\sqrt{2}\right)$-covering number of $\widetilde{\mathcal{H}}_K$ in $\rho_{\mathscr{L}_2}$ is thus bounded by the $\left(\varepsilon/\sqrt{2}\right)$-covering number of the surface of the K-dimension unit sphere with respect to the Euclidean norm $\|\cdot\|_2$. By Lemma 1 of Gallant and Souza (1991), this is bounded above by $2K((2\sqrt{2}/\varepsilon)+1)^{K-1}$.

The ε-covering number of Θ_K with respect to ρ is bounded by the product of $\left(\varepsilon/\sqrt{2}\right)$-covering numbers of $\mathcal{A}$ and $\widetilde{\mathcal{H}}_K$ in their respective norms. To see this, given $\theta_1, \theta_2 \in \Theta$ with $\theta_j = (\alpha_j, h_j)$,

$$\begin{aligned}\rho\left(\theta_1,\theta_2\right)^2 &= \|\alpha_1-\alpha_2\|_2^2 + \int \left(|h_1(b)|-|h_2(b)|\right)^2 \mathrm{d}b \\ &\le \|\alpha_1-\alpha_2\|_2^2 + \int \left(h_1(b)-h_2(b)\right)^2 \mathrm{d}b \\ &= \|\alpha_1-\alpha_2\|_2^2 + \rho_{\mathscr{L}_2}\left(h_1,h_2\right)^2 .\end{aligned}$$

The inequality above is due to $||a|-|b|| \le |a-b|$. Given finite $(\varepsilon/\sqrt{2})$-coverings of $\mathcal{A}$ and $\widetilde{\mathcal{H}}_K$, the above inequality implies that the product of these two coverings constitutes a ε-cover of Θ_K in the metric ρ. The cardinality of this product covering is the product of the individual cardinalities. Thus, (39) follows by taking the products of the covering numbers for $\mathcal{A}$ and $\widetilde{\mathcal{H}}_K$ from the previous paragraph and then taking logs. □

B.3 Proof of Theorem 4.2

This result is proven by verifying the conditions of Theorem 3.2 in Chen (2007). In Appendix B.2.1, I state three additional lemmas required for the proof of Theorem 4.2. Of these lemmas, two serve as verification of the stated conditions of Theorem 3.2 in Chen (2007) and one characterizes the modulus of continuity of the empirical process indexed by log-likelihood differences. The actual proof of Theorem 4.2 is in Appendix B.3.2. The remaining parts of this subsection prove the additional lemmas stated in Appendix B.2.1.

B.3.1 Additional lemmas required for the proof of Theorem 4.2

Lemma B.11. *Under Assumptions 2.1, 4.1 and 4.2, there is $C_1 \in (0, \infty)$ such that for all $\varepsilon > 0$,*

$$\sup_{\theta \in \Theta_{K_n} : \rho(\theta, \theta_0) \leq \varepsilon} \operatorname{Var}\left[\log P(Y, W, X; \theta) - \log P\left(Y, W, X; \theta_0\right)\right] \leq C_1 \varepsilon^2. \tag{40}$$

Proof of Lemma B.11. See Appendix B.3.3. □

Lemma B.12. *Under Assumptions 2.1, 4.1 and 4.2, there is a function $U(Y, W, X)$ such that $\mathbb{E}\left[U(Y, W, X)^2\right] < \infty$ and for any $\delta > 0$,*

$$\sup_{\{\theta \in \Theta_{K_n} : \rho(\theta, \theta_0) \leq \delta\}} \left|\log P(Y, W, X; \theta) - \log P\left(Y, W, X; \theta_0\right)\right| \leq U(Y, W, X)\delta. \tag{41}$$

Proof of Lemma B.12. See Appendix B.3.4. □

Next, let $a > 0$ and $b \geq 0$ be constants to be specified subsequently (their exact values are not important, they are required only to be finite). For $\delta \in (0, 1)$, define the entropy integral used in Theorem 3.2 of Chen (2007) (and Condition A.4 for Theorem 1 of Chen and Shen (1998)) by

$$\mathcal{J}_{n,\delta} := \int_{b\delta^2}^{a\delta} \sqrt{\log N_{[]}\left(\varepsilon, \mathcal{L}_{n,\delta}, \mathscr{L}_2\left(\nu_0\right)\right)}\, \mathrm{d}\varepsilon. \tag{42}$$

For some $C_2 \in (0, \infty)$, let

$$\delta_n = \inf\left\{\delta \in (0, 1) : \frac{1}{\sqrt{n}\delta^2} \mathcal{J}_{n,\delta} \leq C_2\right\}. \tag{43}$$

The rate of convergence shall be determined in part by δ_n above.

Lemma B.13. *There exists a constant $C \in (0, \infty)$, such that setting $a = 3C$ and $b = 0$, $\mathcal{J}_{n,\delta}$ in (42) can be bounded as follows:*

$$\mathcal{J}_{n,\delta} \leq \frac{3C\delta\sqrt{\pi}}{2} \cdot \sqrt{d_W + K_n}. \tag{44}$$

This implies that for a constant $C_3 > 0$, δ_n in (43) can be bounded as:

$$\delta_n \leq C_3 \sqrt{\frac{d_W + K_n}{n}}. \tag{45}$$

Proof of Lemma B.13. See Appendix B.3.5. □

B.3.2 The actual proof of Theorem 4.2

Proof of Theorem 4.2. We start by verifying the conditions of Theorem 3.2 in Chen (2007), in particular their conditions 3.6 to 3.8. Here, data are assumed to be i.i.d. and hence, Condition 3.6 of Chen (2007) holds by assumption. Condition 3.7 of Chen (2007) follows directly from Lemma B.11. Condition 3.8 of Chen (2007) follows directly from Lemma B.12.

By Assumption 4.4 (ii), there is $\widetilde{\gamma}_{0,n} \in \widetilde{\Gamma}_{K_n}$ such that setting $\widetilde{h}_{0,n} = \widetilde{\gamma}'_{0,n}\boldsymbol{\psi}_{K_n}$, $\rho_{\mathscr{L}_2}\left(\widetilde{h}_{0,n}, h_0\right) = O(K_n^{-s/d_X})$. Without loss of generality, we can always set $\widetilde{\gamma}_{0,n}$ equal to the $\mathscr{L}_2$ projection coefficients of h_0 onto the span of $\boldsymbol{\psi}_{K_n}$. There is no requirement for $\widetilde{\gamma}_{0,n}$ to produce a density, i.e. it may (and indeed usually will)[13] be the case that $\int \widetilde{h}_{0,n}(b)^2 \, \mathrm{d}b \neq 1$ so that $\widetilde{h}_{0,n} \notin \mathcal{H}_n$. This is not difficult to resolve. By orthonormality of $\boldsymbol{\psi}_{K_n}$, i.e. Assumption 4.4 (i), $\int \widetilde{h}_{0,n}(b)^2 \, \mathrm{d}b = \|\widetilde{\gamma}_{0,n}\|_2^2$.

$$\begin{aligned} \gamma_{0,n} &= \widetilde{\gamma}_{0,n} / \left\|\widetilde{\gamma}_{0,n}\right\|_2, \\ h_{0,n} &= \gamma'_{0,n}\boldsymbol{\psi}_{K_n}. \end{aligned} \tag{46}$$

Clearly, $\gamma'_{0,n}\gamma_{0,n} = 1$ and $\int h_{0,n}(b)^2 \, \mathrm{d}b = 1$. By Lemma B.14 below $\rho_{\mathscr{L}_2}\left(h_{0,n}, h_0\right) = O(K_n^{-s/d_X})$ so that the approximation rate is unchanged by this normalization. Since $\alpha_0 \in \mathcal{A}$, we can set $\theta_{0,n} = (\alpha_0, h_{0,n})$ so that $\theta_{0,n} \in \Theta_{K_n}$. Furthermore, given the approximation rate by $h_{0,n}$ it follows that $\rho\left(\theta_{0,n}, \theta_0\right) = O(K_n^{-s/d_X})$. This characterizes the “bias” part of the convergence rate.

By Lemma B.13 (in particular (45)), the “standard deviation” (or “variance”) part of the convergence rate is $O_{\mathrm{p}}\left(\delta_n\right) = O_{\mathrm{p}}\left(\sqrt{\left(d_W + K_n\right)/n}\right) = O_{\mathrm{p}}\left(\sqrt{K_n/n}\right)$, where the second equality follows since d_W is constant as $n \to \infty$. By Theorem 3.2 of Chen (2007),

$$\rho\left(\widehat{\theta}_n, \theta_0\right) = O_{\mathrm{p}}\left(\max\left\{K_n^{-s/d_X}, \sqrt{\frac{K_n}{n}}\right\}\right).$$

Thus, (13) holds. Since $\int \left|\widehat{h}_n(b)^2 - h_0(b)^2\right| \mathrm{d}b \leq 2\sqrt{\int \left(|\widehat{h}_n(b)| - h_0(b)\right)^2 \mathrm{d}b}$, (14) also follows. □

Lemma B.14. *Let $h \in \mathscr{L}_2$ satisfy $\int h(b)^2 \mathrm{d}b = 1$ and let $\boldsymbol{\psi} = (\psi_1, \ldots, \psi_K)'$ be a vector of orthonormal functions so that $\int \psi_j(b)\psi_k(b)\mathrm{d}b = \mathbb{I}\{j = k\}$. Let $\widetilde{\gamma}_* = \int \boldsymbol{\psi}(b)h(b) \, \mathrm{d}b$ be the $\mathscr{L}_2$ linear projection coefficients of h onto the linear span of $\boldsymbol{\psi}$ and define $\gamma_* = \widetilde{\gamma}_* / \|\widetilde{\gamma}_*\|_2$. Furthermore, let*

[13] If $\widetilde{\gamma}_{0,n}$ are equal to the $\mathscr{L}_2$ projection coefficients, then it can be shown that $\int \widetilde{h}_{0,n}(b)^2 \, \mathrm{d}b \leq 1$ with equality only if h_0 is contained in the linear span of $\boldsymbol{\psi}_{K_n}$. The “if part” follows from Parseval’s identity, and the “only if part” follows from either one of Parseval’s identity or Bessel’s inequality.

$\widetilde{h}_* = \widetilde{\gamma}_*'\boldsymbol{\psi}$ and $h_* = \gamma_*'\boldsymbol{\psi}$. Then,

$$0 \leq 1 - \|\widetilde{\gamma}_*\|_2 \leq \rho_{\mathscr{L}_2}\left(h, \widetilde{h}_*\right)^2 \leq \rho_{\mathscr{L}_2}(h, h_*)^2 = 2\left(1 - \|\widetilde{\gamma}_*\|_2\right). \tag{47}$$

Remark B.1. The qualitative content of Lemma B.14 is as follows. Given a function h whose square integrates to one, suppose we consider two approximations to it. One projects h onto a finite-dimensional linear subspace without norm preservation, i.e. unconstrained projection. The other takes this projection and normalizes it to unit norm, i.e. unit-norm constrained projection.[14] The approximation error of both is tightly characterized by the difference between one and the size of the unconstrained projection coefficients. Tightness means bounded above and below by universal constants: here the constants are 1 and 2. That is, if the rate of approximation can be written as a function of dimension so that $1 - \|\widetilde{\gamma}_*\|_2 = \zeta(K)$ for some $\zeta(\cdot)$, then both $\rho\left(h, \widetilde{h}_*\right) = O(\zeta(K))$ and $\rho(h, h_*) = O(\zeta(K))$.

Proof of Lemma B.14. Let $\mathbf{I}$ be the $K \times K$ identity matrix. I shall drop b and $\mathrm{d}b$ everywhere in integrals for this proof because they are unnecessary. Orthonormality of $\boldsymbol{\psi}$ implies $\int \boldsymbol{\psi}\boldsymbol{\psi}' = \mathbf{I}$. By $h = \widetilde{h}_* + h - \widetilde{h}_*$, and $\widetilde{h}_* = \widetilde{\gamma}_*'\boldsymbol{\psi}$,

$$\begin{aligned} 1 = \int h^2 &= \int \widetilde{h}_*^2 + 2\int \widetilde{h}_* \cdot \left(h - \widetilde{h}_*\right) + \int \left(h - \widetilde{h}_*\right)^2 \\ &= \widetilde{\gamma}_*'\left[\int \boldsymbol{\psi}\boldsymbol{\psi}'\right]\widetilde{\gamma}_* + 2\widetilde{\gamma}_*'\left[\int \boldsymbol{\psi} \cdot \left(h - \widetilde{h}_*\right)\right] + \int \left(h - \widetilde{h}_*\right)^2 \\ &= \|\widetilde{\gamma}_*\|_2^2 + \int \left(h - \widetilde{h}_*\right)^2, \end{aligned}$$

since $\int \boldsymbol{\psi} \cdot \left(h - \widetilde{h}_*\right) = \mathbf{0}$ (by $\widetilde{h}_*$ being a linear projection). Therefore, since $\int \left(h - \widetilde{h}_*\right)^2 \geq 0$, $\|\widetilde{\gamma}_*\|_2 \leq 1$ proving the far left non-negativity claim of (47). Next, the above display also shows that

$$\rho_{\mathscr{L}_2}\left(h, \widetilde{h}_*\right)^2 = \int \left(h - \widetilde{h}_*\right)^2 = 1 - \|\widetilde{\gamma}_*\|_2^2 = \left(1 + \|\widetilde{\gamma}_*\|_2\right)\left(1 - \|\widetilde{\gamma}_*\|_2\right).$$

Combining the above and $0 \leq \|\widetilde{\gamma}_*\|_2 \leq 1$, it readily follows that

$$0 \leq 1 - \|\widetilde{\gamma}_*\|_2 \leq \rho_{\mathscr{L}_2}\left(h, \widetilde{h}_*\right)^2 \leq 2\left(1 - \|\widetilde{\gamma}_*\|_2\right).$$

The full conclusion of (47) follows if we show $\rho_{\mathscr{L}_2}(h, h_*)^2 = 2\left(1 - \|\widetilde{\gamma}_*\|_2\right)$. To that end,

$$\begin{aligned} \rho_{\mathscr{L}_2}(h, h_*)^2 = \int (h - h_*)^2 &= \underbrace{\int h^2}_{=1} + \underbrace{\int h_*^2}_{=1} - 2\int h \cdot h_* \\ &= 2 - 2\gamma_*'\int \boldsymbol{\psi} \cdot h \end{aligned}$$

[14] When the basis is orthonormal, it can be shown that the solution to least squares projection with a unit norm constraint is the unconstrained least squares coefficients normalized to have unit norm.

$$
\begin{aligned}
&= 2\left(1-\gamma_*'\widetilde{\gamma}_*\right) \\
&= 2\left(1-\frac{\widetilde{\gamma}_*'\widetilde{\gamma}_*}{\|\widetilde{\gamma}_*\|_2}\right) \\
&= 2\left(1-\|\widetilde{\gamma}_*\|_2\right).
\end{aligned}
$$

□

B.3.3 Proof of Lemma B.11

Proof of Lemma B.11. (35) and (36) in Lemma B.9 show that

$$
\left|\log P\left(y,w,x;\theta_1\right)-\log P\left(y,w,x;\theta_2\right)\right| \leq U(y,w,x)\rho\left(\theta_1,\theta_2\right),
$$

for a non-negative function $U(\cdot)$. By Assumption 4.2, U is square-integrable against ν_0, i.e. $\nu_0\left[U^2\right]<\infty$. Therefore

$$
\begin{aligned}
&\operatorname{Var}\left[\log P(Y,W,X;\theta)-\log P\left(Y,W,X;\theta_0\right)\right] \\
&\leq \mathbb{E}\left[\left\{\log P(Y,W,X;\theta)-\log P\left(Y,W,X;\theta_0\right)\right\}^2\right] \\
&\leq \mathbb{E}\left[U(Y,W,X)^2\right]\rho\left(\theta,\theta_0\right)^2.
\end{aligned}
$$

Hence, (40) holds with $C_1=\mathbb{E}\left[U(Y,W,X)^2\right]$ which is finite by Assumption 4.2. □

B.3.4 Proof of Lemma B.12

Proof of Lemma B.12. (35) and (36) in Lemma B.9 show that

$$
\left|\log P\left(y,w,x;\theta_1\right)-\log P\left(y,w,x;\theta_2\right)\right| \leq U(y,w,x)\rho\left(\theta_1,\theta_2\right),
$$

for a non-negative function $U(\cdot)$. By Assumption 4.2, U is square-integrable against ν_0, i.e. $\mathbb{E}\left[U(Y,W,X)^2\right]<\infty$. Hence, (41) holds. □

B.3.5 Proof of Lemma B.13

Proof of Lemma B.13. For $\delta>0$, denote

$$
\begin{aligned}
\Theta_{K_n,\delta} &= \left\{\theta\in\Theta_{K_n}:\rho\left(\theta,\theta_0\right)\leq\delta\right\}, \\
\mathcal{L}_{n,\delta} &= \left\{\log P(\cdot;\theta)-\log P\left(\cdot;\theta_0\right):\theta\in\Theta_{K_n,\delta}\right\}.
\end{aligned}
\tag{48}
$$

By Lemma B.12 and Theorem 2.7.11 in van der Vaart and Wellner (1996), for any norm $\|\cdot\|$,

$$
\log N_{[]}\left(\varepsilon,\mathcal{L}_{n,\delta},\|\cdot\|\right)\leq\log N\left(\varepsilon/\left\|U\right\|,\Theta_{K_n,\delta},\rho\right).
$$

By Lemma B.15 below,

$$\log N_{[]}\left(\varepsilon, \mathcal{L}_{n,\delta}, \|\cdot\|\right) \leq \left(d_W + K_n\right)\log\left(3\delta\|U\|/\varepsilon\right). \tag{49}$$

Let ν_0 be the distribution of (Y, W, X). Working with the normed space $\mathscr{L}_2(\nu_0)$, denote $C^2 = \nu_0\left[U^2\right]$. Then, the above bracketing entropy bound is

$$\log N_{[]}\left(\varepsilon, \mathcal{L}_{n,\delta}, \mathscr{L}_2(\nu_0)\right) \leq \left(d_W + K_n\right) \cdot \log\left(3C\delta/\varepsilon\right).$$

We can thus bound $\mathcal{J}_{n,\delta}$ in (42)

$$\begin{aligned}
\mathcal{J}_{n,\delta} &:= \int_{b\delta^2}^{a\delta} \sqrt{\log N_{[]}\left(\varepsilon, \mathcal{L}_{n,\delta}, \mathscr{L}_2(\nu_0)\right)}\, \mathrm{d}\varepsilon \\
&\leq \sqrt{d_W + K_n} \cdot \int_{b\delta^2}^{a\delta} \log\left(3C\delta/\varepsilon\right)\, \mathrm{d}\varepsilon \\
&\leq \sqrt{d_W + K_n} \cdot \int_{0}^{a\delta} \log\left(3C\delta/\varepsilon\right)\, \mathrm{d}\varepsilon.
\end{aligned}$$

Then, set $a = 3C$ and use (158) in Lemma F.4 to conclude that

$$\mathcal{J}_{n,\delta} \leq \frac{3C\delta\sqrt{\pi}}{2} \cdot \sqrt{d_W + K_n}.$$

Hence (44) follows.

Next, apply this bound to δ_n in (43) as follows:

$$\begin{aligned}
\delta_n &:= \inf\left\{\delta \in (0,1) : \frac{1}{\sqrt{n}\delta^2}\mathcal{J}_{n,a,b,\delta} \leq C_2\right\} \\
&\leq \inf\left\{\delta \in (0,1) : \frac{3C\sqrt{\pi}}{2\sqrt{n}\delta} \cdot \sqrt{d_W + K_n} \leq C_2\right\}.
\end{aligned}$$

The solution to the last infimum problem above yields $\delta_n \leq C_3\sqrt{\frac{d_W + K_n}{n}}$ for a constant $C_3 > 0$. Hence, (45) follows. □

Lemma B.15. *Let $\Theta_{K,\delta}$ be as defined in (48). Then,*

$$\log N\left(\varepsilon, \Theta_{K,\delta}, \rho\right) \leq \left(d_W + K\right)\log\left(3\delta/\varepsilon\right). \tag{50}$$

Proof of Lemma B.15. Let

$$\mathcal{A}_{\delta/\sqrt{2}} = \left\{\alpha \in \mathcal{A} : \|\alpha - \alpha_0\|_2 \leq \delta/\sqrt{2}\right\}, \tag{51}$$

$$\widetilde{\mathcal{H}}_{K,\delta/\sqrt{2}} = \left\{h \in \widetilde{\mathcal{H}}_K : \rho_{\mathscr{L}_2}\left(h, h_0\right) \leq \delta/\sqrt{2}\right\}. \tag{52}$$

Then, $\Theta_{K,\delta} \subseteq \mathcal{A}_{\delta/\sqrt{2}} \times \widetilde{\mathcal{H}}_{K,\delta/\sqrt{2}}$. Hence, the covering number of $\Theta_{K,\delta}$ is less than that of $\mathcal{A}_{\delta/\sqrt{2}} \times$

$\widetilde{\mathcal{H}}_{K,\delta/\sqrt{2}}$. As argued in the proof of Lemma B.10, the ε-covering number of $\mathcal{A}_{\delta/\sqrt{2}} \times \widetilde{\mathcal{H}}_{K,\delta/\sqrt{2}}$ is bounded by the product of the $(\varepsilon/\sqrt{2})$-covering number of $\mathcal{A}_{\delta/\sqrt{2}}$ and the $(\varepsilon/\sqrt{2})$-covering number of $\widetilde{\mathcal{H}}_{K,\delta/\sqrt{2}}$.

The $(\varepsilon/\sqrt{2})$-covering number of $\mathcal{A}_{\delta/\sqrt{2}}$ is bounded by $(3\delta/\varepsilon)^{d_W}$ (the $1/\sqrt{2}$ on both terms cancel out) — see for example Exercise 2.1.6 on page 94 of van der Vaart and Wellner (1996). The $\left(\varepsilon/\sqrt{2}\right)$-covering number of $\widetilde{\mathcal{H}}_{K,\delta/\sqrt{2}}$ with respect to $\rho_{\mathscr{L}_2}$ is bounded by $(3\delta/\varepsilon)^K$. To see this, take any $h_1, h_2 \in \widetilde{\mathcal{H}}_{K,\delta/\sqrt{2}}$, where $h_j = \boldsymbol{\psi}_K'\gamma_j$ for $j = 1, 2$. By orthonormality of $\boldsymbol{\psi}_K$ (Assumption 4.4 (i)),

$$\begin{aligned}\rho_{\mathscr{L}_2}(h_1, h_2) &= \int (h_1(b) - h_2(b))^2 \,\mathrm{d}b = \int \left([\gamma_1 - \gamma_2]' \boldsymbol{\psi}_K(b)\right)^2 \mathrm{d}b \\ &= (\gamma_1 - \gamma_2)' \left[\int \boldsymbol{\psi}_K(b)\boldsymbol{\psi}_K(b)' \mathrm{d}b\right] (\gamma_1 - \gamma_2) \\ &= \|\gamma_1 - \gamma_2\|_2^2 .\end{aligned}$$

Thus, the $\left(\varepsilon/\sqrt{2}\right)$-covering number of $\widetilde{\mathcal{H}}_{K,\delta/2}$ is bounded above by the $\left(\varepsilon/\sqrt{2}\right)$-covering number of the Euclidean ball of radius $\delta/\sqrt{2}$ centered at the origin. This is because, $\widetilde{\mathcal{H}}_{K,\delta/\sqrt{2}}$ is contained in the ball of radius $\delta/\sqrt{2}$ around h_0, and translation by $-h_0$ does not change the covering number. By Exercise 2.1.6 on page 94 of van der Vaart and Wellner (1996), the $\left(\varepsilon/\sqrt{2}\right)$-covering number of the Euclidean ball of radius $\delta/\sqrt{2}$ centered at the origin is $(3\delta/\varepsilon)^K$. Take the product of the two covering numbers for $\mathcal{A}_{\delta/\sqrt{2}}$ and $\widetilde{\mathcal{H}}_{K,\delta/\sqrt{2}}$ respectively and take logs to get (50). □

C Asymptotic Normality

Throughout, let ν_n denote the empirical distribution associated with the data $\{Y_i, W_i, X_i\}_{i=1}^n$ and ν_0 denote the true population distribution of (Y, W, X). Furthermore, let $\mu_n = \nu_n - \nu_0$ be the centered empirical process with respect to observed data. The following linear functional notation will be used throughout: for a (measurable) function g,

$$\nu_n[g] := \frac{1}{n}\sum_{i=1}^n g\left(Y_i, W_i, X_i\right), \quad \nu_0[g] := \mathbb{E}_{\nu_0}\left[g(Y, W, X)\right] = \int g\left(y, w, x\right) \nu_n\left(\mathrm{d}y, \mathrm{d}w, \mathrm{d}x\right)$$
$$\mu_n[g] := (\nu_n - \nu_0)[g] = \frac{1}{n}\sum_{i=1}^n \left\{g\left(Y_i, W_i, X_i\right) - \mathbb{E}_{\nu_0}\left[g(Y, W, X)\right]\right\}.$$

In this part of the present section, the error breakdown of plug-in estimation introduced in Section 4.2 is stated formally as Lemma C.1. The proof of Lemma C.1 is given in Appendix D.1 and involves little more than algebra and two applications of Fubini's Theorem. Immediately afterwards, the proof of Theorem 4.3 is given. This proof is broken down into various parts, which are all provided in subsections of the present appendix section. Some of these parts will themselves involve additional results which are stated as additional lemmas with proofs deferred to Appendix D. By and large, the proofs deferred to Appendix D are ones that are lengthy.

