Ordered Differencing with a Generated Index in Partially Linear Models

Ryo Kimoto*

August 2026

Abstract

Ordered differencing offers a simple way to estimate partially linear models without choosing a sample-size-dependent tuning sequence, such as a bandwidth or series length. By comparing neighboring observations along a scalar index, it removes a smooth nuisance component without estimating that component directly. When the index is estimated, however, small first-stage errors can change the ordering and hence the pairs used in estimation. Just as in matching on an estimated propensity score studied by Abadie and Imbens (2016), this discontinuity prevents direct application of the usual smooth two-step argument. We develop asymptotic theory that accounts for the generated ordering while keeping the differencing order fixed. We establish asymptotic normality, characterize the contribution of first-stage estimation to the limiting distribution, and provide valid inference through subsampling. We apply the method to semiparametric sample selection and examine its performance through simulations and a wage-equation illustration.

1 Introduction

Semiparametric methods provide robustness to functional-form misspecification while retaining the possibility of root-n estimation and inference for finite-dimensional parameters. Their implementation often involves sample-size-dependent tuning parameters. For example, kernel methods require a bandwidth, series methods require a choice of dimension, and spline and related procedures involve smoothing or regularization choices. These choices govern the estimation of nonparametric components or auxiliary regressions and can affect finite-sample performance. Selection rules help make such procedures useful in practice, but remain part of the estimation problem.

In this paper, we consider the partially linear index model

$$Y_i = X_i^\mathsf{T}\theta_0 + g_0(V_i) + \varepsilon_i, \qquad V_i = W_i^\mathsf{T}\gamma_0,$$

where θ_0 is the finite-dimensional parameter of interest, g_0 is an unknown smooth function, and γ_0 is an unknown index coefficient. If γ_0 were known, the method in Yatchew (1997) would order and relabel the observations so that $V_1 < \cdots < V_n$, then regress $Y_i - Y_{i-1}$ on $X_i - X_{i-1}$ by least squares, for $i = 2, \ldots, n$. We instead estimate γ_0 in a preliminary step and apply the same procedure after ordering observations by $\widehat{V}_i = W_i^\mathsf{T}\widehat{\gamma}$. We establish conditions under which the resulting estimator for the object of interest θ_0 permits root-n estimation and inference.

*Pennsylvania State University. Email ryo.kimoto@psu.edu.

This differencing procedure avoids a sample-size-dependent tuning sequence. When the index is observed, smoothness of g_0 makes its contribution to the differenced estimating equation asymptotically negligible, allowing estimation of θ_0 without a bandwidth or a series dimension (Yatchew, 1997, 2003). Related uses of ordered observations appear in the difference-based variance estimator in Rice (1984) and the inverse-density-weighted estimators in Lewbel and Schennach (2007). These methods share the principle of ordering observations and comparing a fixed number of neighbors. Retaining a fixed order when the index is estimated is the central concern of this paper.

Our framework applies to semiparametric sample selection (Heckman, 1979; Powell, 2001; Newey, 2009), which we use to illustrate the theory. Other potential applications include single-index propensity scores (Rosenbaum and Rubin, 1983) and residual-based control functions in triangular models (Newey, Powell, and Vella, 1999), subject to our assumptions.

The difficulty in extending the argument of Yatchew (1997) arises because estimating the index can change which observations are neighbors. When the index is known, we order the observations so that $V_1 < \cdots < V_n$ and form differences between consecutive observations. Ordering by the estimated index can bring different observations next to one another, changing the differences used in the regression. Even a small first-stage estimation error can change many of these pairs. The map from the index coefficient γ to the ordering of $\{W_i^{\mathsf{T}}\gamma\}_{i=1}^n$ is nonsmooth. This prevents direct application of the classical two-step correction based on a Taylor expansion, as in Pagan (1984) and Newey and McFadden (1994).

The closest parallel is matching on an estimated propensity score in Abadie and Imbens (2016). In both settings, the first-stage estimate determines which observations are compared in the second stage. Even a small first-stage estimation error can change these comparisons. Consequently, treating the pairs as fixed does not account for the effect of first-stage estimation. Their analysis builds on Andreou and Werker (2012) and invokes Le Cam's third lemma. Their argument requires the first-stage estimator to behave regularly under local changes in the data distribution. Our analysis instead uses the information available when constructing the first-stage estimate. We require the outcome errors to have conditional mean zero given this information and the outcome regressors. Conditioning on this information fixes the estimated ordering, allowing us to analyze the differenced outcome errors. We also require an asymptotically linear first-stage estimator, but do not require regularity under local alternatives.

Lewbel and Schennach (2007) provide a second comparison. They use ordered spacings to estimate inverse-density-weighted expectations. Although their estimator permits a fixed spacing order when the ordering variable is observed, their extension to an estimated variable requires the spacing order to increase with the sample size. We establish asymptotic normality for the partially linear estimator while keeping the differencing order fixed, even when the index is estimated. The second stage therefore requires neither a bandwidth, a series length, nor a growing differencing order. When the first stage also requires no sample-size-dependent tuning sequence, this property extends to the entire two-step point estimator. In the sample-selection application, combining the tuning-parameter-free first stage of Liu and Yu (2022) with fixed differences in the outcome equation avoids sample-size-dependent tuning in both stages, while allowing a smooth, nonmonotone outcome control function.

The efficiency cost of fixed-order differencing is not specific to the partially linear model. It is a common feature of difference-based estimators, including those studied by Rice (1984), Yatchew (1997, 2003), and Lewbel and Schennach (2007). We show that the usual remedy for this efficiency loss continues to apply when the ordering index is estimated. In particular, under

conditional homoskedasticity, higher-order differences with suitable weights reduce the variance inflation due to differencing, as in Hall, Kay, and Titterington (1990) and Dette, Munk, and Wagner (1998). Both the number of observations and the weights in each difference remain fixed as the sample grows.

The choice of differencing order is analogous to the choice of instruments in conditional moment models with a finite-dimensional parameter. An expanding set of instruments can be used to approach the efficiency bound (Donald, Imbens, and Newey, 2009), but a fixed set suffices for root-n consistency and asymptotic normality under standard regularity conditions, provided that it maintains identification. In this sense, choosing a fixed differencing order is a choice among estimators with different efficiency properties, rather than a smoothing parameter that must vary with the sample size to deliver valid inference. Our procedure therefore requires a choice of order, but not a sample-size-dependent tuning sequence.

Our asymptotic distribution accounts explicitly for first-stage estimation. Direct use of this distribution for inference generally requires estimating unknown functions entering the asymptotic variance. We instead establish validity of inference by subsampling, recomputing both stages within each subsample. This avoids nonparametric variance estimation, although it still requires a subsample-size choice.

Simulations examine the contribution of first-stage estimation to sampling variability, the efficiency gains from higher-order differences, and the coverage of subsampling confidence intervals. We also compare point-estimation accuracy with kernel partialling-out using the same estimated index. Using a sample of married women ages 25–60 in California from the 2024 American Community Survey, we illustrate the procedure by estimating a wage equation in a semiparametric sample-selection model.

The remainder of the paper is organized as follows. Section 2 introduces the model and the oracle and feasible estimators. Section 3 develops the large-sample theory and subsampling inference. Section 4 presents sample selection as the main application. Section 5 reports the simulation evidence, Section 6 presents the wage-equation point estimates, and Section 7 concludes.

2 Model and estimators

We study the partially linear model

$$Y_i = X_i^\mathsf{T}\theta_0 + g_0(V_i) + \varepsilon_i, \qquad \mathbb{E}(\varepsilon_i \mid X_i, W_i) = 0, \tag{1}$$

$$V_i \equiv W_i^\mathsf{T}\gamma_0. \tag{2}$$

Here Y_i and V_i are scalar, $X_i \in \mathbb{R}^d$, and $W_i \in \mathbb{R}^q$. The scale and sign of γ_0 are normalized. The object of interest is θ_0, and the scalar-valued unknown function g_0 is a nuisance parameter. Because γ_0 is unknown, the index V_i is not observed. We first consider the oracle benchmark in which γ_0 is known and hence V_i is observed. The model then reduces to the classical partially linear model, with $\mathbb{E}(\varepsilon_i \mid X_i, V_i) = 0$, studied in Yatchew (1997).