Lemma C.1. *Let*

$$\begin{aligned} T_1(\alpha, b) &= \int t(w, x; \alpha, b) G_0(\mathrm{d}w, \mathrm{d}x), \\ T_{2,0}(w, x) &= \int t\left(w, x; \alpha_0, b\right) h_0(b)^2 \mathrm{d}b, \qquad (53) \\ \widehat{T}_{2,n}(w, x) &= \int t\left(w, x; \widehat{\alpha}_n, b\right) \widehat{h}_n(b)^2 \mathrm{d}b. \end{aligned}$$

Then

$$\begin{aligned} \widehat{\tau}_n - \tau_0 &= R_{1,n} + R_{2,n} + R_{3,n}, \qquad (54) \\ \textit{where} \quad R_{1,n} &= \int T_1\left(\widehat{\alpha}_n, b\right) \widehat{h}_n(b)^2 \mathrm{d}b - \int T_1\left(\alpha_0, b\right) h_0(b)^2 \mathrm{d}b, \qquad (55) \\ R_{2,n} &= \mu_n\left[T_{2,0}\right] = \frac{1}{n} \sum_{i=1}^n \left\{T_{2,0}\left(W_i, X_i\right) - \mathbb{E}\left[T_{2,0}(W, X)\right]\right\}, \qquad (56) \\ R_{3,n} &= \mu_n\left[\widehat{T}_{2,n} - T_{2,0}\right]. \qquad (57) \end{aligned}$$

C.1 Proof of Theorem 4.3

Proof of Theorem 4.3. By Lemma C.4, $\sqrt{n} R_{3,n} \xrightarrow{\mathrm{p}} 0$, and so from (54),

$$\sqrt{n} \cdot \left(\widehat{\tau}_n - \tau_0\right) = \sqrt{n} \cdot \left(R_{1,n} + R_{2,n}\right) + o_{\mathrm{p}}(1).$$

By (89) of Theorem C.1, there is a non-decreasing sequence of constants $V_{\phi,n}$ and a sequence of functions $\dot{\phi}_n$ that have zero mean against ν_0 (see Lemma C.12) such that

$$\frac{\sqrt{n} R_{1,n}}{V_{\phi,n}} = \frac{1}{\sqrt{n}} \sum_{i=1}^n \dot{\phi}_n\left(Y_i, W_i, X_i\right) + o_{\mathrm{p}}(1) \xrightarrow{\mathrm{d}} \mathcal{N}(0,1). \qquad (58)$$

In (58), asymptotic normality follows from showing that Lindeberg's condition holds for the triangular array $\dot{\phi}_n\left(Y_i, W_i, X_i\right)$. Since $T_{2,0}\left(W_i, X_i\right)$ has finite second moment and does not change with n, it follows that Lindeberg's condition holds for the triangular array of random vectors $\left(\dot{\phi}_n\left(Y_i, W_i, X_i\right), T_{2,0}\left(W_i, X_i\right)\right)$. Furthermore, by Lemma C.12, $\mathbb{E}\left[\dot{\phi}_n(Y, W, X) \middle| W, X\right] = 0$, so that

$$\operatorname{Cov}\left[\dot{\phi}_n(Y, W, X), T_{2,0}(W, X)\right] = 0. \qquad (59)$$

Therefore, since $R_{2,n} = \mu_n\left[T_{2,0}\right]$ in (56),

$$\sqrt{n}\begin{pmatrix} R_{1,n}/V_{\phi,n} \\ R_{2,n} \end{pmatrix} = \sqrt{n}\mu_n\left[\begin{pmatrix} \dot{\phi}_n \\ T_{2,0} \end{pmatrix}\right] + o_{\mathrm{p}}(1) \xrightarrow{\mathrm{d}} \mathcal{N}_2\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} 1 & 0 \\ 0 & \operatorname{Var}\left[T_{2,0}(W, X)\right] \end{bmatrix}\right). \qquad (60)$$

By (90) of Theorem C.1, $V_{\phi,n} \nearrow V_\phi$ for some $V_\phi \in (0,\infty)$. Combining the Cramér-Wold device and Slutsky's Theorem,

$$\sqrt{n} \cdot (R_{1,n} + R_{2,n}) = V_{\phi,n} \cdot \frac{\sqrt{n}R_{1,n}}{V_{\phi,n}} + \sqrt{n}R_{2,n} \xrightarrow{\mathrm{d}} \mathcal{N}\left(0, V_\tau\right),$$

where

$$V_\tau = V_\phi + \mathrm{Var}\left[T_{2,0}(W,X)\right].$$

Hence,

$$\sqrt{n} \cdot (\widehat{\tau}_n - \tau_0) = \sqrt{n} \cdot (R_{1,n} + R_{2,n}) + o_{\mathrm{p}}(1) \xrightarrow{\mathrm{d}} \mathcal{N}\left(0, V_\tau\right). \tag{61}$$

Thus, set

$$V_{\tau,n} = V_{\phi,n} + \mathrm{Var}\left[T_{2,0}(W,X)\right].$$

By $V_{\phi,n} \nearrow V_\phi$, it is clear that $V_{\tau,n} \nearrow V_\tau$. By (97) of Theorem C.2, $\widehat{V}_{\tau,n}$ in (18) satisfies

$$\frac{\widehat{V}_{\tau,n}}{V_{\tau,n}} = 1 + o_{\mathrm{p}}(1).$$

By the Continuous Mapping Theorem,

$$\sqrt{\widehat{V}_{\tau,n}} \xrightarrow{\mathrm{p}} \sqrt{V_\tau}.$$

By Slutsky's Theorem and (61)

$$\frac{\sqrt{n} \cdot (\widehat{\tau}_n - \tau_0)}{\sqrt{\widehat{V}_{\tau,n}}} \xrightarrow{\mathrm{d}} \mathcal{N}(0,1),$$

which is exactly (16). The rate claim has already been proven since $\widehat{V}_{\tau,n}/V_{\tau,n} = 1 + o_{\mathrm{p}}(1)$ and $V_{\tau,n} \nearrow V_\tau < \infty$. □

The remainder of this section will proceed as follows. First, it will be shown that $R_{3,n}$ is asymptotically negligible in the sense that $\sqrt{n} \cdot R_{3,n} = o_{\mathrm{p}}(1)$. Next, asymptotic normality of $R_{1,n}$ will be characterized by verifying conditions in Chen and Liao (2014). A number of additional definitions and concepts have to be introduced for this, and thus, this subsection is the longest one. Finally, $\widehat{V}_{\tau,n}$ in (18) will be shown to be a consistent estimator of V_τ by combining results from Chen and Liao (2014) with a Donsker property derived during the proof that $\sqrt{n} \cdot R_{3,n} = o_{\mathrm{p}}(1)$.

C.2 Asymptotic negligibility of the third remainder term $R_{3,n}$ in (57)

Denote the set of probability measures over $\mathcal{B}$ by $\mathrm{PM}(\mathcal{B})$. Denote the following families:

$$\mathscr{T} = \{t(\cdot;\alpha,b) : \alpha \in \mathcal{A}, b \in \mathcal{B}\}, \tag{62}$$

$$\mathscr{T}_* = \left\{\int t(\cdot;\alpha,b)F(\mathrm{d}b) : \alpha \in \mathcal{A}, F \in \mathrm{PM}(\mathcal{B})\right\}. \tag{63}$$

Lemma C.2. *Let Assumptions 2.1, 4.1 and 4.6 hold. Then the families $\mathscr{T}$ and $\mathscr{T}_*$ in (62) and (63) respectively are both G_0-Donsker. Furthermore, define the function $\overline{T}^*$ by*

$$\overline{T}^*(w,x) = |t(w,x,\overline{\alpha},\overline{b})| + 2\left(\mathcal{M}_{\mathcal{A}}^2 + \mathcal{M}_{\mathcal{B}}^2\right)^{1/2}\overline{T}(w,x) \tag{64}$$

where $\mathcal{M}_{\mathcal{A}} = \sup_{\alpha\in\mathcal{A}}\|\alpha\|$, $\mathcal{M}_{\mathcal{B}} = \sup_{b\in\mathcal{B}}\|b\|$ and $\overline{T}(\cdot),\overline{\alpha},\overline{b}$ are defined in Assumption 4.6. Then, $\int\left(\overline{T}^*\right)^2\mathrm{d}G_0 < \infty$, and $\overline{T}^*$ is an envelope function for both $\mathscr{T}$ and $\mathscr{T}_*$.

Lemma C.3. *Let Assumptions 2.1, 4.1 and 4.6 hold. If $\rho\left(\widehat{\theta}_n,\theta_0\right)\xrightarrow{\text{p}}0$, then*

$$\int\left(\widehat{T}_{2,n}(w,x) - T_{2,0}(w,x)\right)^2 G_0(\mathrm{d}w,\mathrm{d}x)\xrightarrow{\text{p}}0 \quad \textit{as } n\to\infty. \tag{65}$$

Together, these two lemmas imply asymptotic negligibility of $R_{3,n}$ in (57).

Lemma C.4. *Let $R_{3,n}$ be as defined in (57). Let ρ be the metric in Definition B.1. Under Assumptions 2.1, 4.1 and 4.6, if $\rho\left(\widehat{\theta}_n,\theta_0\right)\xrightarrow{\text{p}}0$, then $\sqrt{n}\cdot R_{3,n}\xrightarrow{\text{p}}0$.*

Proof of Lemma C.4. Since h_0^2 and $\widehat{h}_n^2$ are probability densities on $\mathcal{B}$, it follows that $T_{2,0},\widehat{T}_{2,n}$ in (53) are members of $\mathscr{T}_*$ in (63). By Lemma C.2, $\mathscr{T}_*$ is a G_0-Donsker class. By Lemma C.3, Assumptions 2.1, 4.1 and 4.6 and $\rho\left(\widehat{\theta}_n,\theta_0\right)\xrightarrow{\text{p}}0$ imply that

$$\int\left(\widehat{T}_{2,n}(w,x) - T_{2,0}(w,x)\right)^2 G_0(\mathrm{d}w,\mathrm{d}x)\xrightarrow{\text{p}}0 \quad \text{as } n\to\infty.$$

Thus by Lemma 19.24 of van der Vaart (1998), $\sqrt{n}\cdot R_{3,n} = \sqrt{n}\cdot\mu_n\left[\widehat{T}_n - T_0\right]\xrightarrow{\text{p}}0$. □

C.3 Asymptotic normality of the leading term

For $\theta = (\alpha,h)\in\mathcal{A}\times\mathcal{H}$, denote

$$\phi(\theta) = \phi(\alpha,h) = \int T_1(\alpha,b)h(b)^2\mathrm{d}b. \tag{66}$$

Then, letting $\widehat{\theta}_n = \left(\widehat{\alpha}_n,\widehat{h}_n\right)$ and $\theta_0 = (\alpha_0,h_0)$, we have

$$R_{1,n} = \phi\left(\widehat{\theta}_n\right) - \phi\left(\theta_0\right). \tag{67}$$

C.3.1 Local parameter spaces and the Fisher norm

Given consistency and convergence rates in Theorems 4.1 and 4.2, we can focus on "local" and "rate-local" versions of the parameter and sieve spaces. For small $\varepsilon > 0$, the "local" spaces are:

$$\begin{aligned}\Theta_\varepsilon &= \left\{\theta\in\Theta:\rho\left(\theta,\theta_0\right)<\varepsilon\right\},\\ \Theta_{n,\varepsilon} &= \Theta_\varepsilon\cap\Theta_{K_n} = \left\{\theta\in\Theta_{K_n}:\rho\left(\theta,\theta_0\right)<\varepsilon\right\}.\end{aligned} \tag{68}$$

Next, let $\zeta_n \geq 1$ be a non-decreasing, slowly growing sequence such that

$$\begin{aligned}
&\zeta_n \nearrow \infty, \\
&\zeta_n^2 \cdot \max\left\{\sqrt{n} \cdot K_n^{-2s/d_X}, \frac{K_n}{\sqrt{n}}\right\} \to 0.
\end{aligned} \tag{69}$$

The "rate-local" spaces are:

$$\begin{aligned}
\mathcal{N}_{0,n} &= \left\{\theta \in \Theta : \rho\left(\theta, \theta_0\right) \leq \max\left\{K_n^{-s/d_X}, \sqrt{\frac{K_n}{n}}\right\} \cdot \zeta_n.\right\}, \\
\mathcal{N}_n &= \mathcal{N}_{0,n} \cap \Theta_{K_n}.
\end{aligned} \tag{70}$$

Then, $\widehat{\theta}_n \in \mathcal{N}_n$ with probability approaching 1. As an example, if $K_n = K_0 n^\delta$ for some $0 < \delta < 1$ chosen according to the conditions of Theorem 4.3, then we can use

$$\zeta_n = \log\log\max\left\{e^e, n\right\}.$$

Remark C.1. Note that ζ_n satisfying these conditions always exists. Indeed, given any positive sequence $\xi_n \to 0$ (not necessarily monotone in n), let $\xi_{*,n} = \sup_{m \geq n} \xi_m$. Then $\xi_{*,n} \downarrow 0$, i.e. $\xi_{*,n}$ is (weakly) monotonically decreasing and limits to zero. We can set $\zeta_n = \xi_{*,n}^{-1/4}$. Then, $\zeta_n \uparrow \infty$ (by $\xi_{*,n} \downarrow 0$) and since $\xi_n \leq \xi_{*,n}$, $\zeta_n^2 \cdot \xi_n \leq \zeta_n^2 \cdot \xi_{*,n} = \sqrt{\xi_{*,n}} \to 0$. With $\xi_n = \max\left\{\sqrt{n} \cdot K_n^{-2s/d_X}, \frac{K_n}{\sqrt{n}}\right\}$, we have a specific instance of ζ_n from the above reasoning.

The likelihood function is pathwise differentiable at $\theta_0 \in \Theta$ and its pathwise derivative is linear in $v = \theta - \theta_0$ for $\theta = (\alpha, h) \in \Theta$. To see this, let $\ell(\cdot) \equiv \log P(\cdot)$ and $v = (v_\alpha, v_h) = (\alpha - \alpha_0, h - h_0)$. Then,

$$\begin{aligned}
\Delta\left(y, w, x; \theta_0\right)[v] &:= \lim_{\varepsilon \downarrow 0} \frac{1}{\varepsilon} \cdot \left(\ell\left(y, w, x; \alpha_0 + \varepsilon v_\alpha, h_0 + \varepsilon v_h\right) - \ell\left(y, w, x; \alpha_0, h_0\right)\right) \\
&= \Delta_\alpha\left(y, w, x; \theta_0\right)' v_\alpha + \Delta_h\left(y, w, x; \theta_0\right)\left[v_h\right],
\end{aligned} \tag{71}$$

where

$$\begin{aligned}
\Delta_\alpha\left(y, w, x; \theta_0\right) &= \frac{\int \frac{\partial}{\partial\alpha}\kappa\left(y, w, x; \alpha_0, b\right) h_0(b)^2 \, \mathrm{d}b}{P\left(y, w, x; \alpha_0, h_0\right)}, \\
\text{and} \qquad \Delta_h\left(y, w, x; \theta_0\right)\left[v_h\right] &= \frac{2\int \kappa\left(y, w, x; \alpha_0, b\right) h_0(b) v_h(b) \mathrm{d}b}{P\left(y, w, x; \alpha_0, h_0\right)}.
\end{aligned} \tag{72}$$

For any y, w, x, $\Delta\left(y, w, x; \theta_0\right)[v]$ is linear in v. The above defines Δ_h as a functional taking a real-valued function v_h as its argument. If instead v_h is vector valued, say $v_h = (v_{h,1}, \ldots, v_{h,d})'$, we set

$$\Delta_h\left(y, w, x; \theta_0\right)\left[v_h\right] = \begin{pmatrix} \Delta_h\left(y, w, x; \theta_0\right)\left[v_{h,1}\right] \\ \vdots \\ \Delta_h\left(y, w, x; \theta_0\right)\left[v_{h,d}\right] \end{pmatrix}.$$

Throughout, we will utilize the fact that the information equality holds in all submodels (or paths) passing through θ_0. Hence, the negative expected Hessian of the log-likelihood is always equal to the expected outer product of the gradient (the expected squared pathwise derivative here). Given that the information equality holds along any path containing θ_0, the Fisher inner product is

$$\langle v_1, v_2 \rangle = \mathbb{E}\left[\Delta\left(Y, W, X; \theta_0\right)[v_1] \cdot \Delta\left(Y, W, X; \theta_0\right)[v_2]\right],$$

with the associated norm $\|\cdot\|$ defined by

$$\|v\|^2 = \langle v, v \rangle = \mathbb{E}\left[\left(\Delta\left(Y, W, X; \theta_0\right)[v]\right)^2\right]. \tag{73}$$

Definition C.1 (Local parameter space and directions)**.** Let $\|\cdot\|$ be the Fisher norm in (73). The space $\mathcal{V}$ is the closed linear span of $\Theta_\varepsilon - \theta_0 = \{\theta - \theta_0 : \theta \in \Theta_\varepsilon\}$ where the closure is computed under the Fisher norm $\|\cdot\|$. The space $\mathcal{V}_n$ is the closed linear span of $\Theta_{n,\varepsilon} - \theta_0 = \{\theta - \theta_0 : \theta \in \Theta_{n,\varepsilon}\}$ where the closure is computed under the Fisher norm $\|\cdot\|$.

Lemma C.5 shows that the $\|\cdot\|$ is weaker than ρ. The proof is deferred to Appendix D.3.

Lemma C.5. *There is a $C_{\|\cdot\|,\rho} \in (0, \infty)$ such that $\|\theta - \theta_0\| \leq C_{\|\cdot\|,\rho} \cdot \rho\left(\theta, \theta_0\right)$.*

Define the best sieve ρ-approximation to θ_0 by

$$\theta_{0,n} \in \underset{\theta \in \Theta_{n,\varepsilon}}{\operatorname{argmin}}\, \rho\left(\theta, \theta_0\right). \tag{74}$$

A necessary condition is that $\theta_{0,n} = (\alpha_0, h_{0,n})$ for some $h_{0,n} \in \widetilde{\mathcal{H}}_{K_n}$ since $\alpha_0 \in \mathcal{A}$. By Assumption 4.4 (ii) and Lemma C.5, it follows that

$$\|\theta_{0,n} - \theta_0\| \leq C_{\|\cdot\|,\rho}\, \rho\left(\theta_{0,n}, \theta_0\right) \leq O\left(K_n^{-s/d_X}\right). \tag{75}$$

C.3.2 Pathwise derivatives, Riesz representers and their estimators

The pathwise derivative of $\phi(\cdot)$ in (66) at θ_0 in the direction $v = (v_\alpha, v_h) \in \mathcal{V}$ is

$$\begin{aligned} \frac{\partial \phi\left(\theta_0\right)}{\partial \theta}[v] &:= \lim_{\varepsilon \downarrow 0} \frac{1}{\varepsilon} \cdot \left(\phi\left(\theta_0 + \varepsilon \cdot v\right) - \phi\left(\theta_0\right)\right) \\ &= \left[\int \frac{\partial}{\partial \alpha} T_1\left(\alpha_0, b\right) h_0(b)^2 \mathrm{d}b\right]' v_\alpha + 2 \int T_1\left(\alpha_0, b\right) h_0(b) v_h(b)\, \mathrm{d}b. \end{aligned} \tag{76}$$

Since $\mathcal{V}_n$ is finite dimensional, by the Riesz Representation Theorem the pathwise derivative $\frac{\partial \phi(\theta_0)}{\partial \theta}[\cdot]$ has a Riesz representer, v_n^*, in the inner product space $(\mathcal{V}_n, \langle \cdot, \cdot \rangle)$. This Riesz representer on $\mathcal{V}_n$ is termed the *sieve Riesz representer* of $\frac{\partial\phi(\theta_0)}{\partial\theta}[\cdot]$ and is defined by the relation

$$\frac{\partial\phi\left(\theta_0\right)}{\partial\theta}[v] = \langle v_n^*, v\rangle \quad \text{for all } v\in\mathcal{V}_n, \tag{77}$$

$$\text{and} \quad \frac{\partial\phi\left(\theta_0\right)}{\partial\theta}\left[v_n^*\right] = \left\|v_n^*\right\|^2 = \sup_{v\in\mathcal{V}_n,\|v\|\neq 0}\frac{\left|\frac{\partial\phi(\theta_0)}{\partial\theta}[v]\right|^2}{\|v\|^2}. \tag{78}$$

The sieve Riesz representer has a closed form expression in terms of a "sieve information matrix". To describe this closed form, let

$$\Phi_n = \left[\begin{array}{c}\int\frac{\partial}{\partial\alpha}T_1\left(\alpha_0,b\right)h_0(b)^2\mathrm{d}b\\ 2\int T_1\left(\alpha_0,b\right)h_0(b)\boldsymbol{\psi}_{K_n}(b)\,\mathrm{d}b\end{array}\right], \tag{79}$$

$$\mathcal{I}_n = \left[\begin{array}{cc}\mathcal{I}_{n,11} & \mathcal{I}_{n,12}\\ \mathcal{I}_{n,21} & \mathcal{I}_{n,22}\end{array}\right] \quad\text{and}\quad \mathcal{I}_n^{-1} = \left[\begin{array}{cc}\mathcal{I}_n^{11} & \mathcal{I}_n^{12}\\ \mathcal{I}_n^{21} & \mathcal{I}_n^{22}\end{array}\right], \tag{80}$$

where with Δ_α and Δ_h defined in (72),

$$\begin{aligned}
\mathcal{I}_{n,11} &= \mathbb{E}\left[\Delta_\alpha\left(Y,W,X;\theta_0\right)\Delta_\alpha\left(Y,W,X;\theta_0\right)'\right]\\
\mathcal{I}_{n,12} &= \mathbb{E}\left[\Delta_\alpha\left(Y,W,X;\theta_0\right)\Delta_h\left(Y,W,X;\theta_0\right)\left[\boldsymbol{\psi}_{K_n}\right]'\right]\\
\mathcal{I}_{n,21} &= \mathcal{I}_{n,12}'\\
\mathcal{I}_{n,22} &= \mathbb{E}\left[\Delta_h\left(Y,W,X;\theta_0\right)\left[\boldsymbol{\psi}_{K_n}\right]\Delta_h\left(Y,W,X;\theta_0\right)\left[\boldsymbol{\psi}_{K_n}\right]'\right].
\end{aligned}$$

For $v\in\mathcal{V}_n$, there are $v_\alpha\in\mathbb{R}^{d_W}$ and $\lambda_h\in\mathbb{R}^{K_n}$ such that $v=(v_\alpha,v_h(\cdot))=(v_\alpha,\boldsymbol{\psi}_{K_n}(\cdot)'\lambda_h)$. Let $\lambda'=(v_\alpha',\lambda_h')$. Then,

$$\|v\|^2 = \lambda'\mathcal{I}_n\lambda. \tag{81}$$

Using equation (32) of Chen et al. (2014), the sieve Riesz representer, v_n^*, is

$$v_n^* = \left[\begin{array}{c}v_{n,\alpha}^*\\ v_{n,h}^*(\cdot)\end{array}\right] = \left[\begin{array}{c}v_{n,\alpha}^*\\ \boldsymbol{\psi}_{K_n}(\cdot)'\boldsymbol{\lambda}_{n,h}\end{array}\right], \quad\text{where}\quad \boldsymbol{\lambda}_n = \left[\begin{array}{c}v_{n,\alpha}^*\\ \boldsymbol{\lambda}_{n,h}\end{array}\right] = \mathcal{I}_n^{-1}\Phi_n. \tag{82}$$

In the above, $v_{n,\alpha}^*$ is comprised of the first d_W components of $\mathcal{I}_n^{-1}\Phi_n$. It is straightforward to show that $\|v_n^*\|$ is a non-decreasing sequence. The results of Chen et al. (2014) and Chen and Liao (2014) show that the behavior of Δ and the sieve Riesz representer essentially pin down the limiting behavior of $\widehat{\phi}_n$ in terms of asymptotic distribution. In particular, under regularity conditions we will verify later on, the following holds:

$$\frac{\sqrt{n}\cdot\left[\phi\left(\widehat{\theta}_n\right)-\phi\left(\theta_0\right)\right]}{\|v_n^*\|} = \frac{1}{\sqrt{n}}\sum_{i=1}^n\Delta\left(Y_i,W_i,X_i;\theta_0\right)\left[\frac{v_n^*}{\|v_n^*\|}\right]+o_{\mathrm{p}}(1)$$

The first term on the right has unit second moment for all n and hence by Chebychev's inequality, the rate of convergence of $\phi\left(\widehat{\theta}_n\right)$ to $\phi\left(\theta_0\right)$ is $O_{\mathrm{p}}\left(\|v_n^*\|/\sqrt{n}\right)$. Remark C.2 below expands on two relevant cases for the rate. In addition, if a Lindeberg or Lyapunov condition holds for the first term on the right above (the summands can be shown to have mean zero), we get

$$\frac{\sqrt{n}\cdot\left[\phi\left(\widehat{\theta}_n\right)-\phi\left(\theta_0\right)\right]}{\|v_n^*\|}=\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\Delta\left(Y_i,W_i,X_i;\theta_0\right)\left[\frac{v_n^*}{\|v_n^*\|}\right]+o_{\mathrm{p}}(1)\xrightarrow{\mathrm{d}}\mathcal{N}(0,1). \tag{83}$$

Thus, $\|v_n^*\|^2$ corresponds to the “asymptotic variance” of $\sqrt{n}\cdot\left[\phi\left(\widehat{\theta}_n\right)-\phi\left(\theta_0\right)\right]$.

Remark C.2 (The role of the sieve Riesz representer in determining the rate of convergence)**.** When $\frac{\partial\phi(\theta_0)}{\partial\theta}[\cdot]$ is a bounded functional (in the sense of operator norm) on the Hilbert space $(\mathcal{V},\langle\cdot,\cdot\rangle)$, the rate of convergence of $\phi\left(\widehat{\theta}_n\right)$ to $\phi\left(\theta_0\right)$ is the parametric $1/\sqrt{n}$ rate. That is, $\phi\left(\theta_0\right)$ is *regularly estimable*. This boundedness occurs if and only if $\limsup_{n\to\infty}\|v_n^*\|<\infty$. Otherwise, if $\frac{\partial\phi(\theta_0)}{\partial\theta}[\cdot]$ is an unbounded functional (again in the sense of operator norm) on $(\mathcal{V},\langle\cdot,\cdot\rangle)$, then the rate of convergence of $\phi\left(\widehat{\theta}_n\right)$ to $\phi\left(\theta_0\right)$ is slower than $1/\sqrt{n}$. In this case, $\limsup_{n\to\infty}\|v_n^*\|=\infty$. For details, see Lemma 3.3 of Chen and Pouzo (2015).