The partially linear model has been studied extensively in the econometrics and statistics literature. The partialling-out procedure in Robinson (1988) removes g_0 from the estimating equation through the identity

$$Y_i - \mathbb{E}(Y_i \mid V_i) = \{X_i - \mathbb{E}(X_i \mid V_i)\}^\mathsf{T}\theta_0 + \varepsilon_i.$$

Implementing this identity requires estimating the auxiliary conditional means $\mathbb{E}(Y_i \mid V_i)$ and $\mathbb{E}(X_i \mid V_i)$, rather than g_0 itself. The series estimator in Chen (1988) instead approximates g_0 by a polynomial series. These approaches require the researcher to choose a bandwidth or a series dimension. In the scalar-index case considered here, the key insight in Yatchew (1997), together with the extensions in Yatchew (2003), is that the contribution of the nonparametric component g_0 can be made asymptotically negligible through differencing without a sample-size-dependent tuning parameter.

To introduce the estimator in Yatchew (1997), consider the following ordering of the observations. Let π_0 denote the permutation satisfying

$$V_{\pi_0(1)} \leq \cdots \leq V_{\pi_0(n)},$$

so that the observations are arranged in increasing order of V_i. Ties may be broken by the original observation number. Under the density conditions below, ties occur with probability zero.

For a permutation π, define

$$\Delta_\pi Y_j = Y_{\pi(j)} - Y_{\pi(j-1)}, \qquad \Delta_\pi X_j = X_{\pi(j)} - X_{\pi(j-1)}, \qquad j = 2, \ldots, n.$$

The simple difference estimator in Yatchew (1997) is

$$\widetilde{\theta} \equiv \left\{ \sum_{j=2}^{n} \Delta_{\pi_0} X_j \Delta_{\pi_0} X_j^{\mathsf{T}} \right\}^{-1} \sum_{j=2}^{n} \Delta_{\pi_0} X_j \Delta_{\pi_0} Y_j. \tag{3}$$

Equivalently, $\widetilde{\theta}$ minimizes the adjacent-pair least-squares objective

$$\widetilde{\theta} \equiv \underset{\theta \in \mathbb{R}^d}{\arg\min} \sum_{j=2}^{n} \left(\Delta_{\pi_0} Y_j - \Delta_{\pi_0} X_j^{\mathsf{T}} \theta \right)^2. \tag{4}$$

This objective admits an interpretation of the estimator as a fixed-neighbor implementation of the pairwise differencing approach studied in Honoré and Powell (2005) and Aradillas-López, Honoré, and Powell (2007). Rather than constructing all pairs and weighting pairs with similar control variables, the estimator in Yatchew (1997) orders observations by the scalar V_i and retains consecutive pairs. The number of neighbors remains fixed as the sample grows, while the distance between their index values shrinks automatically. When g_0 is smooth, $\Delta_{\pi_0} g_0(V)$ is consequently small, while differencing preserves enough variation in X_i to estimate θ_0 at the root-n rate.

We now return to the feasible setting in which γ_0 is unknown. Let $\widehat{\gamma}$ be a first-stage estimator and define $\widehat{V}_i \equiv W_i^{\mathsf{T}} \widehat{\gamma}$. Let $\widehat{\pi}$ denote the permutation obtained by ordering the sample according to $\widehat{V}_i$. The feasible estimator is

$$\widehat{\theta} \equiv \left\{ \sum_{j=2}^{n} \Delta_{\widehat{\pi}} X_j \Delta_{\widehat{\pi}} X_j^{\mathsf{T}} \right\}^{-1} \sum_{j=2}^{n} \Delta_{\widehat{\pi}} X_j \Delta_{\widehat{\pi}} Y_j. \tag{5}$$

This setup, in which γ_0 is unknown, arises naturally in many economic models. We give two prominent examples.

Example 1 (Semiparametric sample selection)**.** *In a latent-outcome model with selection determined by* $D_i = 1\{W_i^\mathsf{T}\gamma_0 + \eta_i > 0\}$*, the conditional mean among selected observations can take the form*

$$\mathbb{E}(Y_i \mid X_i, W_i, D_i = 1) = X_i^\mathsf{T}\beta_0 + \lambda_0(W_i^\mathsf{T}\gamma_0).$$

Thus the selected population satisfies (1)*, with the selection index determining the order. We develop the model in section 4.1.*