There are empirical counterparts of all of the objects in (79), (80) and (82). Replacing population objects with sample analogues, we can first define the empirical counterparts of the Fisher inner product and norm by

$$\begin{aligned}\langle v_1,v_2\rangle_n&=\frac{1}{n}\sum_{i=1}^{n}\Delta\left(Y_i,W_i,X_i;\widehat{\theta}_n\right)[v_1]\cdot\Delta\left(Y_i,W_i,X_i;\widehat{\theta}_n\right)[v_2],\\ \|v\|_n^2=\langle v,v\rangle_n&=\frac{1}{n}\sum_{i=1}^{n}\left(\Delta\left(Y_i,W_i,X_i;\widehat{\theta}_n\right)[v]\right)^2.\end{aligned} \tag{84}$$

Next, we can define the following:

$$\begin{aligned}\widehat{\Phi}_n&=\left[\begin{array}{c}\int\frac{\partial}{\partial\alpha}T_1\left(\widehat{\alpha}_n,b\right)\widehat{h}_n(b)^2\mathrm{d}b\\ \int T_1\left(\widehat{\alpha}_n,b\right)\widehat{h}_n(b)\boldsymbol{\psi}_{K_n}(b)\,\mathrm{d}b\end{array}\right],\\ \widehat{\mathcal{I}}_n&=\left[\begin{array}{cc}\widehat{\mathcal{I}}_{n,11}&\widehat{\mathcal{I}}_{n,12}\\ \widehat{\mathcal{I}}_{n,21}&\widehat{\mathcal{I}}_{n,22}\end{array}\right]\end{aligned}$$

where,

$$\begin{aligned}
\widehat{\mathcal{I}}_{n,11} &= \frac{1}{n}\sum_{i=1}^{n} \Delta_\alpha\left(Y_i, W_i, X_i; \widehat{\theta}_0\right) \Delta_\alpha\left(Y_i, W_i, X_i; \widehat{\theta}_n\right)' \\
\widehat{\mathcal{I}}_{n,12} &= \frac{1}{n}\sum_{i=1}^{n} \Delta_\alpha\left(Y_i, W_i, X_i; \widehat{\theta}_n\right) \Delta_h\left(Y_i, W_i, X_i; \widehat{\theta}_n\right) [\boldsymbol{\psi}_{K_n}]' \\
\widehat{\mathcal{I}}_{n,21} &= \widehat{\mathcal{I}}'_{n,12} \\
\widehat{\mathcal{I}}_{n,22} &= \frac{1}{n}\sum_{i=1}^{n} \Delta_h\left(Y_i, W_i, X_i; \widehat{\theta}_n\right) [\boldsymbol{\psi}_{K_n}] \, \Delta_h\left(Y_i, W_i, X_i; \widehat{\theta}_n\right) [\boldsymbol{\psi}_{K_n}]' .
\end{aligned}$$

By analogy, for $v \in \mathcal{V}_n$, such that $v = (v_\alpha, v_h(\cdot)) = (v_\alpha, \boldsymbol{\psi}_{K_n}(\cdot)'\lambda_h)$, letting $\lambda' = (v'_\alpha, \lambda'_h)$, we get

$$\|v\|_n^2 = \lambda' \widehat{\mathcal{I}}_n \lambda.$$

The empirical counterpart to the sieve Riesz representer v_n^* is $\widehat{v}_n^*$ where

$$\widehat{v}_n^* = \begin{bmatrix} \widehat{v}_{n,\alpha}^* \\ \widehat{v}_{n,h}^*(\cdot) \end{bmatrix} = \begin{bmatrix} \widehat{v}_{n,\alpha}^* \\ \boldsymbol{\psi}_{K_n}(\cdot)'\widehat{\boldsymbol{\lambda}}_{n,h} \end{bmatrix}, \quad \text{where} \quad \widehat{\boldsymbol{\lambda}}_n = \begin{bmatrix} \widehat{v}_{n,\alpha}^* \\ \widehat{\boldsymbol{\lambda}}_{n,h} \end{bmatrix} = \widehat{\mathcal{I}}_n^{-1}\widehat{\Phi}_n. \tag{85}$$

Remark C.3 (On the factor $\widehat{\boldsymbol{\lambda}}_n$)**.** The $\widehat{\boldsymbol{\lambda}}_n$ defined in (85) and in (19) are the same object. Hence, the same notation is used.

C.3.3 Results for asymptotic normality of the leading term

The natural estimator for the asymptotic variance term $\|v_n^*\|$ in (83) is $\|\widehat{v}_n^*\|_n$, where from (84),

$$\|\widehat{v}_n^*\|_n^2 = \frac{1}{n}\sum_{i=1}^{n}\left(\Delta\left(Y_i, W_i, X_i; \widehat{\theta}_n\right)[\widehat{v}_n^*]\right)^2. \tag{86}$$

We might expect that $\|\widehat{v}_n^*\|_n$ is a consistent estimator of $\|v_n^*\|$. An additional pair of assumptions is required for both this consistent variance estimation claim and asymptotic normality of $\phi(\theta_n) - \phi(\theta_0)$ as in (83) to be true. To that end, let the pathwise second derivative be defined:

$$r(y, w, x; \theta_0)[v_1, v_2] := \frac{\partial \Delta(y, w, x; \theta_0 + \varepsilon \cdot v_2)}{\partial \tau}[v_1]\bigg|_{\varepsilon=0}. \tag{87}$$

Assumption C.1. The following hold:

(i) There is a $c_{\mathcal{I}} \in (0, \infty)$ such that for all n, the smallest eigenvalue of $\mathcal{I}_n$ is greater than or equal to $c_{\mathcal{I}}$.

(ii) The following holds:

$$\sup_{\theta\in\mathcal{N}_n,\widetilde{\theta}\in\mathcal{N}_{0,n}}\|\theta-\theta_0\|\sup_{v\in\mathcal{N}_{0,n}:\|v\|=1}\mathbb{E}\left[\left|r\left(Y,W,X;\widetilde{\theta}\right)[v,v]-r\left(Y,W,X;\theta_0\right)[v,v]\right|\right]=o\left(\frac{1}{n}\right).$$

Remark C.4. It can be shown that Lemma C.5 implies that that the largest eigenvalue of $\mathcal{I}_n$ is bounded above uniformly in $n\in\mathbb{N}$ by the constant $C_{\|\cdot\|,\rho}\in(0,\infty)$ defined in Lemma C.5. Hence, Assumption C.1 (i) restricts the smallest eigenvalue. The main consequence of this assumption is that the norm $\|\cdot\|$ and the metric ρ become strongly topologically equivalent.

For parity with the notation used in the proof of Theorem 4.3, define

$$\begin{aligned}\dot{\phi}_n(y,w,x)&=\Delta\left(y,w,x;\theta_0\right)\left[\frac{v_n^*}{\|v_n^*\|}\right],\\ V_{\phi,n}&=\|v_n^*\|,\\ \widehat{V}_{\phi,n}&=\|\widehat{v}_n^*\|_n.\end{aligned}\tag{88}$$

Theorem C.1. *Suppose Assumptions 2.1, 2.2, 4.1-4.7 and C.1 all hold. Then, $R_{1,n}$ in (55) satisfies*

$$\frac{\sqrt{n}R_{1,n}}{V_{\phi,n}}=\frac{\sqrt{n}\cdot\left[\phi\left(\widehat{\theta}_n\right)-\phi\left(\theta_0\right)\right]}{V_{\phi,n}}=\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\dot{\phi}_n\left(Y_i,W_i,X_i\right)+o_{\mathrm{p}}(1)\xrightarrow{\mathrm{d}}\mathcal{N}(0,1).\tag{89}$$

Furthermore, $V_{\phi,n}$ is non-decreasing and there exists $V_\phi\in(0,\infty)$ such that

$$\lim_{n\to\infty}V_{\phi,n}=V_\phi.\tag{90}$$

In addition

$$\frac{\widehat{V}_{\phi,n}}{V_{\phi,n}}\xrightarrow{\mathrm{p}}1,\tag{91}$$

so that by Slutsky's Theorem,

$$\frac{\sqrt{n}R_{1,n}}{\widehat{V}_{\phi,n}}\xrightarrow{\mathrm{d}}\mathcal{N}(0,1).\tag{92}$$

Proof of Theorem C.1. Verification of Assumption 2.1 of Chen and Liao (2014) is the content of Lemma C.6 (see also Remark C.5). Verification of Assumption 2.2 of Chen and Liao (2014) Lemma C.7 (see also Remark C.6). Verification of Assumption 2.3 of Chen and Liao (2014) is the content of Lemma C.8. Therefore, (89) follows as a direct consequence of Lemma 2.1 of Chen and Liao (2014). Finally, Assumption 3.3 of Chen and Liao (2014) is the content of Lemma C.9. Therefore, (91) holds by Corollary 3.3 (2) of Chen and Liao (2014). (92) follows from (89), (91) and Slutsky's Theorem. The remaining result (90) is the content of Lemma C.11. □

Lemma C.6. *Suppose Assumptions 2.1, 2.2, 4.1-4.7 all hold. Then* *(i) The following condition holds:*

$$\frac{\sup_{\theta\in\mathcal{N}_n}\left|\phi(\theta)-\phi\left(\theta_0\right)-\frac{\partial\phi(\theta_0)}{\partial\theta}\left[\theta-\theta_0\right]\right|}{\|v_n^*\|}=o\left(n^{-1/2}\right). \tag{93}$$

(ii) The following two (mutually exclusive) conditions both hold:

$$\limsup_{n\to\infty}\|v_n^*\|=\infty \quad \textit{and} \quad \sqrt{n}\cdot\frac{\left|\frac{\partial\phi(\theta_0)}{\partial\theta}\left[\theta-\theta_0\right]\right|}{\|v_n^*\|}=o(1), \tag{94}$$

$$\textit{or} \ \limsup_{n\to\infty}\|v_n^*\|<\infty \quad \textit{and} \quad \sqrt{n}\cdot\|v^*-v_n^*\|\cdot\|\theta_{0,n}-\theta_0\|=o(1). \tag{95}$$

Proof of Lemma C.6. See Appendix D.4.1. □

Remark C.5. Assumption 2.1 of Chen and Liao (2014) has three parts. The conditions verified in Lemma C.6 correspond to parts (ii) and (iii) of that assumption. Part (i) of Assumption 2.1 in Chen and Liao (2014) has two conditions. First, $\frac{\partial\phi(\theta_0)}{\partial\theta}[v]$ is required to be a linear functional, which is immediate from its definition in (76). The second condition is that the norm $\|\cdot\|$ satisfies $\|v_n^*\| \,/\, \|v_n^*\|_{\mathrm{sd}}=O(1)$ where $\|\cdot\|_{\mathrm{sd}}$ is a "standard deviation norm" (see Chen and Liao (2014) for the definition of this norm). In the case of maximum likelihood, $\|\cdot\|_{\mathrm{sd}}$ is exactly equal to the Fisher norm, which we have defined $\|\cdot\|$ to be. Hence, this second condition is satisfied by definition.

Lemma C.7. *Suppose Assumptions 2.1, 2.2, 4.1-4.7 and C.1 all hold. Then the following also hold:*

(i) The functional $v\mapsto\mu_n\left\{\Delta\left(\cdot;\theta_0\right)[v]\right\}$ is linear in $v\in\mathcal{V}$.

(ii) $\sup_{\theta\in\mathcal{N}_n}\left|\mu_n\left\{\Delta(\cdot;\theta)\left[\frac{v_n^*}{\|v_n^*\|}\right]-\Delta\left(\cdot;\theta_0\right)\left[\frac{v_n^*}{\|v_n^*\|}\right]\right\}\right|=o_{\mathrm{p}}\left(n^{-1/2}\right).$

(iii) The following holds:

$$\sup_{\theta\in\mathcal{N}_n}\left|\mathbb{E}\left[\ell\left(Y,W,X;\theta_0\right)-\ell(Y,W,X;\theta)\right]-\frac{\|\theta-\theta_0\|^2}{2}\right|=o\left(n^{-1}\right)$$

Proof of Lemma C.7. See Appendix D.4.3. □

Remark C.6. In Lemma C.7 conditions (ii) and (iii) correspond to Assumption 2.2 (ii)′ and 2.2 (iii)′′′ of Chen and Liao (2014) respectively, each of which are sufficient conditions for the original Assumption 2.2 (ii) and (iii) respectively.

Lemma C.8. *Suppose Assumptions 2.1, 2.2, 4.1-4.7 and C.1 all hold. Then,*

$$\sqrt{n}\mu_n\left\{\Delta\left(\cdot;\theta_0\right)\left[\frac{v_n^*}{\|v_n^*\|}\right]\right\}\xrightarrow{\mathrm{d}}\mathcal{N}(0,1). \tag{96}$$

Proof of Lemma C.8. See Appendix D.4.4. □

Lemma C.9. *Suppose Assumptions 2.1, 2.2, 4.1-4.7 and C.1 all hold. Then the following hold*

(i)

$$\sup_{\theta\in\mathcal{N}_n, v_1, v_2\in\mathcal{V}_n:\|v_1\|=\|v_2\|=1} \left|\mathbb{E}\left[\Delta(Y,W,X;\theta)\left[v_1\right]\Delta(Y,W,X;\theta)\left[v_2\right]\right.\right.$$
$$\left.\left. -\Delta\left(Y,W,X;\theta_0\right)\left[v_1\right]\Delta\left(Y,W,X;\theta_0\right)\left[v_2\right]\right]\right| = o(1).$$

(ii)

$$\sup_{\theta\in\mathcal{N}_n, v_1, v_2\in\mathcal{V}_n:\|v_1\|=\|v_2\|=1} \left|\mu_n\left\{\Delta(Y,W,X;\theta)\left[v_1\right]\Delta(Y,W,X;\theta)\left[v_2\right]\right\}\right| = o_{\mathrm{p}}(1).$$

Lemma C.10 (Proof of Lemma C.9)**.** *See Appendix D.4.5.*

Lemma C.11. *Suppose Assumptions 2.1, Assumption 4.6, Assumption 4.7 and Assumption C.1 (i) all hold. Then there is $V_\phi \in (0,\infty)$ such that $\lim_{n\to\infty} V_{\phi,n} = V_\phi$.*

Proof of Lemma C.11. See Appendix D.4.6. □

C.4 Consistent variance estimation

Theorem C.2. *Suppose Assumptions 2.1, 2.2, 4.1-4.7 and C.1 all hold. Then $\widehat{V}_{\tau,n}$ in (18) satisfies:*

$$\frac{\widehat{V}_{\tau,n}}{V_{\tau,n}} \xrightarrow{\mathrm{p}} 1. \tag{97}$$

Proof of Theorem C.2. Write $\widehat{V}_{\tau,n}$ as

$$\widehat{V}_{\tau,n} = \frac{1}{n}\sum_{i=1}^{n}\left\{\Delta\left(Y_i, W_i, X_i; \widehat{\theta}_n\right)\left[\widehat{v}_n^*\right] + \widetilde{T}_{2,n}\left(W_i, X_i\right)\right\}^2,$$

where

$$\widetilde{T}_{2,n}(w,x) = \widehat{T}_{2,n}(w,x) - \frac{1}{n}\sum_{i=1}^{n}\widehat{T}_{2,n}\left(W_i, X_i\right).$$

Next, write

$$V_{\tau,n} = V_{\phi,n} + \mathrm{Var}\left[T_{2,0}(W,X)\right] + 2\mathrm{Cov}\left[\dot{\phi}_n(Y,W,X), T_{2,0}(W,X)\right].$$

By Lemma C.12 below, $\mathbb{E}\left[\dot{\phi}_n\left(Y,W,X\right)\middle| W,X\right] = 0$ and so,

$$\mathrm{Cov}\left[\dot{\phi}_n(Y,W,X), T_{2,0}(W,X)\right] = 0.$$

Therefore,

$$V_{\tau,n} = V_{\phi,n} + \mathbb{V}_{2,0},$$
$$\text{where } \mathbb{V}_{2,0} = \mathrm{Var}\left[T_{2,0}(W,X)\right].$$

As before, set

$$\widehat{V}_{\phi,n} = \frac{1}{n}\sum_{i=1}^{n}\left(\Delta\left(Y_i, W_i, X_i; \widehat{\theta}_n\right)[\widehat{v}_n^*]\right)^2$$

and for ease of notation, define

$$\widehat{\mathbb{V}}_{2,n} = \frac{1}{n}\sum_{i=1}^{n}\widetilde{T}_{2,n}\left(W_i, X_i\right)^2 = \frac{1}{n}\sum_{i=1}^{n}\widehat{T}_{2,n}\left(W_i, X_i\right)^2 - \left[\frac{1}{n}\sum_{j=1}^{n}\widehat{T}_{2,n}\left(W_i, X_i\right)\right]^2, \tag{98}$$

and finally,

$$\widehat{\mathbb{C}}_n = \frac{1}{n}\sum_{i=1}^{n}\Delta\left(Y_i, W_i, X_i; \widehat{\theta}_n\right)[\widehat{v}_n^*]\,\widetilde{T}_{2,n}\left(W_i, X_i\right). \tag{99}$$

Therefore,

$$\widehat{V}_{\tau,n} = \widehat{V}_{\phi,n} + \widehat{\mathbb{V}}_{2,n} + 2\widehat{\mathbb{C}}_n.$$

By (91) of Theorem C.1, $\widehat{V}_{\phi,n}/V_{\phi,n} \xrightarrow{\mathrm{p}} 1$. By Lemma C.13, $\widehat{\mathbb{V}}_{2,n} \xrightarrow{\mathrm{p}} \mathbb{V}_{2,0}$, and so $\widehat{\mathbb{V}}_{2,n}/\mathbb{V}_{2,0} \xrightarrow{\mathrm{p}} 1$. By Lemma C.14, $\widehat{\mathbb{C}}_n/\sqrt{V_{\phi,n}} \xrightarrow{\mathrm{p}} 0$. Therefore,

$$\begin{aligned}\frac{\widehat{V}_{\tau,n}}{V_{\tau,n}} &= \frac{\widehat{V}_{\phi,n} + \widehat{V}_{2,n} + 2\mathbb{C}_n}{V_{\phi,n} + \mathbb{V}_{2,0}} \\ &= \frac{V_{\phi,n}\cdot(1+o_{\mathrm{p}}(1)) + \mathbb{V}_{2,0}\cdot(1+o_{\mathrm{p}}(1)) + \sqrt{V_{\phi,n}}\cdot o_{\mathrm{p}}(1)}{V_{\phi,n} + \mathbb{V}_{2,0}} = 1 + o_{\mathrm{p}}(1).\end{aligned}$$

□

Lemma C.12. *Given any $\theta \in \Theta$ and $v \in \mathcal{V}$, for all w, x*

$$\sum_{y=0}^{J} P(y, w, x; \theta)\cdot\Delta(y, w, x; \theta)[v] = 0. \tag{100}$$

As a consequence, given any function $T(w, x)$ that depends only on (w, x),

$$\nu_0\left[\Delta\left(\cdot;\theta_0\right)[v]\cdot T(\cdot)\right] = 0 \quad \textit{for any } v \in \mathcal{V}. \tag{101}$$

Proof of Lemma C.12. For any θ, and any w, x, $\sum_{y=0}^{J} P(y, w, x; \theta) = 1$. This implies that along any path, $\sum_{y=0}^{J}(\partial/\partial\varepsilon)P\left(y, w, x; \theta + \varepsilon v\right)|_{\varepsilon=0} = 0$. Hence,

$$\begin{aligned}\sum_{y=0}^{J} P(y, w, x; \theta)\cdot\Delta(y, w, x; \theta)[v] &= \sum_{y=0}^{J} P(y, w, x; \theta)\cdot\frac{(\partial/\partial\varepsilon)P\left(y, w, x; \theta + \varepsilon v\right)|_{\varepsilon=0}}{P(y, w, x; \theta)} \\ &= \sum_{y=0}^{J}(\partial/\partial\varepsilon)P\left(y, w, x; \theta + \varepsilon v\right)|_{\varepsilon=0} \\ &= 0.\end{aligned}$$

This proves (100). Next, applying (100) with $\theta = \theta_0$, given any function $h(w,x)$ that depends only on (w,x),

$$\nu_0\left[\Delta\left(\cdot;\theta_0\right)[v]\cdot T(\cdot)\right] = \int \underbrace{\left[\sum_{y=0}^{J} P\left(y,w,x;\theta_0\right)\cdot\Delta\left(y,w,x;\theta_0\right)[v]\right]}_{=0}\cdot T(w,x)G_0(\mathrm{d}w,\mathrm{d}x) = 0.$$

□

Lemma C.13. *Let Assumptions 2.1, 4.1 and 4.6 hold. If* $\rho\left(\widehat{\theta}_n,\theta_0\right)\xrightarrow{\mathrm{p}}0$*, then* $\widehat{\mathbb{V}}_{2,n}$ *in* (98) *satisfies* $\widehat{\mathbb{V}}_{2,n}\xrightarrow{\mathrm{p}}\mathbb{V}_{2,0}$.

Proof of Lemma C.13. See Appendix D.5.1. □

Lemma C.14. *Suppose Assumptions 2.1, 2.2, 4.1-4.7 and C.1 all hold. Then,* $\widehat{\mathbb{C}}_n$ *in* (99) *satisfies* $\widehat{\mathbb{C}}_n/\sqrt{V_{\phi,n}}\xrightarrow{\mathrm{p}}0$.

Proof of Lemma C.14. See Appendix D.5.2. □

D Proofs of additional results used for asymptotic normality and consistent variance estimation

D.1 Proof of Lemma C.1

Proof of Lemma C.1. To see (54), write $\widehat{\tau}_n = \nu_n\left[\widehat{T}_{2,n}\right]$ and $\tau_0 = \nu_0\left[T_{2,0}\right]$. Hence

$$\begin{aligned}
\widehat{\tau}_n - \tau_0 &= \nu_n\left[\widehat{T}_{2,n}\right] - \nu_0\left[T_{2,0}\right] \\
&= \nu_0\left[\widehat{T}_{2,n}\right] - \nu_0\left[T_{2,0}\right] + (\nu_n-\nu_0)\left[\widehat{T}_{2,n}\right] \\
&= \nu_0\left[\widehat{T}_{2,n}\right] - \nu_0\left[T_{2,0}\right] + (\nu_n-\nu_0)\left[T_{2,0}\right] + (\nu_n-\nu_0)\left[\widehat{T}_{2,n}-T_{2,0}\right] \\
&= \nu_0\left[\widehat{T}_{2,n}\right] - \nu_0\left[T_{2,0}\right] + \mu_n\left[T_{2,0}\right] + \mu_n\left[\widehat{T}_{2,n}-T_{2,0}\right].
\end{aligned}$$

Using (53) and Fubini's Theorem,

$$\begin{aligned}
\nu_0\left[\widehat{T}_{2,n}\right] &= \int \widehat{T}_{2,n}(w,x)G_0(\mathrm{d}w,\mathrm{d}x) \\
&= \int\left\{\int t\left(w,x;\widehat{\alpha}_n,b\right)\widehat{h}_n(b)^2\,\mathrm{d}b\right\}G_0(\mathrm{d}w,\mathrm{d}x) \\
&= \int\left\{\int t\left(w,x;\widehat{\alpha}_n,b\right)G_0(\mathrm{d}w,\mathrm{d}x)\right\}\widehat{h}_n(b)^2\,\mathrm{d}b \\
&= \int T_1\left(\widehat{\alpha}_n,b\right)\widehat{h}_n(b)^2\,\mathrm{d}b.
\end{aligned}$$

Similarly,

$$\begin{aligned}\nu_0\left[T_{2,0}\right] &= \int T_{2,0}(w,x)G_0(\mathrm{d}w,\mathrm{d}x)\\ &= \int\left\{\int t\left(w,x;\alpha_0,b\right)h_0(b)^2\,\mathrm{d}b\right\}G_0(\mathrm{d}w,\mathrm{d}x)\\ &= \int\left\{\int t\left(w,x;\alpha_0,b\right)G_0(\mathrm{d}w,\mathrm{d}x)\right\}h_0(b)^2\,\mathrm{d}b\\ &= \int T_1\left(\alpha_0,b\right)h_0(b)^2\,\mathrm{d}b.\end{aligned}$$

Therefore,

$$\begin{aligned}\widehat{\tau}_n-\tau_0 &= \int T_1\left(\widehat{\alpha}_n,b\right)\widehat{h}_n(b)^2\,\mathrm{d}b-\int T_1\left(\alpha_0,b\right)h_0(b)^2\,\mathrm{d}b+\mu_n\left[T_{2,0}\right]+\mu_n\left[\widehat{T}_{2,n}-T_{2,0}\right]\\ &= R_{1,n}+R_{2,n}+R_{3,n}.\end{aligned}$$

Hence, (54) follows. □

D.2 Proofs of Lemmas C.2 and C.3

Proof of Lemma C.2. Square integrability of $\overline{T}^*$ is immediate from Assumption 4.6. We start with the proof that $\overline{T}^*$ is indeed an envelope function for both $\mathscr{T}$ and $\mathscr{T}_*$ and then move onto proving that both families are Donsker classes. Under Assumption 4.6, and given any α, b,

$$\begin{aligned}|t(w,x,\alpha,b)| &\leq |t(w,x,\overline{\alpha},\overline{b})|+|t(w,x,\alpha,b)-t(w,x,\overline{\alpha},\overline{b})|\\ &\leq |t(w,x,\overline{\alpha},\overline{b})|+\overline{T}(w,x)\left(\|\alpha-\overline{\alpha}\|_2^2+\|b-\overline{b}\|_2^2\right)^{1/2}\\ &\leq \overline{T}^*(w,x).\end{aligned}$$

The above display shows that $\overline{T}^*$ is an envelope function for $\mathscr{T}$ since w,x,α,b are arbitrary. For fixed w,x,α, integrate the far left with respect to any probability measure $F\in\mathrm{PM}(\mathcal{B})$ to see that $\overline{T}^*$ is also an envelope function for $\mathscr{T}_*$.

Under Assumptions 2.1, 4.1 and 4.6, $\mathscr{T}$ is a G_0-Donsker class — see Example 19.7 in van der Vaart (1998). By Theorems 2.10.1 and 2.10.3 of van der Vaart and Wellner (1996), the convex hull of $\mathscr{T}$,

$$\operatorname{conv}(\mathscr{T})=\left\{\sum_{k=1}^m\omega_k t\left(\cdot;\alpha,b_k\right):\sum_{j=1}^m\omega_j=1,\omega_k\geq 0,b_k\in\mathcal{B}\text{ for each }k=1,\ldots,m,m\in\mathbb{N},\alpha\in\mathcal{A}\right\},$$

is also G_0-Donsker. $\mathcal{B}$ is a metric space, and hence (trivially) metrizable. By Theorem 15.10 of Aliprantis and Border (2006), under the topology of weak convergence, the set of finitely supported discrete probability measures on $\mathcal{B}$ is dense in the set of all probability measures on $\mathcal{B}$, i.e. $\mathrm{PM}(\mathcal{B})$. Thus, given any $F\in\mathrm{PM}(\mathcal{B})$, there is a sequence of finitely supported discrete probability measures,

F_m, such that $F_m \rightsquigarrow F$, where $\rightsquigarrow$ denotes weak convergence. Fix $\alpha \in \mathcal{A}$ and denote:

$$t_{*,m}(w,x) = \int t(w,x;\alpha,b)F_m(\mathrm{d}b)$$
$$t_*(w,x) = \int t(w,x;\alpha,b)F(\mathrm{d}b)$$

Then $t_* \in \mathscr{T}_*$ and $t_{*,m} \in \mathrm{conv}(\mathscr{T})$. Under Assumption 4.6, $t(w,x;\alpha,b)$ is a continuous function of b for any w,x and α. By Assumption 2.1, $\mathcal{B}$ is compact and so, $t(w,x;\alpha,b)$ is a bounded function in b. By the Portmanteau Lemma, for any w,x,

$$\lim_{m\to\infty} t_{*,m}(w,x) = \lim_{m\to\infty} \int t(w,x;\alpha,b)\ F_m(\mathrm{d}b) = \int t(w,x;\alpha,b)\ F(\mathrm{d}b) = t_*(w,x).$$

Thus, given any $t_* \in \mathscr{T}_*$, there exists a sequence $t_{*,m}$ in $\mathrm{conv}(\mathscr{T})$ such that $t_{*,m} \to t_*$ pointwise in w,x.

The sequence $t_{*,m}(w,x)$ also converges to $t_*(w,x)$ in $\mathscr{L}_2\left(G_0\right)$. To see this, we have $|t_{*,m}| \leq \left|\overline{T}^*\right|$ and the upper bounding function has already been argued to be square integrable with respect to G_0. By the Dominated Convergence Theorem, our previous pointwise convergence (in w,x) arguments then imply that

$$\lim_{m\to\infty} \int \left(t_{*,m}(w,x) - t_*(w,x)\right)^2 G_0(\mathrm{d}w,\mathrm{d}x) = 0.$$

Therefore, $t_{*,m} \to t_*$ pointwise and in $\mathscr{L}_2\left(G_0\right)$ as functions of w,x. Since the claim was established for arbitrary $\alpha \in \mathcal{A}$, by Theorem 2.10.2 of van der Vaart and Wellner (1996), $\mathscr{T}_*$ is a G_0-Donsker class. □

Proof of Lemma C.3. Since $\rho\left(\widehat{\theta}_n, \theta_0\right) \xrightarrow{\mathrm{p}} 0$, given any subsequence $\{n_k\}$, there is a further subsequence $\{n_{k_m}\}$ such that[15]

$$\rho\left(\widehat{\theta}_{n_{k_m}}, \theta_0\right) \xrightarrow{\text{a.s.}} 0 \quad \text{as } m \to \infty. \tag{102}$$

Thus, there is an event $\mathcal{Z}\left(\{n_{k_m}\}\right)$ of Pr-probability[16] 1 such that on $\mathcal{Z}\left(\{n_{k_m}\}\right)$,

$$\lim_{m\to\infty} \rho_{\mathscr{L}_2}\left(\left|\widehat{h}_{n_{k_m}}\right|, h_0\right) = 0. \tag{103}$$

Thus, on the event $\mathcal{Z}\left(\{n_{k_m}\}\right)$ the density $\widehat{h}^2_{n_{k_m}}$ converges to h_0^2 in the Hellinger topology (the topology of $\mathscr{L}_2$ convergence of root densities). The Hellinger topology is stronger than the Total Variation topology (equation (8) of Gibbs and Su (2002, p. 428)) and the Total Variation topology is stronger than the weak topology (or the Lévy-Prohorov topology, see Gibbs and Su (2002, p. 428) or equation (2.24) of Huber and Ronchetti (2009, p. 36)). Thus, the probability measure on $\mathcal{B}$ associated with $\widehat{h}^2_{n_{k_m}}$ converges weakly to the probability measure associated with h_0^2. Let $\widehat{H}_{n_{k_m}}, H_0$

[15] Convergence in probability is equivalent to every subsequence having an almost surely convergent subsubsequence. See for example Theorem 2.3.2 of Durrett (2019).

[16] To remind the reader, Pr is the underlying probability measure against which all probability and random variables are defined.

be the probability measures on $\mathcal{A} \times \mathcal{B}$ defined by

$$\widehat{H}_{n_{k_m}}(A \times B) = \mathbb{I}\left\{\widehat{\alpha}_{n_{k_m}} \in A\right\} \times \int_B \widehat{h}_{n_{k_m}}(b)^2 \, \mathrm{d}b,$$
$$H_0(A \times B) = \mathbb{I}\left\{\alpha_0 \in A\right\} \times \int_B h_0(b)^2 \, \mathrm{d}b.$$

Combining Euclidean (almost sure) convergence of $\widehat{\alpha}_{n_{k_m}}$ to α_0 implied by (102) with weak convergence of (the measures associated with) $\widehat{h}^2_{n_{k_m}}$ to h_0^2 implied by (103), it follows that $\widehat{H}_{n_{k_m}}$ converges weakly to H_0. Note that we can write

$$\widehat{T}_{2,n_{k_m}}(w,x) = \int t(w,x;\alpha,b) \; \widehat{H}_{k_m}(\mathrm{d}\alpha, \mathrm{d}b),$$
$$T_{2,0}(w,x) = \int t(w,x;\alpha,b) \; H_0(\mathrm{d}\alpha, \mathrm{d}b).$$

Under Assumption 4.6, $t(w,x;\alpha,b)$ is a continuous function of b for any w, x and α. By Assumptions 2.1 and 4.1, $\mathcal{A}$ and $\mathcal{B}$ are both compact (so that $\mathcal{A} \times \mathcal{B}$ is compact in the Euclidean norm) and so, $t(w,x;\alpha,b)$ is a bounded function in b. By the Portmanteau Lemma, for any w, x, on the event $\mathcal{Z}\left(\{n_{k_m}\}\right)$,

$$\lim_{m\to\infty} \widehat{T}_{2,n_{k_m}}(w,x) = \lim_{m\to\infty} \int t(w,x;\alpha,b) \; \widehat{H}_{k_m}(\mathrm{d}\alpha, \mathrm{d}b) = \int t(w,x;\alpha,b) \; H_0(\mathrm{d}\alpha, \mathrm{d}b) = T_{2,0}(w,x).$$

Let $\overline{T}^*$ be as defined in (64). Under Assumption 4.6, $\int \left(\overline{T}^*\right)^2 \mathrm{d}G_0 < \infty$ and $|t(w,x,\alpha,b)| \leq \overline{T}^*(w,x)$ given any w, x, α, b as argued in the proof of Lemma C.2. Integrating against $\widehat{h}^2_{n_{k_m}}$ and h_0^2 respectively, we get $\left|\widehat{T}_{2,n_{k_m}}(w,x)\right| \leq \overline{T}^*(w,x)$ and $|T_{2,0}(w,x)| \leq \overline{T}^*(w,x)$. By the Dominated Convergence Theorem, on the event $\mathcal{Z}\left(\{n_{k_m}\}\right)$,

$$\lim_{m\to\infty} \int \left(\widehat{T}_{2,n_{k_m}}(w,x) - T_{2,0}(w,x)\right)^2 G_0(\mathrm{d}w, \mathrm{d}x) = 0.$$

Since the event $\mathcal{Z}\left(\{n_{k_m}\}\right)$ has Pr-probability 1, it follows that

$$\int \left(\widehat{T}_{2,n_{k_m}}(w,x) - T_{2,0}(w,x)\right)^2 G_0(\mathrm{d}w, \mathrm{d}x) \overset{\text{a.s.}}{\to} 0.$$

Thus, every subsequence $\{n_k\}$ has a further subsequence $\{n_{k_m}\}$ on which $\widehat{T}_{2,n_{k_m}}(\cdot)$ converges to $T_{2,0}$ in $\mathscr{L}_2(G_0)$ almost surely. By Theorem 2.3.2 of Durrett (2019), it follows that

$$\int \left(\widehat{T}_{2,n}(w,x) - T_{2,0}(w,x)\right)^2 G_0(\mathrm{d}w, \mathrm{d}x) \overset{\text{p}}{\to} 0,$$

which is exactly (65). □

D.3 Proof of Lemma C.5

Proof of Lemma C.5. Set $v = \theta - \theta_0$ and write $v = (v_\alpha, v_h)$. Then

$$\begin{aligned}\|v\|^2 &= \mathbb{E}\left[\left(\Delta_\alpha\left(Y, W, X; \theta_0\right)' v_\alpha + \Delta_h\left(Y, W, X; \theta_0\right)\left[v_h\right]\right)^2\right] \\ &\leq \mathbb{E}\left[\left(\left|\Delta_\alpha\left(Y, W, X; \theta_0\right)' v_\alpha\right| + \left|\Delta_h\left(Y, W, X; \theta_0\right)\left[v_h\right]\right|\right)^2\right].\end{aligned}$$

By Jensen's inequality,

$$\|v\|^2 \leq 2\left\{\mathbb{E}\left[\left(\Delta_\alpha\left(Y, W, X; \theta_0\right)' v_\alpha\right)^2\right] + \mathbb{E}\left[\left(\Delta_h\left(Y, W, X; \theta_0\right)\left[v_h\right]\right)^2\right]\right\}. \tag{104}$$

For the first term in (104), by the Cauchy-Schwarz inequality,

$$\mathbb{E}\left[\left(\Delta_\alpha\left(Y, W, X; \theta_0\right)' v_\alpha\right)^2\right] \leq \mathbb{E}\left[\left\|\Delta_\alpha\left(Y, W, X; \theta_0\right)\right\|_2^2\right] \cdot \left\|v_\alpha\right\|_2^2. \tag{105}$$

For the second term in (104),

$$\begin{aligned}\mathbb{E}\left[\left(\Delta_h\left(Y, W, X; \theta_0\right)\left[v_h\right]\right)^2\right] &= 4\mathbb{E}\left[\left\{\frac{\int \kappa\left(Y, W, X; \alpha_0, b\right) h_0(b) v_h(b)\, \mathrm{d}b}{P\left(Y, W, X; \alpha_0, h_0\right)}\right\}^2\right] \\ &\leq 4\mathbb{E}\left[\frac{1}{P\left(Y, W, X; \alpha_0, h_0\right)^2}\right]\left\{\int h_0(b)^2 \mathrm{d}b\right\}\left\{\int v_h(b)^2\, \mathrm{d}b\right\}.\end{aligned}$$

Apply the Cauchy-Schwarz inequality and use the fact that $\kappa(\cdot) \in (0,1)$ to get

$$\mathbb{E}\left[\left(\Delta_h\left(Y, W, X; \theta_0\right)\left[v_h\right]\right)^2\right] \leq 4\mathbb{E}\left[\frac{1}{P\left(Y, W, X; \alpha_0, h_0\right)^2}\right]\left\{\int h_0(b)^2 \mathrm{d}b\right\}\left\{\int v_h(b)^2\, \mathrm{d}b\right\}. \tag{106}$$

Therefore, define $C_{\|\cdot\|,\rho}$ by

$$C_{\|\cdot\|,\rho}^2 = 2\max\left\{\mathbb{E}\left[\left\|\Delta_\alpha\left(Y, W, X; \theta_0\right)\right\|_2^2\right], 4\mathbb{E}\left[\frac{1}{P\left(Y, W, X; \alpha_0, h_0\right)^2}\right] \cdot \int h_0(b)^2 \mathrm{d}b\right\}. \tag{107}$$

Combine (104), (105), and (106) with (107) to get

$$\|v\|^2 \leq C_{\|\cdot\|,\rho}^2 \cdot \left(\left\|v_\alpha\right\|_2^2 + \int v_h(b)^2\, \mathrm{d}b\right).$$

Since $v = \theta - \theta_0$, the conclusion of Lemma C.5 follows. □

D.4 Proofs of Lemmas required for Theorem C.1

D.4.1 Proof of Lemma C.6

Proof of Lemma C.6. Part (i): We wish to show

$$\sqrt{n}\frac{\sup_{\theta\in\mathcal{N}_n}\left|\phi(\theta)-\phi\left(\theta_0\right)-\frac{\partial\phi(\theta_0)}{\partial\theta}\left[\theta-\theta_0\right]\right|}{\left\|v_n^*\right\|}=o(1). \tag{108}$$

Since $\|v_n^*\|$ is non-decreasing, it suffices to show

$$\sqrt{n}\sup_{\theta\in\mathcal{N}_n}\left|\phi(\theta)-\phi\left(\theta_0\right)-\frac{\partial\phi\left(\theta_0\right)}{\partial\theta}\left[\theta-\theta_0\right]\right|=o(1). \tag{109}$$

Given $\theta=(\alpha,h)$, with some algebra using (66) and (76),

$$\begin{aligned}
&\phi(\theta)-\phi\left(\theta_0\right)-\frac{\partial\phi\left(\theta_0\right)}{\partial\theta}\left[\theta-\theta_0\right]\\
=&\int\left\{T_1(\alpha,b)-T_1\left(\alpha_0,b\right)-\left[\frac{\partial}{\partial\alpha}T_1\left(\alpha_0,b\right)\right]'\left(\alpha-\alpha_0\right)\right\}h_0(b)^2\,\mathrm{d}b\\
&+\int T_1\left(\alpha_0,b\right)\left(h(b)-h_0(b)\right)^2\mathrm{d}b\\
&+\int\left[T_1\left(\alpha,b\right)-T_1\left(\alpha_0,b\right)\right]\left(h(b)^2-h_0(b)^2\right)\mathrm{d}b.
\end{aligned} \tag{110}$$

Using Assumption 4.7, a second order Taylor expansion of $T_1(\alpha,b)$ around $\alpha=\alpha_0$ gives

$$\begin{aligned}
&\int\left\{T_1(\alpha,b)-T_1\left(\alpha_0,b\right)-\left[\frac{\partial}{\partial\alpha}T_1\left(\alpha_0,b\right)\right]'\left(\alpha-\alpha_0\right)\right\}h_0(b)^2\,\mathrm{d}b\\
=&\left(\alpha-\alpha_0\right)'\left[\int\frac{\partial}{\partial\alpha\partial\alpha}T_1\left(\widetilde{\alpha}(b),b\right)\,\mathrm{d}b\right]\left(\alpha-\alpha_0\right),
\end{aligned}$$

for a midpoint $\widetilde{\alpha}(b)$ between α and α_0. Thus, for a constant $C_1\in(0,\infty)$ determined by the maximal eigenvalue of the Hessian matrix in the display above,

$$\begin{aligned}
\left|\int\left\{T_1(\alpha,b)-T_1\left(\alpha_0,b\right)-\left[\frac{\partial}{\partial\alpha}T_1\left(\alpha_0,b\right)\right]'\left(\alpha-\alpha_0\right)\right\}h_0(b)^2\,\mathrm{d}b\right|&\leq C_1\left\|\alpha-\alpha_0\right\|_2^2\\
&\leq C_1\rho\left(\theta,\theta_0\right)^2.
\end{aligned} \tag{111}$$

Note that the last inequality follows from $\rho\left(\theta,\theta_0\right)^2=\left\|\alpha-\alpha_0\right\|_2^2+\rho_{\mathscr{L}_2}\left(h,h_0\right)^2$.

$T_1(\cdot)$ is continuous on $\mathcal{A}\times\mathcal{B}$, a compact set, and is therefore a bounded function. So,

$$\begin{aligned}
\left|\int T_1\left(\alpha_0,b\right)\left(h(b)-h_0(b)\right)^2\mathrm{d}b\right|&\leq C_2\rho_{\mathscr{L}_2}\left(h,h_0\right)^2\\
&\leq C_2\rho\left(\theta,\theta_0\right)^2,
\end{aligned} \tag{112}$$

where $C_2 = \sup_{(\alpha,b)\in\mathcal{A}\times\mathcal{B}} |T_1(\alpha,b)|$.

Next,

$$\begin{aligned}
&\left|\int \left[T_1(\alpha,b) - T_1(\alpha_0,b)\right]\left(h(b)^2 - h_0(b)^2\right) \mathrm{d}b\right| \\
=& \left|(\alpha-\alpha_0)' \int \frac{\partial}{\partial\alpha} T_1(\widetilde{\alpha}(b),b)\,(h(b)+h_0(b))\,(h(b)-h_0(b))\,\mathrm{d}b\right| \\
\leq& \ \|\alpha-\alpha_0\|_2 \cdot \int \left\|\frac{\partial}{\partial\alpha} T_1(\widetilde{\alpha}(b),b)\right\| |h(b)+h_0(b)|\,|h(b)-h_0(b)|\,\mathrm{d}b
\end{aligned}$$

By continuity of the first derivative $(\partial/\partial\alpha)T_1(\cdot)$ and compactness of $\mathcal{A}\times\mathcal{B}$, there is a $C_3 \in (0,\infty)$ such that

$$\begin{aligned}
&\left|\int \left[T_1(\alpha,b) - T_1(\alpha_0,b)\right]\left(h(b)^2 - h_0(b)^2\right) \mathrm{d}b\right| \\
&\leq C_3 \|\alpha-\alpha_0\|_2 \int |h(b)+h_0(b)|\,|h(b)-h_0(b)|\,\mathrm{d}b \\
\text{(by Cauchy-Schwarz)}\quad &\leq C_3 \|\alpha-\alpha_0\|_2 \left\{\int |h(b)+h_0(b)|^2 \mathrm{d}b\right\}^{1/2} \cdot \left\{\int |h(b)-h_0(b)|^2 \mathrm{d}b\right\}^{1/2} \\
&\leq 2C_3 \|\alpha-\alpha_0\|_2\, \rho_{\mathscr{L}_2}(h,h_0)
\end{aligned}$$

since $\int |h+h_0|^2 \leq 4$ by Lemma B.6. By $uv \leq \left(u^2+v^2\right)/2$, we have

$$\begin{aligned}
\left|\int \left[T_1(\alpha,b) - T_1(\alpha_0,b)\right]\left(h(b)^2 - h_0(b)^2\right) \mathrm{d}b\right| &\leq C_3\left(\|\alpha-\alpha_0\|_2^2 + \rho_{\mathscr{L}_2}(h,h_0)^2\right) \\
&= C_3 \rho(\theta,\theta_0)^2 .
\end{aligned} \tag{113}$$

Combine (111), (112) and (113) with (110) to get

$$\left|\phi(\theta) - \phi(\theta_0) - \frac{\partial\phi(\theta_0)}{\partial\theta}[\theta-\theta_0]\right| \leq C_4 \rho(\theta,\theta_0)^2 , \tag{114}$$

for $C_4 = C_1 + C_2 + C_3$. Let $\theta \in \mathcal{N}_n$ defined in (70). By definition of $\mathcal{N}_n$,

$$\begin{aligned}
\sqrt{n}\rho(\theta,\theta_0)^2 &\leq \sqrt{n}\max\left\{K_n^{-\frac{2s}{d_X}}, \frac{K_n}{n}\right\} \cdot \zeta_n^2 \\
&= \max\left\{\sqrt{n}K_n^{-\frac{2s}{d_X}}, \frac{K_n}{\sqrt{n}}\right\} \zeta_n^2 \\
&= o(1).
\end{aligned}$$

where the last line follows by (69). Hence, using (114),

$$\sqrt{n} \sup_{\theta \in \mathcal{N}_n} \left| \phi(\theta) - \phi(\theta_0) - \frac{\partial \phi(\theta_0)}{\partial \theta} [\theta - \theta_0] \right| \leq C_4 \sup_{\theta \in \mathcal{N}_n} \sqrt{n} \rho(\theta, \theta_0)^2$$
$$= o(1)$$

This shows (109), which in turn is sufficient for (108).

Part (ii): Recall that we wish to show either

$$\begin{aligned} &\limsup_{n \to \infty} \|v_n^*\| = \infty \quad \text{and} \quad \sqrt{n} \cdot \frac{\left| \frac{\partial \phi(\theta_0)}{\partial \theta} [\theta - \theta_0] \right|}{\|v_n^*\|} = o(1), \\ \text{or } &\limsup_{n \to \infty} \|v_n^*\| < \infty \quad \text{and} \quad \sqrt{n} \cdot \|v^* - v_n^*\| \cdot \|\theta_{0,n} - \theta_0\| = o(1). \end{aligned}$$

Again, since $\|v_n^*\|$ is non-decreasing and positive, it suffices to show the following respectively:

$$\text{if } \limsup_{n \to \infty} \|v_n^*\| = \infty \quad \text{then} \quad \sqrt{n} \cdot \left| \frac{\partial \phi(\theta_0)}{\partial \theta} [\theta_{0,n} - \theta_0] \right| = o(1), \tag{115}$$

$$\text{else, if } \limsup_{n \to \infty} \|v_n^*\| < \infty \quad \text{then} \quad \sqrt{n} \cdot \|\theta_{0,n} - \theta_0\| = o(1). \tag{116}$$

Recall from (76)

$$\frac{\partial \phi(\theta_0)}{\partial \theta} [\theta - \theta_0] = \left[\int \frac{\partial}{\partial \alpha} T_1(\alpha_0, b) h_0(b)^2 \mathrm{d}b \right]' (\alpha - \alpha_0) + 2 \int T_1(\alpha_0, b) h_0(b) (h(b) - h_0(b)) \; \mathrm{d}b.$$

Applying the triangle and Cauchy-Schwarz inequalities

$$\begin{aligned} \left| \frac{\partial \phi(\theta_0)}{\partial \theta} [\theta - \theta_0] \right| \leq & \left\| \int \frac{\partial}{\partial \alpha} T_1(\alpha_0, b) h_0(b)^2 \mathrm{d}b \right\|_2 \|\alpha - \alpha_0\|_2 \\ & + 2C_2 \left\{ \int h_0(b)^2 \; \mathrm{d}b \right\}^{1/2} \left\{ \int (h(b) - h_0(b))^2 \; \mathrm{d}b \right\}^{1/2}, \end{aligned}$$

where as before, $C_2 = \sup_{(\alpha,b) \in \mathcal{A} \times \mathcal{B}} |T_1(\alpha, b)|$. In addition, $\int h_0^2 = 1$. Therefore, using continuity of the first derivative $(\partial/\partial\alpha) T_1(\cdot)$ and compactness of $\mathcal{A} \times \mathcal{B}$ to bound the first component of the sum above, and applying it to $\theta_{0,n}$ in (74), we get that for some $C_5 \in (0, \infty)$,

$$\left| \frac{\partial \phi(\theta_0)}{\partial \theta} [\theta_{0,n} - \theta_0] \right| \leq C_5 \rho(\theta_{0,n}, \theta_0).$$

Furthermore, using the inequality (75), we have $\|\theta_{0,n} - \theta_0\| \leq C_{\|\cdot\|,\rho} \cdot \rho(\theta_{0,n}, \theta_0)$ for positive and finite $C_{\|\cdot\|,\rho}$. Then, it suffices to show $\sqrt{n} \cdot \rho(\theta_{0,n}, \theta_0) = o(1)$ for both (115) and (116). To that end, by Assumption 4.4 (i), $\sqrt{n}\rho(\theta_{0,n}, \theta_0) = O\left(K_n^{-s/d_X}\right) = o(1)$, since $\sqrt{n} \cdot K_n^{-s/d_X} = o(1)$ by hypothesis. As previously argued, (115) and (116) imply (94) and (95) respectively. □

D.4.2 A useful consequence of Assumption C.1 (i)

Lemma D.1. *Let Assumption 4.4 (i) and Assumption C.1 (i) hold. Take any* $n \in \mathbb{N}$ *and* $v \in \mathcal{V}_n$ *with* $v = (v_\alpha, v_h(\cdot)) = (v_\alpha, \lambda_h' \boldsymbol{\psi}_{K_n}(\cdot))$. *Then, with* $c_{\mathcal{I}} \in (0, \infty)$ *as defined in Assumption C.1 (i),*

$$\|v\| \geq \sqrt{c_{\mathcal{I}}} \cdot \max \{\|v_\alpha\|_2, \|v_h\|_2\}, \tag{117}$$

where $\|v_h\|_2^2 = \int v_h^2$.

Proof of Lemma D.1. Take any $n \in \mathbb{N}$ and $v \in \mathcal{V}_n$ with $v = (v_\alpha, v_h(\cdot)) = (v_\alpha, \lambda_h' \boldsymbol{\psi}_{K_n}(\cdot))$. By orthonormality of $\boldsymbol{\psi}_{K_n}$, $\|v_h\|_2^2 = \|\lambda_h\|_2^2$. Therefore,

$$\begin{aligned}\|v\|^2 = \left(v_\alpha', \lambda_h'\right) \mathcal{I}_n \begin{pmatrix} v_\alpha \\ \lambda_h \end{pmatrix} &\geq c_{\mathcal{I}} \cdot \left(\|v_\alpha\|_2^2 + \|\lambda_h\|_2^2\right) = c_{\mathcal{I}} \cdot \left(\|v_\alpha\|_2^2 + \|v_h\|_2^2\right) \\ &\geq c_{\mathcal{I}} \cdot \max\left\{\|v_\alpha\|_2^2, \|v_h\|_2^2\right\}.\end{aligned}$$

Thus, (117) follows. □

D.4.3 Proof of Lemma C.7

In this subsection and the next, define

$$u_n^* = \frac{v_n^*}{\|v_n^*\|}. \tag{118}$$

Here, we verify each of the three parts of Assumption 2.2 of Chen and Liao (2014).

Proof of Lemma C.7. Part (i) requires the functional $\mu_n \{\Delta(Y, W, X; \theta_0)[v]\}$ to be linear in $v \in \mathcal{V}$. This is immediately satisfied since Δ in (71) is linear in v and $f \mapsto \mu_n f$ is linear.

Part (ii) $\sup_{\theta \in \mathcal{N}_n} |\mu_n \{\Delta(\cdot; \theta)[u_n^*] - \Delta(\cdot; \theta_0)[u_n^*]\}| = o_{\mathrm{p}}\left(n^{-1/2}\right)$. By (138) in Lemma D.4, given any $\theta_1, \theta_2 \in \Theta$ and $v \in \mathcal{V}$,

$$\begin{aligned}|\Delta(y, w, x; \theta_1)[v] - \Delta(y, w, x; \theta_2)[v]| \leq\ & \overline{U}_{\Delta,1}(y, w, x) \cdot \max\{\|v_\alpha\|_2, \|v_h\|_2\} \cdot \rho(\theta_1, \theta_2) \\ & + \overline{U}_{\Delta,2}(y, w, x) \cdot \max\left\{\|v_\alpha\|_2^3, \|v_h\|_2^3\right\} \rho(\theta_1, \theta_2)^2,\end{aligned}$$

for non-negative functions $\overline{U}_{\Delta,j}$, $j \in \{1, 2\}$ that are square-integrable against ν_0 under Assumption 4.2. Hence, squaring and dividing through by $\|v\|$, we get that for any $u \in \mathcal{V}$ with $\|u\| = 1$,

$$\begin{aligned}&|\Delta(y, w, x; \theta_1)[u] - \Delta(y, w, x; \theta_2)[u]|^2 \\ &\leq 2 \max_{j=1,2} \left\{\overline{U}_{\Delta,j}(y, w, x)^2\right\} \times \\ &\quad \times \max\left\{\|u_\alpha\|_2, \|u_h\|_2, \|u_\alpha\|_2^3, \|u_h\|_2^3\right\}^2 \times \\ &\quad \times \max\left\{\rho(\theta_1, \theta_2)^2, \rho(\theta_1, \theta_2)^4\right\}.\end{aligned}$$

Hence,

$$\left|\Delta\left(y,w,x;\theta_1\right)[u]-\Delta\left(y,w,x;\theta_2\right)[u]\right|^2\leq C_*\cdot\overline{U}_{\Delta,*}(y,w,x)^2\cdot\max\left\{\rho\left(\theta_1,\theta_2\right)^2,\rho\left(\theta_1,\theta_2\right)^4\right\},\quad(119)$$

for a constant $C_*>0$ and a function $\overline{U}_{\Delta,*}$ that is square-integrable against ν_0.

For $\delta\in(0,1]$, denote the class of functions

$$\mathcal{D}_{n,\delta}=\left\{\Delta(\cdot;\theta)\left[u_n^*\right]-\Delta\left(\cdot;\theta_0\right)\left[u_n^*\right]:\theta\in\mathcal{N}_n,\rho\left(\theta,\theta_0\right)\leq\delta\right\}.\quad(120)$$

By (119) and Theorem 2.7.11 in van der Vaart and Wellner (1996)

$$\log N_{[]}\left(\varepsilon,\mathcal{D}_{n,\delta},\mathscr{L}_2\left(\nu_0\right)\right)\leq\log N\left(\frac{\varepsilon}{C_{*,1}},\Theta_{K_n,\delta},\rho\right),$$

where

$$C_{*,1}^2=C_*\nu_0\left[\overline{U}_{\Delta,*}^2\right],$$

$$\text{and}\qquad\Theta_{K_n,\delta}=\left\{\theta\in\Theta_{K_n}^2:\rho\left(\theta,\theta_0\right)\leq\delta\right\}.$$

Effectively the same arguments as those used to establish (50) in Lemma B.15 show that

$$\log N\left(\frac{\varepsilon}{C_{*,1}},\overline{\Theta}_{K_n,\delta},\rho\right)\leq\left(d_W+K_n\right)\cdot\max\left\{0,\log\left(\frac{3C_{*,1}\delta}{\varepsilon}\right)\right\}.$$

and hence

$$\log N_{[]}\left(\varepsilon,\mathcal{D}_{n,\delta},\mathscr{L}_2\left(\nu_0\right)\right)\leq\left(d_W+K_n\right)\cdot\max\left\{0,\log\left(\frac{3C_{*,1}\delta}{\varepsilon}\right)\right\}.$$

Therefore,

$$\begin{aligned}\int_0^\infty\sqrt{\log N_{[]}\left(\varepsilon,\mathcal{D}_{n,\delta},\mathscr{L}_2\left(\nu_0\right)\right)}\,\mathrm{d}\varepsilon&\leq2\left(d_W+K_n\right)\cdot\int_0^{3C_{*,1}\delta}\sqrt{\log\left(\frac{3C_{*,1}\delta}{\varepsilon}\right)}\,\mathrm{d}\varepsilon\\&=\frac{\sqrt{\pi}}{2}\cdot\left(d_W+K_n\right)<\infty.\end{aligned}$$

where the last equality follows from (158) in Lemma F.4.

In addition, with $u=u_n^*$, taking suprema and then integrating,

$$\begin{aligned}&\mathbb{E}\left[\sup_{\theta_1,\theta_2\in\mathcal{N}_n:\rho(\theta_1,\theta_2)\leq\delta}\left|\Delta\left(Y,W,X;\theta_1\right)\left[u_n^*\right]-\Delta\left(Y,W,X;\theta_2\right)\left[u_n^*\right]\right|^2\right]\\&\leq C_*\cdot\mathbb{E}\left[\overline{U}_{\Delta,*}(Y,W,X)\right]\cdot\max\left\{\delta^2,\delta^4\right\}.\end{aligned}$$

As $\delta \downarrow 0$, since $\delta^2 \geq \delta^4$ (as soon as $\delta \leq 1$), we further have

$$\mathbb{E}\left[\sup_{\theta_1,\theta_2 \in \mathcal{N}_n : \rho(\theta_1,\theta_2) \leq \delta} \left|\Delta\left(Y, W, X; \theta_1\right)\left[u_n^*\right] - \Delta\left(Y, W, X; \theta_2\right)\left[u_n^*\right]\right|^2\right] \leq C_* \cdot \mathbb{E}\left[\overline{U}_{\Delta,*}(Y, W, X)\right] \delta^2.$$

Thus, the conditions of Theorem 3 in Chen et al. (2003) (or Lemma 4.2 in Chen (2007)) are met, which implies that

$$\sup_{\theta \in \mathcal{N}_n} \mu_n \left\{\Delta(\cdot;\theta)\left[u_n^*\right] - \Delta\left(\cdot;\theta_0\right)\left[u_n^*\right]\right\} = o_{\mathrm{p}}\left(\frac{1}{\sqrt{n}}\right).$$

This proves part (ii) of Lemma C.7.

Part (iii): Let $\theta \in \mathcal{N}_n$, and let $v = \theta - \theta_0$. By a second order (pathwise) Taylor expansion,

$$\begin{aligned}\ell(y, w, x; \theta) = {} & \ell\left(y, w, x; \theta_0\right) + \Delta\left(y, w, x; \theta_0\right)[v] + \frac{1}{2} r\left(y, w, x; \theta_0\right)[v, v] \\ & + \frac{1}{2}\left[r\left(y, w, x; (1 - \widetilde{\varepsilon}) \cdot \theta_0 + \widetilde{\varepsilon} \cdot \theta\right)[v, v] - r\left(y, w, x; \theta_0\right)[v, v]\right]\end{aligned}$$

where $\widetilde{\varepsilon} = \widetilde{\varepsilon}\left(y, w, x; \theta, \theta_0\right) \in (0, 1)$. By Lemma C.12 and Lemma D.2 (below)

$$\begin{aligned}\mathbb{E}\left[\Delta\left(Y, W, X; \theta_0\right)[v]\right] &= 0 \\ \mathbb{E}\left[r\left(Y, W, X; \theta_0\right)[v, v]\right] = \mathbb{E}\left[-\left(\Delta\left(Y, W, X; \theta_0\right)[v]\right)^2\right] &= -\|v\|^2.\end{aligned}$$

Hence,

$$\begin{aligned}& \mathbb{E}\left[\ell\left(Y, W, X; \theta_0\right) - \ell(Y, W, X; \theta)\right] \\ = {} & -\mathbb{E}\left[\Delta\left(y, w, x; \theta_0\right)[v]\right] - \frac{1}{2}\mathbb{E}\left[r\left(y, w, x; \theta_0\right)[v, v]\right] \\ & + \frac{1}{2}\mathbb{E}\left[r\left(y, w, x; (1 - \widetilde{\varepsilon}) \cdot \theta_0 + \widetilde{\varepsilon} \cdot \theta\right)[v, v] - r\left(y, w, x; \theta_0\right)[v, v]\right] \\ = {} & \frac{\|v\|^2}{2} + \frac{\|v\|^2}{2}\mathbb{E}\left[r\left(y, w, x; (1 - \widetilde{\varepsilon}) \cdot \theta_0 + \widetilde{\varepsilon} \cdot \theta\right)[u, u] - r\left(y, w, x; \theta_0\right)[u, u]\right]\end{aligned}$$

where $u = v/\|v\|$. Therefore,

$$\begin{aligned}& \mathbb{E}\left[\ell\left(Y, W, X; \theta_0\right) - \ell(Y, W, X; \theta)\right] - \frac{\|v\|^2}{2} \\ = {} & \frac{\|v\|^2}{2}\mathbb{E}\left[r\left(Y, W, X; (1 - \widetilde{\varepsilon}) \cdot \theta_0 + \widetilde{\varepsilon} \cdot \theta\right)[u, u] - r\left(Y, W, X; \theta_0\right)[u, u]\right]\end{aligned}$$

Substituting $v = \theta - \theta_0$, bounding both sides and using Assumption C.1 (ii),

$$\begin{aligned}
&\sup_{\theta \in \mathcal{N}_n} \left| \mathbb{E}\left[\ell\left(Y, W, X; \theta_0\right) - \ell(Y, W, X; \theta)\right] - \frac{\|\theta - \theta_0\|^2}{2} \right| \\
&= \sup_{\theta \in \mathcal{N}_n, \widetilde{\theta} \in \mathcal{N}_{0,n}} \frac{\|\theta - \theta_0\|^2}{2} \sup_{u \in \mathcal{N}_{0,n} : \|u\| = 1} \mathbb{E}\left[\left| r\left(Y, W, X; \widetilde{\theta}\right)[u, u] - r\left(Y, W, X; \theta_0\right)[u, u]\right|\right] \\
&= o\left(\frac{1}{n}\right).
\end{aligned}$$

This proves part (iii) of Lemma C.7. □

Lemma D.2. *Given any $\theta \in \Theta$ and $v_1, v_2 \in \mathcal{V}$, for all w, x,*

$$\sum_{y=0}^{J} P(y, w, x; \theta) \cdot r(y, w, x; \theta)\left[v_1, v_2\right] = -\sum_{y=0}^{J} P(y, w, x; \theta) \cdot \Delta(y, w, x; \theta)\left[v_1\right] \cdot \Delta(y, w, x; \theta)\left[v_2\right]. \quad (121)$$

Proof of Lemma D.2. For any θ, and any w, x, $\sum_{y=0}^{J} P(y, w, x; \theta) = 1$. This implies that along any path,

$$\sum_{y=0}^{J} (\partial/\partial\varepsilon) P\left(y, w, x; \theta + \varepsilon v\right)|_{\varepsilon=0} = 0$$

$$\sum_{y=0}^{J} \left(\partial/\partial\varepsilon_2\right)\left(\partial/\partial\varepsilon_1\right) P\left(y, w, x; \theta + \varepsilon_1 v_1 + \varepsilon_2 v_2\right)|_{\varepsilon_1=0, \varepsilon_2=0} = 0$$

Hence,

$$\begin{aligned}
&\sum_{y=0}^{J} P(y, w, x; \theta) \cdot r(y, w, x; \theta)\left[v_1, v_2\right] \\
&= -\sum_{y=0}^{J} P(y, w, x; \theta) \cdot \Delta(y, w, x; \theta)\left[v_1\right] \cdot \Delta(y, w, x; \theta)\left[v_2\right] \\
&\quad + \sum_{y=0}^{J} P(y, w, x; \theta) \cdot \frac{\left(\partial/\partial\varepsilon_2\right)\left(\partial/\partial\varepsilon_1\right) P\left(y, w, x; \theta + \varepsilon_1 v_1 + \varepsilon_2 v_2\right)|_{\varepsilon_1=0, \varepsilon_2=0}}{P(y, w, x; \theta)} \\
&= -\sum_{y=0}^{J} P(y, w, x; \theta) \cdot \Delta(y, w, x; \theta)\left[v_1\right] \cdot \Delta(y, w, x; \theta)\left[v_2\right],
\end{aligned}$$

which proves (121). □

D.4.4 Proof of Lemma C.8

Proof of Lemma C.8. Let

$$\Delta_{0,n}(y, w, x) = \Delta\left(Y, W, X; \theta_0\right)\left[u_n^*\right], \quad (122)$$

where as in (118), $u_n^* = v_n^* / \|v_n^*\|$. Then, we can rewrite (96) as the requirement that

$$\sqrt{n}\mu_n \{\Delta_{0,n}\} \xrightarrow{\mathrm{d}} \mathcal{N}(0,1). \tag{123}$$

For this, it is sufficient to verify Lindeberg's condition, which in this context is:

$$\lim_{n\to\infty} \nu_0 \left[\Delta_{0,n}^2 \cdot \mathbb{I}\left\{|\Delta_{0,n}| > \varepsilon\sqrt{n}\right\}\right] = 0 \quad \text{for every } \varepsilon > 0. \tag{124}$$

Lemma D.3 shows (in (137)) that the following bound holds for any $v \in \mathcal{V}$ with $v = (v_\alpha, v_h)$:

$$|\Delta(y, w, x; \theta_0)[v]| \leq U_\Delta(y, w, x) \cdot \max\left\{\|v_\alpha\|_2, \|v_h\|_2\right\}.$$

Under Assumption 4.2, the function $U_\Delta(\cdot)$ is square-integrable under ν_0. By (117) of Lemma D.1, given any $n \in \mathbb{N}$ and $v \in \mathcal{V}_n$,

$$\left|\Delta(y, w, x; \theta_0)\left[\frac{v}{\|v\|}\right]\right| \leq \frac{U_\Delta(y, w, x)}{\sqrt{c_{\mathcal{I}}}}.$$

Therefore, given any $n \in \mathbb{N}$ and $\varepsilon > 0$,

$$\Delta_{0,n}^2 \cdot \mathbb{I}\left\{|\Delta_{0,n}| > \varepsilon\sqrt{n}\right\} \leq \frac{U_\Delta^2}{c_{\mathcal{I}}} \cdot \mathbb{I}\left\{|U_\Delta| > \varepsilon\sqrt{c_{\mathcal{I}} n}\right\}.$$

Therefore, by ν_0-integrability of U_Δ and dominated convergence, i.e.

$$\lim_{n\to\infty} \nu_0 \left[\Delta_{0,n}^2 \cdot \mathbb{I}\left\{|\Delta_{0,n}| > \varepsilon\sqrt{n}\right\}\right] = 0 \quad \text{for every } \varepsilon > 0.$$

Hence, (124) holds which implies (123), i.e. (96). □

D.4.5 Proof of Lemma C.9

Proof of Lemma C.9. Part (i):

$$\begin{aligned}
&\Delta(\cdot;\theta)[v_1]\Delta(\cdot;\theta)[v_2] - \Delta(\cdot;\theta_0)[v_1]\Delta(\cdot;\theta_0)[v_2] \\
&= \Delta(\cdot;\theta)[v_1]\Delta(\cdot;\theta)[v_2] - \Delta(\cdot;\theta)[v_1]\Delta(\cdot;\theta_0)[v_2] \\
&\quad + \Delta(\cdot;\theta)[v_1]\Delta(\cdot;\theta_0)[v_2] - \Delta(\cdot;\theta_0)[v_1]\Delta(\cdot;\theta_0)[v_2] \\
&= \Delta(\cdot;\theta)[v_1] \cdot \{\Delta(\cdot;\theta)[v_2] - \Delta(\cdot;\theta_0)[v_2]\} \\
&\quad + \Delta(\cdot;\theta_0)[v_2]\{\Delta(\cdot;\theta)[v_1] - \Delta(\cdot;\theta_0)[v_1]\}.
\end{aligned}$$

Thus, by the triangle inequality

$$\begin{aligned}
& \left|\Delta(\cdot;\theta)\left[v_1\right]\Delta(\cdot;\theta)\left[v_2\right]-\Delta\left(\cdot;\theta_0\right)\left[v_1\right]\Delta\left(\cdot;\theta_0\right)\left[v_2\right]\right| \\
\leq\ & \left|\Delta(\cdot;\theta)\left[v_1\right]\right|\left|\Delta(\cdot;\theta)\left[v_2\right]-\Delta\left(\cdot;\theta_0\right)\left[v_2\right]\right| \\
& +\left|\Delta\left(\cdot;\theta_0\right)\left[v_2\right]\right|\left|\Delta(\cdot;\theta)\left[v_1\right]-\Delta\left(\cdot;\theta_0\right)\left[v_1\right]\Delta\left(\cdot;\theta_0\right)\left[v_2\right]\right|
\end{aligned}$$

By Lemma D.3 and Lemma D.4, we have respectively

$$\begin{aligned}
|\Delta(\cdot;\theta)[v]| &\leq U_\Delta(\cdot)\cdot\max\left\{\left\|v_\alpha\right\|_2,\left\|v_h\right\|_2\right\}, \\
\left|\Delta(\cdot;\theta)[v]-\Delta\left(\cdot;\theta_0\right)[v]\right| &\leq \overline{U}_{\Delta,1}(\cdot)\cdot\max\left\{\left\|v_\alpha\right\|_2,\left\|v_h\right\|_2\right\}\cdot\rho\left(\theta,\theta_0\right) \\
&\quad+\overline{U}_{\Delta,2}(\cdot)\cdot\max\left\{\left\|v_\alpha\right\|_2^3,\left\|v_h\right\|_2^3\right\}\rho\left(\theta_1,\theta_2\right)^2,
\end{aligned}$$

so that with $v_j=\left(v_{\alpha,j},v_{h,j}\right)$ for $j\in\{1,2\}$

$$\begin{aligned}
& \left|\Delta(\cdot;\theta)\left[v_1\right]\Delta(\cdot;\theta)\left[v_2\right]-\Delta\left(\cdot;\theta_0\right)\left[v_1\right]\Delta\left(\cdot;\theta_0\right)\left[v_2\right]\right| \\
\leq\ & 2U_\Delta(\cdot)\max_{j=1,2}\left\{\overline{U}_{\Delta,j}(\cdot)\right\} \\
& \times\max_{j=1,2}\left\{\left\|v_{\alpha,j}\right\|_2,\left\|v_{h,j}\right\|_2\right\} \\
& \times\max_{j=1,2}\left\{\left\|v_{\alpha,j}\right\|_2,\left\|v_{h,j}\right\|_2,\left\|v_{\alpha,j}\right\|_2^3,\left\|v_{h,j}\right\|_2^3\right\} \\
& \times\max\left\{\rho\left(\theta,\theta_0\right),\rho\left(\theta,\theta_0\right)^2\right\}
\end{aligned}$$

Thus, for $\left\|v_j\right\|=1$, the terms containing $\left\|v_{\alpha,j}\right\|_2,\left\|v_{h,j}\right\|_2$ are all bounded and collecting the envelope functions into one term, we have

$$\left|\Delta(\cdot;\theta)\left[v_1\right]\Delta(\cdot;\theta)\left[v_2\right]-\Delta\left(\cdot;\theta_0\right)\left[v_1\right]\Delta\left(\cdot;\theta_0\right)\left[v_2\right]\right|\leq C_*U_{\Delta,*}(\cdot)\max\left\{\rho\left(\theta,\theta_0\right),\rho\left(\theta,\theta_0\right)^2\right\}, \tag{125}$$

where C_* is a finite, positive constant and

$$U_{\Delta,*}(\cdot)=U_\Delta(\cdot)\max_{j=1,2}\left\{\overline{U}_{\Delta,j}(\cdot)\right\}.$$

Now, $U_\Delta(\cdot)$ and $\overline{U}_{\Delta,j}(\cdot)$ for $j\in\{1,2\}$ are all square-integrable against ν_0 by Assumption 4.2 and hence, $U_{\Delta,*}(\cdot)$ is ν_0-integrable by the Cauchy-Schwarz inequality. Therefore, by (125)

$$\begin{aligned}
& \sup_{\theta\in\mathcal{N}_n,v_1,v_2\in\mathcal{V}_n:\left\|v_1\right\|=\left\|v_2\right\|=1}\left|\nu_0\left\{\Delta(\cdot;\theta)\left[v_1\right]\Delta(\cdot;\theta)\left[v_2\right]-\Delta\left(\cdot;\theta_0\right)\left[v_1\right]\Delta\left(\cdot;\theta_0\right)\left[v_2\right]\right\}\right| \\
\leq\ & C_*\nu_0\left[U_{\Delta,*}\right]\cdot\sup_{\theta\in\mathcal{N}_n}\max\left\{\rho\left(\theta,\theta_0\right)\rho\left(\theta,\theta_0\right)^2\right\} \\
\leq\ & \max\left\{K_n^{-s/d_X},\sqrt{\frac{K_n}{n}}\right\}\cdot\zeta_n \\
\to\ & 0,
\end{aligned}$$

where the inequality in the penultimate line follows from the definition of $\mathcal{N}_n$ in (70) and the limit claim in the final line follows from the definition of ζ_n in (69). Hence, the requirement in Lemma C.9 (i) follows.

Part (ii): define the family of functions

$$\mathcal{D}_n = \left\{\Delta(\cdot;\theta)\left[v_1\right]\Delta(\cdot;\theta)\left[v_2\right] : \theta \in \mathcal{N}_n, v_1, v_2 \in \mathcal{V}_n, \|v_1\| = \|v_2\| = 1\right\}. \tag{126}$$

Then, using (125), the fact that covering numbers are dominated by bracketing numbers for $\mathscr{L}_p$-norms and Theorem 2.7.11 in van der Vaart and Wellner (1996), given any ε

$$\begin{aligned}\log N\left(\varepsilon, \mathcal{D}_n, \mathscr{L}_1\left(\nu_n\right)\right) &\leq \log N_{[]}\left(\varepsilon, \mathcal{D}_n, \mathscr{L}_1\left(\nu_n\right)\right)\\ &\leq \log N\left(\frac{\varepsilon}{C_{*,n}}, \mathcal{N}_n, \rho\right),\end{aligned}$$

where we define

$$C_{*,n} = C_*\nu_n\left[\overline{U}_{\Delta,*}\right].$$

Effectively the same arguments as those used to establish (50) in Lemma B.15 show that

$$\log N\left(\varepsilon, \mathcal{D}_n, \mathscr{L}_1\left(\nu_n\right)\right) \leq \log N\left(\frac{\varepsilon}{C_{*,n}}, \mathcal{N}_n, \rho\right) \leq \left(d_W + K_n\right)\cdot \max\left\{0, \log\left(\frac{3C_{*,n}\delta_{*,n}}{\varepsilon}\right)\right\}.$$

where

$$\delta_{*,n} = \max\left\{K_n^{-s/d_X}, \sqrt{\frac{K_n}{n}}\right\}\cdot\zeta_n$$

from the definition of $\mathcal{N}_n$ in (70). By the Strong Law of Large Numbers, $C_{*,n} \overset{\text{a.s.}}{\to} C_*\nu_0\left[\overline{U}_{\Delta,*}\right] < \infty$ and by assumption, $\delta_{*,n} \to 0$. By the Continuous Mapping Theorem,

$$\max\left\{0, \log\left(\frac{3C_{*,n}\delta_{*,n}}{\varepsilon}\right)\right\} \overset{\text{a.s.}}{\to} 0.$$

Also by assumption, $K_n/n \to 0$ and so, also by the Continuous Mapping Theorem,

$$\frac{\log N\left(\varepsilon, \mathcal{D}_n, \mathscr{L}_1\left(\nu_n\right)\right)}{n} \leq \frac{d_W + K_n}{n}\cdot \max\left\{0, \log\left(\frac{3C_{*,n}\delta_{*,n}}{\varepsilon}\right)\right\} = o_{\mathrm{p}}(1).$$

Since $\varepsilon > 0$ was arbitrary, by Theorem 2.4.3 of van der Vaart and Wellner (1996),

$$\sup_{\theta\in\mathcal{N}_n, v_1, v_2\in\mathcal{V}_n : \|v_1\|=\|v_2\|=1}\left|\mu_n\left\{\Delta(\cdot;\theta)\left[v_1\right]\Delta(\cdot;\theta)\left[v_2\right]\right\}\right| = o_{\mathrm{p}}(1).$$

□

D.4.6 Proof of Lemma C.11

Proof of Lemma C.11. From Equation (88),

$$V_{\phi,n} = \|v_n^*\|, \tag{127}$$

which is non-decreasing since $\mathcal{V}_n$ is a sequence of nested non-decreasing sets and from (78) $\|v_n^*\|$ can be defined as a supremum over $\mathcal{V}_n$. Since any monotone and bounded sequence is convergent, we need only prove then that $\|v_n^*\|$ is bounded to establish the claim. To that end, combining (81) and (82), we have

$$\|v_n^*\|^2 = \Phi_n' \mathcal{I}_n^{-1} \Phi_n,$$

where Φ_n is defined in (79) as

$$\Phi_n = \begin{bmatrix} \int \frac{\partial}{\partial\alpha} T_1\left(\alpha_0, b\right) h_0(b)^2 \mathrm{d}b \\ 2\int T_1\left(\alpha_0, b\right) h_0(b) \boldsymbol{\psi}_{K_n}(b)\ \mathrm{d}b \end{bmatrix}.$$

Then, from Assumption C.1 (i), it follows that

$$\begin{aligned} \|v_n^*\|^2 &\leq c_{\mathcal{I}}^{-1} \|\Phi_n\|_2 \\ &= c_{\mathcal{I}}^{-1} \left\{ \left\| \int \frac{\partial}{\partial\alpha} T_1\left(\alpha_0, b\right) h_0(b)^2 \mathrm{d}b \right\|_2^2 + \left\| \int T_1\left(\alpha_0, b\right) h_0(b) \boldsymbol{\psi}_{K_n}(b)\ \mathrm{d}b \right\|_2^2 \right\} \end{aligned}$$

By Assumption 4.7, the component of Φ_n corresponding to α is finite, i.e.

$$\left\| \int \frac{\partial}{\partial\alpha} T_1\left(\alpha_0, b\right) h_0(b)^2 \mathrm{d}b \right\|_2 < \infty.$$

Therefore, we need to argue that the following is a bounded sequence:

$$\left\| \int T_1\left(\alpha_0, b\right) h_0(b) \boldsymbol{\psi}_{K_n}(b)\ \mathrm{d}b \right\|_2^2,$$

To that end, $\int T_1\left(\alpha_0, b\right) h_0(b) \boldsymbol{\psi}_{K_n}(b)\ \mathrm{d}b$ are generalized Fourier coefficients and $\boldsymbol{\psi}_{K_n}$ are orthonormal. By Bessel's inequality,

$$\left\| \int T_1\left(\alpha_0, b\right) h_0(b) \boldsymbol{\psi}_{K_n}(b)\ \mathrm{d}b \right\|_2^2 \leq \int \int T_1\left(\alpha_0, b\right)^2 h_0(b)^2 \mathrm{d}b.$$

By Assumption 4.6, and dominated convergence, $T_1\left(\alpha_0, b\right)$ is continuous in b and hence compactness of $\mathcal{B}$ implies that the right hand side of the above inequality is finite. Thus, $\|v_n^*\|$ is a bounded sequence. □

D.5 Proofs of Lemmas required for Theorem C.2

D.5.1 Proof of Lemma C.13

Proof of Lemma C.13. From Lemma C.2, $\mathscr{T}_*$ is a Donsker class with a square integrable envelope. Any Donsker class is also Glivenko-Cantelli (by Slutsky's Theorem). By Lemma 2.10.14 of van der Vaart and Wellner (1996), the class $\mathscr{T}_*^2 = \left\{t_*^2 : t \in \mathscr{T}_*\right\}$ is also Glivenko-Cantelli. Hence, we have both

$$\sup_{t_* \in \mathscr{T}_*} |\mu_n(t_*)| \xrightarrow{\mathrm{p}} 0, \quad \text{and} \quad \sup_{t_* \in \mathscr{T}_*} \left|\mu_n\left(t_*^2\right)\right| \xrightarrow{\mathrm{p}} 0. \tag{128}$$

Therefore,

$$\begin{aligned}
\widehat{\mathbb{V}}_{2,n} &= \nu_n\left[\widehat{T}_{2,n}^2\right] - \nu_n\left[\widehat{T}_{2,n}\right]^2 \\
&= \nu_0\left[\widehat{T}_{2,n}^2\right] + \mu_n\left[\widehat{T}_{2,n}^2\right] - \left\{\nu_0\left[\widehat{T}_{2,n}\right] + \mu_n\left[\widehat{T}_{2,n}\right]\right\}^2 \\
&= \nu_0\left[\widehat{T}_{2,n}^2\right] + o_{\mathrm{p}}(1) - \left\{\nu_0\left[\widehat{T}_{2,n}\right] + o_{\mathrm{p}}(1)\right\}^2
\end{aligned}$$

where the last line follows by (128). Hence,

$$\widehat{\mathbb{V}}_{2,n} = \nu_0\left[T_{2,0}^2\right] + \nu_0\left[\widehat{T}_{2,n}^2 - T_{2,0}^2\right] + o_{\mathrm{p}}(1) - \left\{\nu_0\left[T_{2,0}\right] + \nu_0\left[\widehat{T}_{2,n} - T_{2,0}\right] + o_{\mathrm{p}}(1)\right\}^2. \tag{129}$$

Thus, we are done if we can show that

$$\nu_0\left[\widehat{T}_{2,n} - T_{2,0}\right] = o_{\mathrm{p}}(1), \tag{130}$$

$$\nu_0\left[\widehat{T}_{2,n}^2 - T_{2,0}^2\right] = o_{\mathrm{p}}(1). \tag{131}$$

For (130), by Jensen's inequality,

$$\left|\nu_0\left[\widehat{T}_{2,n} - T_{2,0}\right]\right| \leq \sqrt{\nu_0\left[\left(\widehat{T}_{2,n} - T_{2,0}\right)^2\right]} = o_{\mathrm{p}}(1),$$

where the last equality follows from $\rho\left(\widehat{\theta}_n, \theta_0\right) \xrightarrow{\mathrm{p}} 0$ and (65) in Lemma C.3. For (131), by the Cauchy-Schwarz inequality,

$$\begin{aligned}
\left|\nu_0\left[\widehat{T}_{2,n}^2 - T_{2,0}^2\right]\right| &= \left|\nu_0\left[\left(\widehat{T}_{2,n} + T_{2,0}\right)\left(\widehat{T}_{2,n} - T_{2,0}\right)\right]\right| \\
&\leq \left(\nu_0\left[\left(\widehat{T}_{2,n} + T_{2,0}\right)^2\right]\right)^{1/2}\left(\nu_0\left[\left(\widehat{T}_{2,n} - T_{2,0}\right)^2\right]\right)^{1/2} \\
&\leq \left(2\nu_0\left[\left(\overline{T}^*\right)^2\right]\right)^{1/2}\left(\nu_0\left[\left(\widehat{T}_{2,n} - T_{2,0}\right)^2\right]\right)^{1/2}.
\end{aligned}$$

where $\overline{T}^*$ is the square-integrable envelope function for $\mathscr{T}_*$ from Lemma C.2. Again, combining $\rho\left(\widehat{\theta}_n, \theta_0\right) \xrightarrow{\mathrm{p}} 0$ and (65) in Lemma C.3,

$$\left|\nu_0\left[\widehat{T}_{2,n}^2 - T_{2,0}^2\right]\right| \leq o_{\mathrm{p}}(1).$$

Therefore, from (129), we get by the Continuous Mapping Theorem that

$$\widehat{\mathbb{V}}_{2,n} = \nu_0\left[T_{2,0}^2\right] + o_{\mathrm{p}}(1) - \left\{\nu_0\left[T_{2,0}\right] + o_{\mathrm{p}}(1)\right\}^2 = \nu_0\left[T_{2,0}^2\right] - \nu_0\left[T_{2,0}\right]^2 + o_{\mathrm{p}}(1) = \mathbb{V}_{2,0} + o_{\mathrm{p}}(1).$$

□

D.5.2 Proof of Lemma C.14

Proof of Lemma C.14. From (99),

$$\begin{aligned}
\frac{\widehat{\mathbb{C}}_n}{\|\widehat{v}_n^*\|} &= \frac{\frac{1}{n}\sum_{i=1}^n \Delta\left(Y_i, W_i, X_i; \widehat{\theta}_n\right)\left[\widehat{v}_n^*\right]\widetilde{T}_{2,n}\left(W_i, X_i\right)}{\|\widehat{v}_n^*\|} \\
&= \frac{1}{n}\sum_{i=1}^n \Delta\left(Y_i, W_i, X_i; \widehat{\theta}_n\right)\left[\frac{\widehat{v}_n^*}{\|\widehat{v}_n^*\|}\right]\widetilde{T}_{2,n}\left(W_i, X_i\right) \\
&= \nu_n\left[\widehat{\Delta}_n \cdot \widehat{T}_{2,n}\right] - \nu_n\left[\widehat{\Delta}_n\right] \cdot \nu_n\left[\widehat{T}_{2,n}\right],
\end{aligned}$$

where

$$\widehat{\Delta}_n\left(y, w, x\right) = \Delta\left(y, w, x; \widehat{\theta}_n\right)\left[\frac{\widehat{v}_n^*}{\|\widehat{v}_n^*\|}\right].$$

Thus,

$$\frac{\widehat{\mathbb{C}}_n}{\|\widehat{v}_n^*\|} = \nu_0\left[\widehat{\Delta}_n \cdot \widehat{T}_{2,n}\right] + \mu_n\left[\widehat{\Delta}_n \cdot \widehat{T}_{2,n}\right] - \nu_n\left[\widehat{\Delta}_n\right] \cdot \nu_n\left[\widehat{T}_{2,n}\right]. \tag{132}$$

We are hence tasked with showing the following:

$$\begin{aligned}
\nu_0\left[\widehat{\Delta}_n \cdot \widehat{T}_{2,n}\right] &= o_{\mathrm{p}}(1), & (133)\\
\nu_n\left[\widehat{\Delta}_n\right] \cdot \nu_n\left[\widehat{T}_{2,n}\right] &= o_{\mathrm{p}}(1), & (134)\\
\mu_n\left[\widehat{\Delta}_n \cdot \widehat{T}_{2,n}\right] &= o_{\mathrm{p}}(1). & (135)
\end{aligned}$$

We start with (133). Set

$$\widehat{\Delta}_{0,n}\left(y, w, x\right) = \Delta\left(y, w, x; \theta_0\right)\left[\frac{\widehat{v}_n^*}{\|\widehat{v}_n^*\|}\right].$$

and write

$$\nu_0\left[\widehat{\Delta}_n \cdot \widehat{T}_{2,n}\right] = \nu_0\left[\left(\widehat{\Delta}_n - \widehat{\Delta}_{0,n}\right) \cdot \widehat{T}_{2,n}\right] + \nu_0\left[\widehat{\Delta}_{0,n} \cdot \widehat{T}_{2,n}\right].$$

By (101) of Lemma C.12, for any function $T(w,x)$ that only depends on (w,x), $\nu_0\left[\widehat{\Delta}_{0,n} \cdot T\right] = 0$.

Hence, since $\widehat{T}_{2,n}$ only takes (w, x) as arguments, $\nu_0\left[\widehat{\Delta}_{0,n} \cdot \widehat{T}_{2,n}\right]$ so that

$$\left|\nu_0\left[\widehat{\Delta}_n \cdot \widehat{T}_{2,n}\right]\right| = \left|\nu_0\left[\left(\widehat{\Delta}_n - \widehat{\Delta}_{0,n}\right) \cdot \widehat{T}_{2,n}\right]\right| \leq \nu_0\left[\left(\widehat{\Delta}_n - \widehat{\Delta}_{0,n}\right)^2\right]^{1/2} \nu_0\left[\widehat{T}^2_{2,n}\right]^{1/2},$$

where the inequality follows by Cauchy-Schwarz. Using the fact that $\left|\widehat{T}_{2,n}\right| \leq \overline{T}^*$,

$$\left|\nu_0\left[\widehat{\Delta}_n \cdot \widehat{T}_{2,n}\right]\right| \leq \nu_0\left[\left(\overline{T}^*\right)^2\right]^{1/2} \nu_0\left[\left(\widehat{\Delta}_n - \widehat{\Delta}_{0,n}\right)^2\right]^{1/2}.$$

By (138) in Lemma D.4,

$$\begin{aligned}
&\left|\nu_0\left[\widehat{\Delta}_n \cdot \widehat{T}_{2,n}\right]\right| \\
&\leq 2\nu_0\left[\left(\overline{T}^*\right)^2\right]^{1/2} \nu_0\left[\max\left\{\overline{U}_{\Delta,1}, \overline{U}_{\Delta,2}\right\}^2\right]^{1/2} \times \\
&\times \left(\frac{\max\left\{\left\|\widehat{v}^*_{n,\alpha}\right\|_2, \left\|\widehat{v}^*_{n,h}\right\|_2\right\}}{\left\|\widehat{v}^*_n\right\|} \cdot \rho\left(\widehat{\theta}_n, \theta_0\right) + \frac{\max\left\{\left\|\widehat{v}^*_{n,\alpha}\right\|_2^3, \left\|v_h\right\|_2^3\right\}}{\left\|\widehat{v}^*_n\right\|^3} \rho\left(\widehat{\theta}_n, \theta_0\right)^2\right) \\
&\leq C_\Delta \max\left\{\rho\left(\widehat{\theta}_n, \theta_0\right), \rho\left(\widehat{\theta}_n, \theta_0\right)^2\right\}
\end{aligned}$$

for a constant $C_\Delta < \infty$, since both $\overline{U}_{\Delta,1}$ and $\overline{U}_{\Delta,2}$ are square integrable under Assumption 4.2. By Theorem 4.1, $\rho\left(\widehat{\theta}_n, \theta_0\right) = o_{\mathrm{p}}(1)$ and so, $\nu_0\left[\widehat{\Delta}_n \cdot \widehat{T}_{2,n}\right] = o_{\mathrm{p}}(1)$, i.e. (133) holds.

For (134), first write

$$\begin{aligned}
\left|\nu_n\left[\widehat{T}_{2,n}\right]\right| &= \left|\nu_0\left[\widehat{T}_{2,n}\right] + \mu_n\left[\widehat{T}_{2,n}\right]\right| \\
&\leq \left|\nu_0\left[\widehat{T}_{2,n}\right]\right| + \left|\mu_n\left[\widehat{T}_{2,n}\right]\right| \\
&\leq \left|\nu_0\left[\overline{T}^*\right]\right| + \left|\mu_n\left[\widehat{T}_{2,n}\right]\right| \\
&\leq \nu_0\left[\left(\overline{T}^*\right)^2\right]^{1/2} + \sup_{t_* \in \mathscr{T}_*} \left|\mu_n\left[t_*\right]\right| \\
&= \nu_0\left[\left(\overline{T}^*\right)^2\right]^{1/2} + o_{\mathrm{p}}(1)
\end{aligned}$$

The last negligibility claim is due to (128) and follows from $\mathscr{T}_*$ being a G_0-Donsker class. Therefore, $\nu_n\left[\widehat{T}_{2,n}\right] = O_{\mathrm{p}}(1)$ and (134) follows if we show that $\nu_n\left[\widehat{\Delta}_n\right] = o_{\mathrm{p}}(1)$. To that end, write

$$\nu_n\left[\widehat{\Delta}_n\right] = \nu_n\left[\widehat{\Delta}_{0,n}\right] + \nu_n\left[\widehat{\Delta}_n - \widehat{\Delta}_{0,n}\right]$$

First, by (138) in Lemma D.4,

$$\begin{aligned}\left|\nu_n\left[\widehat{\Delta}_n-\widehat{\Delta}_{0,n}\right]\right| \leq & \nu_n\left[\left(\widehat{\Delta}_n-\widehat{\Delta}_{0,n}\right)^2\right]^{1/2}\\ \leq & \, 2\nu_n\left[\max\left\{\overline{U}_{\Delta,1},\overline{U}_{\Delta,2}\right\}^2\right]^{1/2}\times\\ & \times\left(\frac{\max\left\{\left\|\widehat{v}^*_{n,\alpha}\right\|_2,\left\|\widehat{v}^*_{n,h}\right\|_2\right\}}{\left\|\widehat{v}^*_n\right\|}\cdot\rho\left(\widehat{\theta}_n,\theta_0\right)+\frac{\max\left\{\left\|\widehat{v}^*_{n,\alpha}\right\|_2^3,\left\|v_h\right\|_2^3\right\}}{\left\|\widehat{v}^*_n\right\|^3}\rho\left(\widehat{\theta}_n,\theta_0\right)^2\right)\\ \leq & \, \left(C_{\Delta,2}+o_{\mathrm{p}}(1)\right)\max\left\{\rho\left(\widehat{\theta}_n,\theta_0\right),\rho\left(\widehat{\theta}_n,\theta_0\right)^2\right\}\end{aligned}$$

for a constant $C_{\Delta,2}$ by the Strong Law of Large Numbers. By Theorem 4.1, $\rho\left(\widehat{\theta}_n,\theta_0\right)=o_{\mathrm{p}}(1)$ and so $\nu_n\left[\widehat{\Delta}_n-\widehat{\Delta}_{0,n}\right]=o_{\mathrm{p}}(1)$. Next, for each n, we again have $\nu_0\left[\widehat{\Delta}_{0,n}\right]=0$ by Lemma C.12 and by Lemma D.3, $\widehat{\Delta}_{0,n}(\cdot)$ are uniformly ν_0-integrable. By the Law of Large Numbers under uniform integrability (see for instance Theorem A.7.3 in Bickel et al. (1998)), we have

$$\nu_n\left[\widehat{\Delta}_{0,n}\right]=\mu_n\left[\widehat{\Delta}_{0,n}\right]\xrightarrow{\mathrm{p}}0.$$

And so, $\nu_n\left[\widehat{\Delta}_n\right]=\nu_n\left[\widehat{\Delta}_{0,n}\right]+\nu_n\left[\widehat{\Delta}_n-\widehat{\Delta}_{0,n}\right]=o_{\mathrm{p}}(1)+o_{\mathrm{p}}(1)=o_{\mathrm{p}}(1)$, and so, (134) holds.

Finally (135) follows from the previous arguments and Theorem 2.10.5 in van der Vaart and Wellner (2023) (an element-wise product of Glivenko-Cantelli classes with integrable envelopes is itself Glivenko-Cantelli). □

D.6 Envelopes for score functions and their differences

In what follows, let $z=(y,w,x)$ and $\theta=(\alpha,h)$ for brevity. Let

$$\begin{aligned}\widetilde{\Delta}_\alpha(z;\theta)&=\int\frac{\partial}{\partial\alpha}\kappa(z;\alpha,b)h(b)^2\,\mathrm{d}b,\\ \widetilde{\Delta}_h(z;\theta)[v_h]&=\int\kappa(z;\alpha,b)h(b)v_h(b)\mathrm{d}b,\end{aligned}\tag{136}$$

so that (by (72)) Δ can be written as

$$\Delta\left(z;\theta\right)[v]=\frac{\widetilde{\Delta}_\alpha\left(z;\theta\right)'v_\alpha+2\widetilde{\Delta}_h(z;\theta)\left[v_h\right]}{P(z;\theta)}.$$

Now, we provide results concerning envelope functions for the pathwise derivative Δ. All proofs of lemmas stated here are given in Appendix D.7.

Lemma D.3. *Given any $\theta\in\Theta$ and $v\in\mathcal{V}$,*

$$|\Delta(z;\theta)[v]|\leq U_\Delta(z)\cdot\max\left\{\left\|v_\alpha\right\|_2,\left\|v_h\right\|_2\right\}.\tag{137}$$

where

$$U_{\Delta}(z) = \exp\left(\overline{\ell}^{*}(z)\right) \cdot \left[U_{\alpha}(z) + 2\right],$$

with the envelope function U_α given in Lemma D.5 and $\overline{\ell}^*$ and U_P are defined in Lemmas B.7 and B.8 respectively.

Lemma D.4. *Given any $\theta_1, \theta_2 \in \Theta$ and $v \in \mathcal{V}$,*

$$\begin{aligned} \left|\Delta\left(z;\theta_1\right)[v] - \Delta\left(z;\theta_2\right)[v]\right| \leq\ & \overline{U}_{\Delta,1}(z) \cdot \max\left\{\|v_\alpha\|_2, \|v_h\|_2\right\} \cdot \rho\left(\theta_1,\theta_2\right) \\ & + \overline{U}_{\Delta,2}(z) \cdot \max\left\{\|v_\alpha\|_2^3, \|v_h\|_2^3\right\} \rho\left(\theta_1,\theta_2\right)^2 \end{aligned} \tag{138}$$

where

$$\begin{aligned} \overline{U}_{\Delta,1}(z) &= \widetilde{U}_{\alpha,1}(z) + 2\widetilde{U}_{h,1}(z), \\ \overline{U}_{\Delta,2}(z) &= \widetilde{U}_{\alpha,2}(z) + 2\widetilde{U}_{h,2}(z), \\ \widetilde{U}_{\alpha,1}(z) &= \exp\left(\overline{\ell}^{*}(z)\right) \cdot \overline{U}_{\alpha}(z) + \exp\left(2\overline{\ell}^{*}(z)\right) \cdot U_{\alpha}(z) \cdot U_P(z), \\ \widetilde{U}_{\alpha,2}(z) &= \exp\left(2\overline{\ell}^{*}(z)\right) \cdot \max\left\{\frac{1}{2}, \exp\left(\overline{\ell}^{*}(z)\right) \cdot U_{\alpha}(z)\right\} \cdot \left[\overline{U}_{\alpha}(z)^2 + U_P(z)^2\right], \\ \widetilde{U}_{h,1}(z) &= \exp\left(\overline{\ell}^{*}(z)\right) \cdot \overline{U}_{h}(z) + \exp\left(2\overline{\ell}^{*}(z)\right) \cdot U_P(z), \\ \widetilde{U}_{h,2}(z) &= \exp\left(2\overline{\ell}^{*}(z)\right) \cdot \max\left\{\frac{1}{2}, \exp\left(\overline{\ell}^{*}(z)\right)\right\} \cdot \left[\overline{U}_{h}(z)^2 + U_P(z)^2\right]. \end{aligned}$$

and the envelope functions U_α, $\overline{U}_\alpha$, $\overline{U}_h$, $\overline{\ell}^*$ and U_P are defined in in Lemmas D.5, D.6, D.8, B.7 and B.8 respectively.

The bounds in this previous two lemmas are a combination of the following.

Lemma D.5. *For any $\theta = (\alpha, h) \in \Theta$*

$$\left\|\widetilde{\Delta}_{\alpha}(z;\theta)\right\|_2 \leq U_{\alpha}(z). \tag{139}$$

where

$$U_{\alpha}(y,w,x) = 2(J+1) \max_{j=0,\ldots,J} \left\{\|w_j\|_2\right\}.$$

Lemma D.6. *Given $\theta_1, \theta_2 \in \Theta$,*

$$\left\|\widetilde{\Delta}_{\alpha}\left(z;\theta_1\right) - \widetilde{\Delta}_{\alpha}\left(z;\theta_2\right)\right\|_2 \leq \overline{U}_{\alpha}(z) \cdot \rho\left(\theta_1,\theta_2\right). \tag{140}$$

where

$$\overline{U}_{\alpha}(y,w,x) = \left[8(J+1)^2 \max_{j=0,\ldots,J} \left\{\|w_j\|_2^2\right\} + 4(J+1) \max_{j=0,\ldots,J} \left\{\|w_j\|_2\right\}\right].$$

Lemma D.7. *Given $\theta \in \Theta$ and $v_h \in \mathscr{L}_2$,*

$$\left|\widetilde{\Delta}_{h}(z;\theta)[v_h]\right| \leq \left(\int v_h(b)^2 \mathrm{d}b\right)^{1/2}.$$

Lemma D.8. *Given $\theta_1, \theta_2 \in \Theta$ with $\theta_j = (\alpha_j, h_j)$, and $v_h \in \mathscr{L}_2$,*

$$\left|\widetilde{\Delta}_h\left(z;\theta_1\right)\left[v_h\right] - \widetilde{\Delta}_h\left(z;\theta_2\right)\left[v_h\right]\right| \le \overline{U}_h(z) \cdot \left(\int v_h(b)^2 \mathrm{d}b\right)^{1/2} \cdot \rho\left(\theta_1, \theta_2\right). \tag{141}$$

where

$$\overline{U}_h(y,w,x) = \max\left\{1, 2(J+1)\max_{j=0,\ldots,J}\left\{\left\|w_j\right\|_2\right\}\right\}.$$

D.7 Proofs of envelopes for score functions and their differences

Proof of Lemma D.3. Use (72) and (136) to write

$$\begin{aligned}
|\Delta(z;\theta)[v]| &= \frac{\left|\widetilde{\Delta}_\alpha(z;\theta)' v_\alpha + 2\widetilde{\Delta}_h(z;\theta)[v_h]\right|}{P(z;\theta)} \\
&\le \exp\left(\overline{\ell}^*(z)\right)\left|\widetilde{\Delta}_\alpha(z;\theta)' v_\alpha + 2\widetilde{\Delta}_h(z;\theta)[v_h]\right| \\
&\le \exp\left(\overline{\ell}^*(z)\right)\left\{\left|\widetilde{\Delta}_\alpha(z;\theta)' v_\alpha\right| + 2\left|\widetilde{\Delta}_h(z;\theta)[v_h]\right|\right\} \\
&\le \exp\left(\overline{\ell}^*(z)\right)\left\{\left\|\widetilde{\Delta}_\alpha(z;\theta)\right\|_2 \left\|v_\alpha\right\|_2 + 2\left|\widetilde{\Delta}_h(z;\theta)[v_h]\right|\right\} \\
&\le \exp\left(\overline{\ell}^*(z)\right)\left\{U_\alpha(z)\left\|v_\alpha\right\|_2 + 2\left\|v_h\right\|_2\right\} \\
&\le \exp\left(\overline{\ell}^*(z)\right)\left[U_\alpha(z) + 2\right] \cdot \max\left\{\left\|v_\alpha\right\|_2, \left\|v_h\right\|_2\right\}.
\end{aligned}$$

The penultimate inequality follows from Lemmas D.5 and D.7. Hence, (137) follows. □

Proof of Lemma D.4. Using (72) and (136),

$$\begin{aligned}
&\Delta\left(z;\theta_1\right)[v] - \Delta\left(z;\theta_2\right)[v] \\
&= \left(\frac{\widetilde{\Delta}_\alpha\left(z;\theta_1\right)' v_\alpha}{P\left(z;\theta_1\right)} - \frac{\widetilde{\Delta}_\alpha\left(z;\theta_2\right)' v_\alpha}{P\left(z;\theta_2\right)}\right) + 2\left(\frac{\widetilde{\Delta}_h\left(z;\theta_1\right)\left[v_h\right]}{P\left(z;\theta_1\right)} - \frac{\widetilde{\Delta}_h\left(z;\theta_2\right)\left[v_h\right]}{P\left(z;\theta_2\right)}\right).
\end{aligned}$$

By the triangle inequality,

$$\begin{aligned}
&\left|\Delta\left(z;\theta_1\right)[v] - \Delta\left(z;\theta_2\right)[v]\right| \\
&\le \left|\frac{\widetilde{\Delta}_\alpha\left(z;\theta_1\right)' v_\alpha}{P\left(z;\theta_1\right)} - \frac{\widetilde{\Delta}_\alpha\left(z;\theta_2\right)' v_\alpha}{P\left(z;\theta_2\right)}\right| + 2\left|\frac{\widetilde{\Delta}_h\left(z;\theta_1\right)\left[v_h\right]}{P\left(z;\theta_1\right)} - \frac{\widetilde{\Delta}_h\left(z;\theta_2\right)\left[v_h\right]}{P\left(z;\theta_2\right)}\right|.
\end{aligned} \tag{142}$$

By (161) in Lemma F.5,

$$\begin{aligned}
\left|\frac{a_1}{b_1} - \frac{a_2}{b_2}\right| \le &\frac{1}{b_2}\left|a_1 - a_2\right| + \frac{\left|a_2\right|}{b_2^2}\left|b_1 - b_2\right| \\
&+ \frac{1}{\min\left\{b_1, b_2\right\}^2} \cdot \max\left\{\frac{1}{2}, \left|\frac{a_2}{b_2}\right|\right\}\left[\left|a_1 - a_2\right|^2 + \left|b_1 - b_2\right|^2\right].
\end{aligned}$$

Applying the above to the first summand in (142)

$$
\begin{aligned}
&\left|\frac{\widetilde{\Delta}_\alpha(z;\theta_1)' v_\alpha}{P(z;\theta_1)} - \frac{\widetilde{\Delta}_\alpha(z;\theta_2)' v_\alpha}{P(z;\theta_2)}\right| \\
&\leq \frac{1}{P(z;\theta_2)} \left|\left[\widetilde{\Delta}_\alpha(z;\theta_1) - \widetilde{\Delta}_\alpha(z;\theta_2)\right]' v_\alpha\right| \\
&\quad + \frac{\left|\widetilde{\Delta}_\alpha(z;\theta_2)' v_\alpha\right|}{P(z;\theta_2)^2} \left|P(z;\theta_1) - P(z;\theta_2)\right| \\
&\quad + \frac{1}{\min_{j=1,2}\{P(z;\theta_j)\}^2} \cdot \max\left\{\frac{1}{2}, \left|\frac{\widetilde{\Delta}_\alpha(z;\theta_2)' v_\alpha}{P(z;\theta_2)}\right|\right\} \times \\
&\quad \times \left\{\left|\left[\widetilde{\Delta}_\alpha(z;\theta_1) - \widetilde{\Delta}_\alpha(z;\theta_2)\right]' v_\alpha\right|^2 + \left|P(z;\theta_1) - P(z;\theta_2)\right|^2\right\} \\
&\leq \frac{1}{P(z;\theta_2)} \left\|\widetilde{\Delta}_\alpha(z;\theta_1) - \widetilde{\Delta}_\alpha(z;\theta_2)\right\|_2 \|v_\alpha\|_2 \\
&\quad + \frac{\left\|\widetilde{\Delta}_\alpha(z;\theta_2)\right\|_2 \|v_\alpha\|_2}{P(z;\theta_2)^2} \left|P(z;\theta_1) - P(z;\theta_2)\right| \\
&\quad + \frac{1}{\min_{j=1,2}\{P(z;\theta_j)\}^2} \cdot \max\left\{\frac{1}{2}, \frac{\left\|\widetilde{\Delta}_\alpha(z;\theta_2)\right\|_2 \|v_\alpha\|_2}{P(z;\theta_2)}\right\} \times \\
&\quad \times \left\{\left\|\widetilde{\Delta}_\alpha(z;\theta_1) - \widetilde{\Delta}_\alpha(z;\theta_2)\right\|_2^2 \|v_\alpha\|_2^2 + \left|P(z;\theta_1) - P(z;\theta_2)\right|^2\right\}.
\end{aligned}
$$

Using the envelope $\overline{\ell}^*$ defined in (31) of Lemma B.7,

$$
\begin{aligned}
&\left|\frac{\widetilde{\Delta}_\alpha(z;\theta_1)' v_\alpha}{P(z;\theta_1)} - \frac{\widetilde{\Delta}_\alpha(z;\theta_2)' v_\alpha}{P(z;\theta_2)}\right| \\
&\leq \exp\left(\overline{\ell}^*(z)\right) \left\|\widetilde{\Delta}_\alpha(z;\theta_1) - \widetilde{\Delta}_\alpha(z;\theta_2)\right\|_2 \|v_\alpha\|_2 \\
&\quad + \exp\left(2\overline{\ell}^*(z)\right) \left\|\widetilde{\Delta}_\alpha(z;\theta_2)\right\|_2 \|v_\alpha\|_2 \left|P(z;\theta_1) - P(z;\theta_2)\right| \\
&\quad + \exp\left(2\overline{\ell}^*(z)\right) \cdot \max\left\{\frac{1}{2}, \exp\left(\overline{\ell}^*(z)\right) \left\|\widetilde{\Delta}_\alpha(z;\theta_2)\right\|_2 \|v_\alpha\|_2\right\} \times \\
&\quad \times \left\{\left\|\widetilde{\Delta}_\alpha(z;\theta_1) - \widetilde{\Delta}_\alpha(z;\theta_2)\right\|_2^2 \|v_\alpha\|_2^2 + \left|P(z;\theta_1) - P(z;\theta_2)\right|^2\right\}.
\end{aligned}
$$

Using the definitions of the envelope functions $U_\alpha, \overline{U}_\alpha$ and U_P in Lemma D.5, Lemma D.6 and Lemma B.8 respectively, we get

$$
\begin{aligned}
&\left|\frac{\widetilde{\Delta}_\alpha(z;\theta_1)' v_\alpha}{P(z;\theta_1)} - \frac{\widetilde{\Delta}_\alpha(z;\theta_2)' v_\alpha}{P(z;\theta_2)}\right| \\
&\leq \widetilde{U}_{\alpha,1}(z) \cdot \|v_\alpha\|_2 \cdot \rho(\theta_1,\theta_2) + \widetilde{U}_{\alpha,2}(z) \cdot \max\left\{1, \|v_\alpha\|_2^3\right\} \cdot \rho(\theta_1,\theta_2)^2.
\end{aligned}
$$

Next, we proceed similarly and apply (161) in Lemma F.5 to the second summand in (142):

$$\begin{aligned}
&\left|\frac{\widetilde{\Delta}_h\left(z;\theta_1\right)\left[v_h\right]}{P\left(z;\theta_1\right)}-\frac{\widetilde{\Delta}_h\left(z;\theta_2\right)\left[v_h\right]}{P\left(z;\theta_2\right)}\right| \\
\leq\ & \frac{1}{P\left(z;\theta_2\right)}\left|\widetilde{\Delta}_h\left(z;\theta_1\right)\left[v_h\right]-\widetilde{\Delta}_h\left(z;\theta_2\right)\left[v_h\right]\right| \\
&+\frac{\left|\widetilde{\Delta}_h\left(z;\theta_2\right)\left[v_h\right]\right|}{P\left(z;\theta_2\right)^2}\left|P\left(z;\theta_1\right)-P\left(z;\theta_2\right)\right| \\
&+\frac{1}{\min_{j=1,2}\left\{P\left(z;\theta_j\right)\right\}^2}\cdot\max\left\{\frac{1}{2},\left|\frac{\widetilde{\Delta}_h\left(z;\theta_2\right)\left[v_h\right]}{P\left(z;\theta_2\right)}\right|\right\}\times \\
&\times\left[\left|\widetilde{\Delta}_h\left(z;\theta_1\right)\left[v_h\right]-\widetilde{\Delta}_h\left(z;\theta_2\right)\left[v_h\right]\right|^2+\left|P\left(z;\theta_1\right)-P\left(z;\theta_2\right)\right|^2\right].
\end{aligned}$$

Using the envelope $\overline{\ell}^*$ defined in (31) of Lemma B.7,

$$\begin{aligned}
&\left|\frac{\widetilde{\Delta}_h\left(z;\theta_1\right)\left[v_h\right]}{P\left(z;\theta_1\right)}-\frac{\widetilde{\Delta}_h\left(z;\theta_2\right)\left[v_h\right]}{P\left(z;\theta_2\right)}\right| \\
\leq\ & \exp\left(\overline{\ell}^*(z)\right)\left|\widetilde{\Delta}_h\left(z;\theta_1\right)\left[v_h\right]-\widetilde{\Delta}_h\left(z;\theta_2\right)\left[v_h\right]\right| \\
&+\exp\left(2\overline{\ell}^*(z)\right)\cdot\left|\widetilde{\Delta}_h\left(z;\theta_2\right)\left[v_h\right]\right|\cdot\left|P\left(z;\theta_1\right)-P\left(z;\theta_2\right)\right| \\
&+\exp\left(2\overline{\ell}^*(z)\right)\cdot\max\left\{\frac{1}{2},\exp\left(\overline{\ell}^*(z)\right)\cdot\left|\widetilde{\Delta}_h\left(z;\theta_2\right)\left[v_h\right]\right|\right\}\times \\
&\times\left[\left|\widetilde{\Delta}_h\left(z;\theta_1\right)\left[v_h\right]-\widetilde{\Delta}_h\left(z;\theta_2\right)\left[v_h\right]\right|^2+\left|P\left(z;\theta_1\right)-P\left(z;\theta_2\right)\right|^2\right].
\end{aligned}$$

Using Lemma D.7 and the definitions of the envelope functions $\overline{U}_h$ and U_P in Lemma D.8 and Lemma B.8 respectively, we get

$$\begin{aligned}
&\left|\frac{\widetilde{\Delta}_h\left(z;\theta_1\right)\left[v_h\right]}{P\left(z;\theta_1\right)}-\frac{\widetilde{\Delta}_h\left(z;\theta_2\right)\left[v_h\right]}{P\left(z;\theta_2\right)}\right| \\
\leq\ & \widetilde{U}_{h,1}(z)\cdot\left\|v_h\right\|_2\cdot\rho\left(\theta_1,\theta_2\right)+\widetilde{U}_{h,2}(z)\cdot\max\left\{1,\left\|v_h\right\|_2^3\right\}\cdot\rho\left(\theta_1,\theta_2\right)^2.
\end{aligned}$$

And so, (138) follows from combining Combining (D.7) and (D.7) with (142). □

Proof of Lemma D.5. Lemma F.1 derives the following in (151):

$$\frac{\partial}{\partial\alpha}\kappa(y,w,x;\alpha,b)=\kappa(y,w,x;\alpha,b)\cdot\sum_{j=0}^{J}\left(w_y-w_j\right)\kappa(j,w,x;\alpha,b).$$

As a result, the triangle inequality and $\kappa(\cdot) \in (0,1)$ provide the following norm bound

$$\left\| \frac{\partial}{\partial \alpha} \kappa(y, w, x; \alpha, b) \right\|_2 \leq 2(J+1) \max_{j=0,\ldots,J} \left\{ \|w_j\|_2 \right\}. \tag{143}$$

From (136)

$$\begin{aligned}
\left\| \widetilde{\Delta}_\alpha(z;\theta) \right\|_2 &= \left\| \int \frac{\partial}{\partial \alpha} \kappa(z;\alpha,b) h(b)^2 \, \mathrm{d}b \right\|_2 \\
&\leq \int \left\| \frac{\partial}{\partial \alpha} \kappa(z;\alpha,b) \right\|_2 h(b)^2 \, \mathrm{d}b \\
&\leq 2(J+1) \max_{j=0,\ldots,J} \left\{ \|w_j\|_2 \right\} \int h(b)^2 \, \mathrm{d}b,
\end{aligned}$$

and so (139) follows from $\int h^2 = 1$. □

Proof of Lemma D.6. Let $\theta_j = (\alpha_j, h_j)$ for $j \in \{1,2\}$. The difference can be decomposed

$$\begin{aligned}
&\widetilde{\Delta}_\alpha\left(z;\theta_1\right) - \widetilde{\Delta}_\alpha\left(z;\theta_2\right) \\
&= \int \frac{\partial}{\partial \alpha} \kappa\left(z;\alpha_1,b\right) h_1(b)^2 \, \mathrm{d}b - \int \frac{\partial}{\partial \alpha} \kappa\left(z;\alpha_2,b\right) h_2(b)^2 \, \mathrm{d}b \\
&= \int \left[\frac{\partial}{\partial \alpha} \kappa\left(z;\alpha_1,b\right) - \frac{\partial}{\partial \alpha} \kappa\left(z;\alpha_2,b\right) \right] h_1(b)^2 \, \mathrm{d}b \\
&\quad + \int \frac{\partial}{\partial \alpha} \kappa\left(z;\alpha_2,b\right) \left(h_1(b)^2 - h_2(b)^2 \right) \, \mathrm{d}b \\
&= \left[\int \frac{\partial^2}{\partial \alpha \partial \alpha'} \kappa\left(z;\widetilde{\alpha},b\right) h_1(b)^2 \, \mathrm{d}b \right] \left(\alpha_1 - \alpha_2\right) \\
&\quad + \int \frac{\partial}{\partial \alpha} \kappa\left(z;\alpha_2,b\right) \left(h_1(b) + h_2(b)\right) \left(h_1(b) - h_2(b)\right) \, \mathrm{d}b
\end{aligned}$$

for a midpoint $\widetilde{\alpha} = \widetilde{\alpha}(z,b)$ between α_1 and α_2. Using the triangle inequality,

$$\begin{aligned}
&\left\| \widetilde{\Delta}_\alpha\left(z;\theta_1\right) - \widetilde{\Delta}_\alpha\left(z;\theta_2\right) \right\|_2 \\
&\leq \left\| \int \frac{\partial^2}{\partial \alpha \partial \alpha'} \kappa\left(z;\widetilde{\alpha},b\right) h_1(b)^2 \, \mathrm{d}b \left(\alpha_1 - \alpha_2\right) \right\|_2 \\
&\quad + \left\| \int \frac{\partial}{\partial \alpha} \kappa\left(z;\alpha_2,b\right) \left(h_1(b) + h_2(b)\right) \left(h_1(b) - h_2(b)\right) \, \mathrm{d}b \right\|_2 \\
&\leq \left[\int \left\| \frac{\partial^2}{\partial \alpha \partial \alpha'} \kappa\left(z;\widetilde{\alpha},b\right) \right\|_{\mathrm{op}} h_1(b)^2 \, \mathrm{d}b \right] \left\| \alpha_1 - \alpha_2 \right\|_2 \\
&\quad + \left(\int \left\| \frac{\partial}{\partial \alpha} \kappa\left(z;\alpha_2,b\right) \right\|_2^2 \left(h_1(b) + h_2(b)\right)^2 \, \mathrm{d}b \right)^{1/2} \left(\int \left(h_1(b) - h_2(b)\right)^2 \, \mathrm{d}b \right)^{1/2}
\end{aligned}$$

where $\|\cdot\|_{\mathrm{op}}$ denotes the operator (or spectral) norm of a matrix. As in (143),

$$\left\| \frac{\partial}{\partial \alpha} \kappa(z;\alpha,b) \right\|_2 \leq 2(J+1) \max_{j=0,\ldots,J} \left\{ \|w_j\|_2 \right\}.$$

Furthermore, the second derivative is

$$\begin{aligned} \frac{\partial}{\partial \alpha \partial \alpha'} \kappa(z;\alpha,b) &= \frac{\partial}{\partial \alpha} \kappa(z;\alpha,b) \cdot \sum_{j=0}^{J} (w_y - w_j)' \, \kappa(j,w,x;\alpha,b) \\ &\quad + \kappa(z;\alpha,b) \cdot \sum_{j=0}^{J} (w_y - w_j) \frac{\partial}{\partial \alpha'} \kappa(z;\alpha,b). \end{aligned} \tag{144}$$

It can be shown that given any two vectors v_1, v_2, the operator norm of their outer product is bounded by the product of their Euclidean norms, i.e. $\|v_1 v_2'\|_{\mathrm{op}} \leq \|v_1\|_2 \, \|v_2\|_2$. Repeated application of this to (144) gives

$$\left\| \frac{\partial}{\partial \alpha \partial \alpha'} \kappa(y,w,x;\alpha,b) \right\|_{\mathrm{op}} \leq 8(J+1)^2 \max_{j=0,\ldots,J} \left\{ \|w_j\|_2^2 \right\}. \tag{145}$$

Both norm bounds (143) and (145) do not depend on y, x, α, b. Thus,

$$\begin{aligned} &\left\| \widetilde{\Delta}_\alpha (z;\theta_1) - \widetilde{\Delta}_\alpha (z;\theta_2) \right\|_2 \\ \leq &\left[\int \left\| \frac{\partial^2}{\partial \alpha \partial \alpha'} \kappa\left(z;\widetilde{\alpha},b\right) \right\|_{\mathrm{op}} h_1(b)^2 \, \mathrm{d}b \right] \|\alpha_1 - \alpha_2\|_2 \\ &+ \left(\int \left\| \frac{\partial}{\partial \alpha} \kappa\left(z;\alpha_2,b\right) \right\|_2^2 \left(h_1(b) + h_2(b)\right)^2 \, \mathrm{d}b \right)^{1/2} \left(\int \left(h_1(b) - h_2(b)\right)^2 \, \mathrm{d}b \right)^{1/2} \\ &+ \left(\int \left\| \frac{\partial}{\partial \alpha} \kappa\left(z;\alpha_2,b\right) \right\|_2^2 \left(h_1(b) + h_2(b)\right)^2 \, \mathrm{d}b \right)^{1/2} \rho_{\mathscr{L}_2}\left(h_1, h_2\right) \\ \leq &\, 8(J+1)^2 \max_{j=0,\ldots,J} \left\{ \|w_j\|_2^2 \right\} \cdot \|\alpha_1 - \alpha_2\|_2 \\ &+ 2(J+1) \max_{j=0,\ldots,J} \left\{ \|w_j\|_2 \right\} \left(\int \left(h_1(b) + h_2(b)\right)^2 \, \mathrm{d}b \right)^{1/2} \rho_{\mathscr{L}_2}\left(h_1, h_2\right) \end{aligned}$$

5 Using the inequality $\int (h_1 + h_2)^2 \leq 4$ due to Lemma B.6 and the fact that both $\|\alpha_1 - \alpha_2\|_2 \leq \rho\left(\theta_1, \theta_2\right)$ and $\rho_{\mathscr{L}_2}\left(h_1, h_2\right) \leq \rho\left(\theta_1, \theta_2\right)$, we get

$$\begin{aligned} &\left\| \widetilde{\Delta}_\alpha (z;\theta_1) - \widetilde{\Delta}_\alpha (z;\theta_2) \right\|_2 \\ \leq &\left[8(J+1)^2 \max_{j=0,\ldots,J} \left\{ \|w_j\|_2^2 \right\} + 4(J+1) \max_{j=0,\ldots,J} \left\{ \|w_j\|_2 \right\} \right] \rho\left(\theta_1, \theta_2\right), \end{aligned}$$

which is exactly (140). □

Proof of Lemma D.7. By $\kappa(\cdot) \in (0,1)$ and $\int h^2 = 1$,

$$\begin{aligned}\left|\widetilde{\Delta}_h(z;\theta)[v_h]\right| = \left|\int \kappa(z;\alpha,b)h(b)v_h(b)\mathrm{d}b\right| &\leq \int |h(b)|\,|v_h(b)|\,\mathrm{d}b\\ &\leq \left(\int h(b)^2\mathrm{d}b\right)^{1/2}\left(\int v_h(b)^2\mathrm{d}b\right)^{1/2} = \left(\int v_h(b)^2\mathrm{d}b\right)^{1/2}.\end{aligned}$$

□

Proof of Lemma D.8. Decompose the difference as

$$\begin{aligned}&\widetilde{\Delta}_h\left(z;\theta_1\right)\left[v_h\right] - \widetilde{\Delta}_h\left(z;\theta_2\right)\left[v_h\right]\\ &= \int \kappa\left(z;\alpha_1,b\right)h_1(b)v_h(b)\mathrm{d}b - \int \kappa\left(z;\alpha_2,b\right)h_2(b)v_h(b)\mathrm{d}b\\ &= \int \left[\kappa\left(z;\alpha_1,b\right) - \kappa\left(z;\alpha_2,b\right)\right]h_1(b)v_h(b)\mathrm{d}b\\ &\quad + \int \kappa\left(z;\alpha_2,b\right)\left[h_1(b) - h_2(b)\right]v_h(b)\mathrm{d}b\\ &= \left[\int \frac{\partial}{\partial\alpha'}\kappa\left(z;\widetilde{\alpha}_1,b\right)h_1(b)v_h(b)\mathrm{d}b\right]\left(\alpha_1 - \alpha_2\right)\\ &\quad + \int \kappa\left(z;\alpha_2,b\right)\left[h_1(b) - h_2(b)\right]v_h(b)\mathrm{d}b\end{aligned}$$

By the triangle and Cauchy-Schwarz inequalities as well as the fact that $\kappa(\cdot) \in (0,1)$,

$$\begin{aligned}&\left|\widetilde{\Delta}_h\left(z;\theta_1\right)\left[v_h\right] - \widetilde{\Delta}_h\left(z;\theta_2\right)\left[v_h\right]\right|\\ &\leq \left[\int \left\|\frac{\partial}{\partial\alpha}\kappa\left(z;\widetilde{\alpha}_1,b\right)\right\|_2 |h_1(b)|\,|v_h(b)|\,\mathrm{d}b\right]\left\|\alpha_1 - \alpha_2\right\|_2\\ &\quad + \left(\int \left(h_1(b) - h_2(b)\right)^2\mathrm{d}b\right)^{1/2}\left(\int v_h(b)^2\mathrm{d}b\right)^{1/2}\\ &\leq 2(J+1)\max_{j=0,\ldots,J}\left\{\left\|w_j\right\|_2\right\}\left(\int v_h(b)^2\mathrm{d}b\right)^{1/2}\left\|\alpha_1 - \alpha_2\right\|_2\\ &\quad + \left(\int v_h(b)^2\mathrm{d}b\right)^{1/2}\cdot\rho_{\mathscr{L}_2}\left(h_1,h_2\right).\end{aligned}$$

In the above, the last inequality follows from (143), $\int h_1^2 = 1$ and the Cauchy Schwarz inequality. Next, using the fact that both $\left\|\alpha_1 - \alpha_2\right\|_2 \leq \rho\left(\theta_1,\theta_2\right)$ and $\rho_{\mathscr{L}_2}\left(h_1,h_2\right) \leq \rho\left(\theta_1,\theta_2\right)$

$$\begin{aligned}&\left|\widetilde{\Delta}_h\left(z;\theta_1\right)\left[v_h\right] - \widetilde{\Delta}_h\left(z;\theta_2\right)\left[v_h\right]\right|\\ &\leq \max\left\{1, 2(J+1)\max_{j=0,\ldots,J}\left\{\left\|w_j\right\|_2\right\}\right\}\left(\int v_h(b)^2\mathrm{d}b\right)^{1/2}\cdot\rho\left(\theta_1,\theta_2\right),\end{aligned}$$

which is exactly (141). □

E Proofs of Theorems 4.4 and 4.5

Proof of Theorem 4.4. Let $\{h_{j,l} : j = 1, 2, l = 1, \ldots, d_X\}$ be a finite set of univariate root-densities, i.e. $\int h_{j,l}\left(b_l\right)^2 \mathrm{d}b_l = 1$. Without loss of generality, assume also that $h_{j,l} \geq 0$ everywhere. Denote the j^{th} product root-density by $h_j = \prod_{l=1}^{d_X} h_{j,l}$. Then, by Lemma 3.3.10 (i) in Reiss (1989, p. 100)

$$\begin{aligned}\rho_{\mathscr{L}_2}\left(h_1, h_2\right)^2 = \sqrt{\int \left(h_1(b) - h_2(b)\right)^2 \mathrm{d}b} &\leq \sqrt{\sum_{l=1}^{d_X} \int \left(h_{1,l}\left(b_l\right) - h_{2,l}\left(b_l\right)\right)^2 \mathrm{d}b_l} \\ &= \sqrt{\sum_{l=1}^{d_X} \rho_{\mathscr{L}_2}\left(h_{1,l}, h_{2,l}\right)^2}.\end{aligned}$$

In finite dimensions, the following holds: for any real a_l with $l = 1, \ldots, d_X$,

$$\sqrt{\sum_{l=1}^{d_X} \left|a_l\right|^2} \leq \sum_{l=1}^{d_X} \left|a_l\right|.$$

Combining with the previous display,

$$\rho_{\mathscr{L}_2}\left(h_1, h_2\right) \leq \sum_{l=1}^{d_X} \rho_{\mathscr{L}_2}\left(h_{1,l}, h_{2,l}\right). \tag{146}$$

By the univariate counterpart of Assumption 4.4 (ii), for each $l = 1, \ldots, d_X$, there is $\widetilde{\gamma}_{0,n,l}$ such that setting $\widetilde{h}_{0,n,l} = \gamma_{0,n,l}' \boldsymbol{\psi}_{K_n,l}$, $\rho_{\mathscr{L}_2}\left(\widetilde{h}_{0,n,l}, h_{0,l}\right) = O(K_n^{-s})$. As before, we can set $\widetilde{\gamma}_{0,n,l}$ equal to the $\mathscr{L}_2$ projection coefficients of $h_{0,l}$ onto the span of $\boldsymbol{\psi}_{K_n,l}$ and

$$\begin{aligned}\gamma_{0,n,l} &= \widetilde{\gamma}_{0,n,l} / \left\|\widetilde{\gamma}_{0,n,l}\right\|_2, \\ h_{0,n,l} &= \gamma_{0,n,l}' \boldsymbol{\psi}_{K_n,l}.\end{aligned}$$

Then, $\gamma_{0,n,l}'\gamma_{0,n,l} = 1$ and $\int h_{0,n,l}(b)^2 \mathrm{d}b = 1$. By Lemma B.14, $\rho_{\mathscr{L}_2}\left(h_{0,n,l}, h_{0,l}\right) = O(K_n^{-s})$ (the approximation rate is unchanged by normalization). Let

$$h_{0,n} = \prod_{l=1}^{d_X} h_{0,n,l}.$$

By (146),

$$\rho_{\mathscr{L}_2}\left(h_{0,n}, h_0\right) \leq \sum_{l=1}^{d_X} \rho_{\mathscr{L}_2}\left(h_{1,l}, h_{2,l}\right) = O\left(K_n^{-s}\right)$$

This characterizes the "bias" part of the convergence rate.

Repeat the remainder of the proof of Theorem 4.2 to see that the "variance" part is of order $O_{\mathrm{p}}\left(K_n/\sqrt{n}\right)$. □

Proof of Theorem 4.5. Given Theorem 4.4, let the sequence $\zeta_n \geq 1$ in (69) now be non-decreasing, slowly growing sequence and satisfy

$$\begin{aligned}&\zeta_n \nearrow \infty,\\ &\zeta_n^2 \cdot \max\left\{\sqrt{n} \cdot K_n^{-2s}, \frac{K_n}{\sqrt{n}}\right\} \to 0.\end{aligned}$$

The definition of "rate-local" spaces (70) is unchanged up to this new definition of ζ_n, i.e.

$$\begin{aligned}\mathcal{N}_{0,n} &= \left\{\theta \in \Theta : \rho\left(\theta, \theta_0\right) \leq \max\left\{K_n^{-s/d_X}, \sqrt{\frac{K_n}{n}}\right\} \cdot \zeta_n.\right\},\\ \mathcal{N}_n &= \mathcal{N}_{0,n} \cap \Theta_{K_n}.\end{aligned}$$

Repeat the proof of Theorem 4.3 with these new definitions. □

F Auxiliary Results

F.1 Proofs of envelope functions for the log-likelihood and choice probabilities

F.1.1 Proof of Lemma B.7

Proof of Lemma B.7. Lemma F.3 below shows that for any α and any $h : \mathcal{B} \to \mathbb{R}$ with $\int h(b)^2 \, \mathrm{d}b = 1$, the following bound holds

$$\begin{aligned}|\log P(y, w, x; \alpha, h)| \leq\ & \log(J+1) + 2\left(\sum_{j=0}^{J} \|w_j\|_2\right) \cdot \|\alpha\|_2\\ &+ 2\left(\sum_{j=0}^{J} \|x_j\|_2\right) \cdot \int \|b\|_2 h(b)^2 \mathrm{d}b.\end{aligned}$$

Since $\mathcal{A} \subseteq \mathbb{R}^{d_W}$ and $\mathcal{B} \subseteq \mathbb{R}^{d_X}$ are compact sets and $\int h^2 = 1$, (32) follows. □

F.1.2 Proof of Lemma B.8

Proof of Lemma B.8. Assume without loss of generality throughout that $h_1, h_2 \geq 0$ everywhere.

$$\begin{aligned}&\left|P\left(y, w, x; \alpha_1, h_1\right) - P\left(y, w, x; \alpha_2, h_2\right)\right|\\ =& \left|\int \kappa\left(y, w, x; \alpha_1, b\right) h_1(b)^2 \mathrm{d}b - \int \kappa(y, w, x; \alpha_2, b) h_2(b)^2 \mathrm{d}b\right|\\ =& \left|\int \left(\kappa\left(y, w, x; \alpha_1, b\right) - \kappa\left(y, w, x; \alpha_2, b\right)\right) h_1(b)^2 \mathrm{d}b + \int \kappa\left(y, w, x; \alpha_2, b\right)\left(h_1(b)^2 - h_2(b)^2\right) \mathrm{d}b\right|\\ \leq& \left|\int \left(\kappa\left(y, w, x; \alpha_1, b\right) - \kappa\left(y, w, x; \alpha_2, b\right)\right) h_1(b)^2 \mathrm{d}b\right| + \left|\int \kappa\left(y, w, x; \alpha_2, b\right)\left(h_1(b)^2 - h_2(b)^2\right) \mathrm{d}b\right|\end{aligned}$$

and so,

$$\begin{aligned}|P(y,w,x;\alpha_1,h_1)-P(y,w,x;\alpha_2,h_2)| \leq & \int |\kappa(y,w,x;\alpha_1,b)-\kappa(y,w,x;\alpha_2,b)|\, h_1(b)^2 \mathrm{d}b \\ & + \int \kappa(y,w,x;\alpha_2,b)\left|h_1(b)^2-h_2(b)^2\right| \mathrm{d}b.\end{aligned} \tag{147}$$

Applying mean value theorem to the kernel difference in the first component, there is a midpoint $\widetilde{\alpha}=\widetilde{\alpha}(y,w,x,b;\alpha_1,\alpha_2)$ between α_1 and α_2 such that

$$\kappa(y,w,x;\alpha_1,b)-\kappa(y,w,x;\alpha_2,b)=\kappa(y,w,x;\widetilde{\alpha},b)\left[\sum_{j=0}^{J}(w_y-w_j)\,\kappa(j,w,x;\widetilde{\alpha},b)\right]'(\alpha_1-\alpha_2),$$

where we have used the expression for $\partial\kappa/\partial\alpha$ in (151) from Lemma F.1 below. Thus,

$$|\kappa(y,w,x;\alpha_1,b)-\kappa(y,w,x;\alpha_2,b)| \leq 2\left[\sum_{j=0}^{J}\|w_j\|_2\right]\|\alpha_1-\alpha_2\|_2.$$

Hence, the first component in (147) can be bounded by

$$\int|\kappa(y,w,x;\alpha_1,b)-\kappa(y,w,x;\alpha_2,b)|\,h_1(b)^2\mathrm{d}b \leq \int\left(\left[2\sum_{j=0}^{J}\|w_j\|_2\right]\|\alpha_1-\alpha_2\|_2\right)h_1(b)^2\mathrm{d}b$$

and since $\int h(b)^2\mathrm{d}b=1$,

$$\int|\kappa(y,w,x;\alpha_1,b)-\kappa(y,w,x;\alpha_2,b)|\,h_1(b)^2\mathrm{d}b \leq 2\left[\sum_{j=1}^{J}\|w_j\|_2\right]\|\alpha_1-\alpha_2\|_2. \tag{148}$$

For the second component of (147), $\int\kappa(y,w,x;\alpha_2,b)\left|h_1(b)^2-h_2(b)^2\right|\mathrm{d}b$, note that

$$\begin{aligned}\int\kappa(y,w,x;\alpha_2,b)\left|h_1(b)^2-h_2(b)^2\right|\mathrm{d}b &= \int\kappa(y,w,x;\alpha_2,b)\,(h_1(b)+h_2(b))\,|h_1(b)-h_2(b)|\,\mathrm{d}b \\ \text{(Cauchy-Schwarz)} \qquad &\leq \left(\int\kappa(y,w,x;\alpha_2,b)^2\,(h_1(b)+h_2(b))^2\,\mathrm{d}b\right)^{\frac{1}{2}} \\ &\quad\times\left(\int(h_1(b)-h_2(b))^2\,\mathrm{d}b\right)^{\frac{1}{2}} \\ \text{(by } 0\leq\kappa(\cdot;\cdot)\leq 1\text{)} \qquad &\leq \left(\int(h_1(b)+h_2(b))^2\,\mathrm{d}b\right)^{\frac{1}{2}}\times\rho_{\mathscr{L}_2}(h_1,h_2) \\ \text{(by by Lemma B.6)} \qquad &\leq \left(2\int\left(h_1(b)^2+h_2(b)^2\right)\mathrm{d}b\right)^{\frac{1}{2}}\times\rho_{\mathscr{L}_2}(h_1,h_2).\end{aligned}$$

And so, since $\int h_j(b)^2 \, \mathrm{d}b = 1$ for each $j = 1, 2$,

$$\int \kappa\left(y, w, x; \alpha_2, b\right) \left|h_1(b)^2 - h_2(b)^2\right| \mathrm{d}b \leq 2 \cdot \rho_{\mathscr{L}_2}\left(h_1, h_2\right). \tag{149}$$

Combining (148) and (149) with (147),

$$\begin{aligned}
&\left|P\left(y, w, x; \alpha_1, h_1\right) - P\left(y, w, x; \alpha_2, h_2\right)\right| \\
\leq\ & 2\left[\sum_{j=0}^{J} \left\|w_j\right\|_2\right] \left\|\alpha_1 - \alpha_2\right\|_2 + 2\rho_{\mathscr{L}_2}\left(h_1, h_2\right) \\
\leq\ & 2 \max\left\{1, \sum_{j=0}^{J} \left\|w_j\right\|_2\right\} \left\{\left\|\alpha_1 - \alpha_2\right\|_2 + \rho_{\mathscr{L}_2}\left(\left|h_1\right|, \left|h_2\right|\right)\right\}.
\end{aligned}$$

(33) follows from the above display and the inequality: $u + v \leq \sqrt{2} \cdot \sqrt{u^2 + v^2}$ for any $u, v \geq 0$. This completes the proof. □

F.1.3 Proof of Lemma B.9

Proof of Lemma B.9. Write

$$\log P\left(y, w, x; \theta_1\right) - \log P\left(y, w, x; \theta_2\right) = \log\left(1 + \frac{P\left(y, w, x; \theta_1\right)}{P\left(y, w, x; \theta_2\right)} - 1\right).$$

By $u/(1+u) \leq \log(1+u) \leq u$ for $u > -1$, applied with $u = \left[P\left(y, w, x; \theta_1\right)/P\left(y, w, x; \theta_2\right)\right] - 1$,

$$\begin{aligned}
&\left|\log P\left(y, w, x; \theta_1\right) - \log P\left(y, w, x; \theta_2\right)\right| \\
\leq\ & \frac{1}{\min_{j=1,2}\left\{P\left(y, w, x; \theta_j\right)\right\}} \left|P\left(y, w, x; \theta_1\right) - P\left(y, w, x; \theta_2\right)\right|.
\end{aligned}$$

(35) and (36) now follow from Lemma B.7 and Lemma B.8. □

F.1.4 Additional results required for envelope construction

Lemma F.1 (Derivatives of kernel)**.** *Let*

$$\kappa(y, w, x; \alpha, b) = \frac{\exp\left(w_y'\alpha + x_y'b\right)}{\sum_{j=0}^{J} \exp\left(w_j'\alpha + x_j'b\right)}. \tag{150}$$

Then,

$$\frac{\partial}{\partial\alpha}\kappa(y,w,x;\alpha,b) = \kappa(y,w,x;\alpha,b)\cdot\sum_{j=0}^{J}(w_y - w_j)\,\kappa(j,w,x;\alpha,b), \tag{151}$$

$$\frac{\partial}{\partial b}\kappa(y,w,x;\alpha,b) = \kappa(y,w,x;\alpha,b)\cdot\sum_{j=0}^{J}(x_y - x_j)\,\kappa(j,w,x;\alpha,b). \tag{152}$$

Therefore,

$$\frac{\partial}{\partial\alpha}\log\kappa(y,w,x;\alpha,b) = \sum_{j=0}^{J}(w_y - w_j)\,\kappa(j,w,x;\alpha,b), \tag{153}$$

$$\frac{\partial}{\partial b}\log\kappa(y,w,x;\alpha,b) = \sum_{j=0}^{J}(x_y - x_j)\,\kappa(j,w,x;\alpha,b). \tag{154}$$

Proof of Lemma F.1. We can stack $z_j' = \left(w_j', x_j'\right)$ and $\zeta' = (\alpha', b')$ so that

$$\kappa(y,w,x;\alpha,b) = \kappa(y,z;\zeta) = \frac{\exp\left(z_y'\zeta\right)}{\sum_{j=0}^{J}\exp\left(z_j'\zeta\right)}.$$

Combine the quotient rule and chain rule to get

$$\begin{aligned}\frac{\partial}{\partial\zeta}\kappa(y,z;\zeta) &= \frac{\left(\sum_{j=0}^{J}\exp\left(z_j'\zeta\right)\right)z_y\exp\left(z_y'\zeta\right) - \exp\left(z_y'\zeta\right)\cdot\left(\sum_{j=0}^{J}z_j\exp\left(z_j'\zeta\right)\right)}{\left[\sum_{j=0}^{J}\exp\left(z_j'\zeta\right)\right]^2}\\ &= \kappa(y,z;\zeta)\cdot\frac{\sum_{j=0}^{J}(z_y - z_j)\exp\left(z_j'\zeta\right)}{\sum_{j=0}^{J}\exp\left(z_j'\zeta\right)}\\ &= \kappa(y,z;\zeta)\cdot\sum_{j=0}^{J}(z_y - z_j)\,\kappa(j,z;\zeta).\end{aligned}$$

This proves both (151) and (152), which imply (153) and (154) respectively by the chain rule. □

Lemma F.2. *Let $\kappa(\cdot;\cdot)$ be defined by (150). For any (y,w,x) and any $\alpha\in\mathcal{A}$, $b\in\mathcal{B}$,*

$$0 > \log\kappa(y,w,x;\alpha,b) \geq -\log(J+1) - 2\left(\sum_{j=1}^{J}\|w_j\|_2\right)\cdot\|\alpha\|_2 - 2\left(\sum_{j=1}^{J}\|x_j\|_2\right)\cdot\|b\|_2. \tag{155}$$

Proof of Lemma F.2. The upper bound follows since $0 < \kappa(\cdot;\cdot) < 1$ everywhere. For the lower bound, note that $\log\kappa\left(y,w,x;\mathbf{0}_{d_W+d_X}\right) = -\log\left(J+1\right)$. Using the derivatives in (153) and (154), by the Mean Value Theorem, there is a midpoint $\widetilde{\zeta} = \widetilde{\zeta}(y,w,x;\alpha,b)$ between (α,b) and $\mathbf{0}_{d_W+d_X}$

such that

$$\log\kappa\left(y,w,x;\alpha,b\right) = -\log(J+1)+\left(\sum_{j=0}^{J}\left(w_y-w_j\right)\kappa\left(j,w,x;\widetilde{\zeta}\right)\right)'\alpha + \left(\sum_{j=0}^{J}\left(x_y-x_j\right)\kappa\left(j,w,x;\widetilde{\zeta}\right)\right)' b.$$

Note the expressions for the derivative in the mean value expansion follow from in Lemma F.1. Furthermore, by $\kappa(\cdot;\cdot)\in(0,1)$, $\log\kappa(\cdot;\cdot)<0$ so that $|\log\kappa(\cdot;\cdot)|=-\log\kappa(\cdot;\cdot)$. Therefore, taking absolute values and using the triangle inequality

$$\begin{aligned}
-\log\kappa\left(y,w,x;\alpha,b\right) =& \left|\log\kappa\left(y,w,x;\alpha,b\right)\right| \\
=& \left|\begin{array}{l}\log\kappa\left(y,w,x;\mathbf{0}_{d_W+d_X}\right) \\ +\left(\sum_{j=0}^{J}\left(w_y-w_j\right)\kappa\left(j,w,x;\widetilde{\zeta}\right)\right)'\alpha \\ +\left(\sum_{j=0}^{J}\left(x_y-x_j\right)\kappa\left(j,w,x;\widetilde{\zeta}\right)\right)' b\end{array}\right| \\
\leq& \left|\log\kappa\left(y,x,\mathbf{0}_{d_W+d_X}\right)\right| + \left|\left(\sum_{j=0}^{J}\left(w_y-w_j\right)\kappa\left(j,w,x;\widetilde{\zeta}\right)\right)'\alpha\right| \\
&+\left|\left(\sum_{j=0}^{J}\left(x_y-x_j\right)\kappa\left(j,w,x;\widetilde{\zeta}\right)\right)' b\right|.
\end{aligned}$$

By the Cauchy-Schwarz inequality and the fact that $\kappa(\cdot;\cdot)\in(0,1)$ everywhere,

$$\begin{aligned}
-\log\kappa\left(y,w,x;\alpha,b\right) \leq& \log(J+1)+\left\|\sum_{j=0}^{J}\left(w_y-w_j\right)\kappa\left(j,w,x;\widetilde{\zeta}\right)\right\|_2\cdot\|\alpha\|_2 \\
&+\left\|\sum_{j=0}^{J}\left(x_y-x_j\right)\kappa\left(j,w,x;\widetilde{\zeta}\right)\right\|_2\cdot\|b\|_2 \\
\leq& \log(J+1)+2\left\|\sum_{j=0}^{J}w_j\right\|_2\cdot\|\alpha\|_2+2\left\|\sum_{j=0}^{J}x_j\right\|_2\cdot\|b\|_2
\end{aligned}$$

from which the lower bound in (155) follows after multiplication by -1. □

Lemma F.3. *Let $\kappa(\cdot;\cdot)$ be defined by (150). Let (y,w,x) and $\alpha\in\mathcal{A}$ be given. For any probability* *distribution F over $\mathcal{B}$,*

$$\left|\log \int \kappa(y, w, x; \alpha, b) F(\mathrm{d}b)\right| \\ \leq \log(J+1) + 2\left(\sum_{j=0}^{J} \|w_j\|_2\right) \cdot \|\alpha\|_2 + 2\left(\sum_{j=0}^{J} \|x_j\|_2\right) \cdot \int \|b\|_2 F(\mathrm{d}b). \tag{156}$$

Proof of Lemma F.3. Since $\kappa(y, w, x; \alpha, b) \in (0, 1)$,

$$\int \kappa(y, w, x; \alpha, b) F(\mathrm{d}b) \in [0, 1] \implies \log \int \kappa(y, w, x; \alpha, b) F(\mathrm{d}b) \leq 0.$$

Using concavity of the natural logarithm and applying Jensen's inequality,

$$\int \log \kappa(y, w, x; \alpha, b) F(\mathrm{d}b) \leq \log \int \kappa(y, w, x; \alpha, b) F(\mathrm{d}b) \leq 0.$$

Thus,

$$\left|\log \int \kappa(y, w, x; \alpha, b) F(\mathrm{d}b)\right| \leq \left|\int \log \kappa(y, w, x; \alpha, b) F(\mathrm{d}b)\right| \leq \int |\log \kappa(y, w, x; \alpha, b)| F(\mathrm{d}b).$$

Inequality (156) now follows from integrating (155):

$$\left|\log \int \kappa(y, w, x; \alpha, b) F(\mathrm{d}b)\right| \leq \int |\log \kappa(y, w, x; \alpha, b)| F(\mathrm{d}b) = -\int \log \kappa(y, w, x; \alpha, b) F(\mathrm{d}b) \\ \leq \log(J+1) + 2\left(\sum_{j=0}^{J} \|w_j\|_2\right) \cdot \|\alpha\|_2 + 2\left(\sum_{j=0}^{J} \|x_j\|_2\right) \cdot \int \|b\|_2 F(\mathrm{d}b).$$

□

F.2 Miscellaneous results

Lemma F.4. *For $c, \delta > 0$*

$$\int_0^\delta \sqrt{\log(c/\varepsilon)} \mathrm{d}\varepsilon = \delta \cdot \sqrt{\log(c/\delta)} + c \int_{\sqrt{\log(c/\delta)}}^\infty \exp\left(-y^2\right) \, \mathrm{d}y. \tag{157}$$

Consequently,

$$\int_0^c \sqrt{\log(c/\varepsilon)} \mathrm{d}\varepsilon = \frac{c\sqrt{\pi}}{2}. \tag{158}$$

Proof. Conduct the following change of variables:

$$y = \sqrt{\log(c/\varepsilon)} \quad \text{so that} \quad \varepsilon = c \exp\left(-y^2\right) \quad \text{and} \quad \mathrm{d}\varepsilon = -2cy \exp\left(-y^2\right) \, \mathrm{d}y.$$

The integration bounds are changed as follows: $\varepsilon = 0 \implies y = \infty$, and $\varepsilon = \delta \implies y = y_{c,\delta}$ where for brevity of notation we define

$$y_{c,\delta} = \sqrt{\log(c/\delta)}. \tag{159}$$

By change of variables,

$$\begin{aligned}\int_0^\delta \sqrt{\log(c/\varepsilon)}\ \mathrm{d}\varepsilon &= c\int_\infty^{\sqrt{\log(c/\delta)}} y \cdot \left(-2y\exp\left(-y^2\right)\right)\ \mathrm{d}y \\ &= c\int_{y_{c,\delta}}^\infty y \cdot \left(2y\exp\left(-y^2\right)\right)\ \mathrm{d}y\end{aligned}$$

where the last equality follows from the fact that the negative in the integral will simply interchange the bounds of integration, and $y_{c,\delta} = \sqrt{\log(c/\delta)}$ as in (159). Integration by parts gives

$$\begin{aligned}\int_0^\delta \sqrt{\log(c/\varepsilon)}\ \mathrm{d}\varepsilon &= c\left[-y\cdot\exp\left(-y^2\right)\right]_{y_{c,\delta}}^\infty + c\int_{y_{c,\delta}}^\infty \exp\left(-y^2\right)\ \mathrm{d}y, \\ &= cy_{c,\delta}\cdot\exp\left(-y_{c,\delta}^2\right) + c\int_{y_{c,\delta}}^\infty \exp\left(-y^2\right)\ \mathrm{d}y\end{aligned}$$

where the last equality follows from $\lim_{y\to\infty} y\cdot\exp\left(-y^2\right) = 0$ (using L'Hôpital's rule). Plugging in the definition of $y_{c,\delta}$ from (159) gives (157). To show (158), use (157) with $\delta = c$ to write

$$\int_0^c \sqrt{\log(c/\varepsilon)}\mathrm{d}\varepsilon = c\int_0^\infty \exp\left(-y^2\right)\mathrm{d}y = c\int_0^\infty \frac{1}{\sqrt{2}}\exp\left(-\frac{x^2}{2}\right)\mathrm{d}x,$$

where the last equality follows from the change of variables $y = x/\sqrt{2}$, so that $\mathrm{d}y = (\mathrm{d}x)/\sqrt{2}$. The remaining integral is a multiple of $\sqrt{\pi}$ times the integral of the standard normal density over the non-negative half of the real line. Hence (158) follows since

$$\int_0^c \sqrt{\log(c/\varepsilon)}\mathrm{d}\varepsilon = c\sqrt{\pi}\int_0^\infty \frac{1}{\sqrt{2\pi}}\exp\left(-\frac{x^2}{2}\right)\mathrm{d}x = \frac{c\sqrt{\pi}}{2}.$$

□

Lemma F.5. *Let $a_1, a_2 \in \mathbb{R}$ and $b_1, b_2 \in (0,\infty)$. Then,*

$$\frac{a_1}{b_1} - \frac{a_2}{b_2} = \frac{1}{b_2}\left(a_1 - a_2\right) - \frac{a_2}{b_2^2}\left(b_1 - b_2\right) - \frac{1}{b_1\cdot b_2}\left[\left(a_1 - a_2\right) - \frac{a_2}{b_2}\left(b_1 - b_2\right)\right]\left(b_1 - b_2\right). \tag{160}$$

Hence,

$$\begin{aligned}\left|\frac{a_1}{b_1} - \frac{a_2}{b_2}\right| &\le \frac{1}{b_2}\left|a_1 - a_2\right| + \frac{\left|a_2\right|}{b_2^2}\left|b_1 - b_2\right| \\ &\quad + \frac{1}{\min\left\{b_1, b_2\right\}^2}\cdot\max\left\{\frac{1}{2}, \left|\frac{a_2}{b_2}\right|\right\}\left[\left|a_1 - a_2\right|^2 + \left|b_1 - b_2\right|^2\right].\end{aligned} \tag{161}$$

Proof of Lemma F.5.

$$\begin{aligned}\frac{a_1}{b_1} - \frac{a_2}{b_2} &= \frac{a_1 \cdot b_2 - a_2 \cdot b_1}{b_1 \cdot b_2} \\ &= \frac{a_1 \cdot b_2 - a_2 \cdot b_2 + a_2 \cdot b_2 - a_2 \cdot b_1}{b_1 \cdot b_2} \\ &= \frac{1}{b_1}\left(a_1 - a_2\right) - \frac{a_2}{b_1 \cdot b_2}\left(b_1 - b_2\right) \\ &= \frac{1}{b_2}\left(a_1 - a_2\right) - \frac{a_2}{b_2^2}\left(b_1 - b_2\right) + \left(\frac{1}{b_1} - \frac{1}{b_2}\right)\left[\left(a_1 - a_2\right) - \frac{a_2}{b_2}\left(b_1 - b_2\right)\right] \\ &= \frac{1}{b_2}\left(a_1 - a_2\right) - \frac{a_2}{b_2^2}\left(b_1 - b_2\right) - \frac{1}{b_1 \cdot b_2}\left[\left(a_1 - a_2\right) - \frac{a_2}{b_2}\left(b_1 - b_2\right)\right]\left(b_1 - b_2\right),\end{aligned}$$

where the last line follows from $(1/b_1) - (1/b_2) = -\left(b_1 - b_2\right) / \left(b_1 \cdot b_2\right)$. The last line is exactly (160). Then, (161) follows from the triangle inequality, bounding each term in the last summand of (160) and the inequality $u \cdot v \leq \left(u^2 + v^2\right)/2$. □