Example 2 (Semiparametric treatment effects)**.** *Under selection on observables, suppose the propensity score satisfies*

$$e(X_i) = F(X_i^\mathsf{T}\alpha_0),$$

where F *is strictly increasing and may be unknown. Ordering observations by* $X_i^\mathsf{T}\widehat{\alpha}$ *generates local treated–control comparisons without estimating* F*. In this application, treatment* D_i *plays the role of the regressor of interest in* (1)*, while* X_i *denotes the pretreatment variables entering the index. Section 4.2 states the homogeneous-effects model and the additional conditional-mean restriction needed to apply the main theorem.*

Unlike (3), the ordering in (5) is random through estimation of γ_0. The map from $\widehat{\gamma}$ to $\widehat{\pi}$ is discontinuous, so standard generated-regressor arguments such as those in Pagan (1984) and the general two-step framework in Newey and McFadden (1994, Section 6) do not directly apply. This is the main theoretical complication of the paper.

Section 3 establishes conditions under which the estimator remains root-n consistent and asymptotically normal despite changes in the ordering induced by first-stage estimation.

The strength of the proposed estimator is that its second stage requires neither a sample-size-dependent tuning parameter nor a shape restriction on g_0. Existing semiparametric estimators of sample-selection models use kernel weighting, pairwise differencing, or series approximation and therefore require a bandwidth or a series dimension (Ahn and Powell, 1993; Powell, 2001; Newey, 2009). A notable exception is the estimator in Liu and Yu (2022), which avoids such a smoothing choice by imposing monotonicity on the selection control function. The related estimator in Xu and Otsu (2020) also exploits monotonicity in a partially linear index model. Our estimator requires smoothness but does not require g_0 to be monotone. When γ_0 is estimated using a tuning-free first stage, such as the sample-selection adaptation in Liu and Yu (2022) of the estimator in Groeneboom and Hendrickx (2018), or the maximum rank correlation estimator in Han (1987), the complete point estimator has no sample-size-dependent tuning parameter.

The treatment-effect application builds on the design-based role of the propensity score. Under the selection-on-observables framework in Rosenbaum and Rubin (1983), the propensity score is a balancing score, so conditioning on it balances the distribution of observed covariates across treatment groups. The design principle discussed in Rubin (2007) is to determine the propensity-score specification without using outcome data, mirroring the separation between treatment assignment and outcome analysis in a randomized experiment. Building on this insight, the covariate balancing propensity score proposed in Imai and Ratkovic (2014) estimates a parametric treatment-assignment model while directly targeting covariate balance. Our treatment-effect application follows the same design-stage separation. The index determining the order is estimated from treatment and pretreatment covariates without consulting the outcomes.

3 Large sample properties

Root-n consistency of the first stage controls the average true-index gaps between fitted neighbors. For asymptotic normality we additionally require an outcome-free asymptotically linear first stage and the primitive density and moment conditions in Assumption 4. Under these conditions, lemma 1 establishes the fitted-order design and conditional-variance limits needed for the central limit argument.

We begin by introducing notation that will be used throughout the analysis. Write

$$m_0(v) = \mathbb{E}(X_i \mid V_i = v), \qquad \xi_i = X_i - m_0(V_i), \qquad \Sigma_X(v) = \mathbb{E}(\xi_i \xi_i^\mathsf{T} \mid V_i = v).$$

We first establish the oracle result when the ordering index is observed, then derive the limit theory under an estimated ordering, extend the argument to fixed-order weighted differences, and finally discuss inference based on subsampling.

Assumption 1 (Sampling, support, and moments)**.** *The observations are i.i.d. The support $\mathcal{V}$ of V_i is a compact interval. Its density is continuously differentiable, bounded, and bounded away from zero on $\mathcal{V}$. For some $\eta_s > 0$, $\mathbb{E}(\|X_i\|^{4+\eta_s} + \|W_i\|^{4+\eta_s} + |\varepsilon_i|^{4+\eta_s}) < \infty$. The conditional moment functions $\Sigma_X(v)$, $\mathbb{E}(\varepsilon_i^2 \mid V_i = v)$, and $\mathbb{E}(\xi_i \xi_i^\mathsf{T} \varepsilon_i^2 \mid V_i = v)$ are Lipschitz and uniformly bounded on $\mathcal{V}$.*

Assumption 2 (Smoothness and identification)**.** *The functions g_0 and m_0 are twice continuously differentiable on an open set containing the support $\mathcal{V}$ of V_i, with bounded derivatives. The matrix*

$$Q = 2\mathbb{E}\{\Sigma_X(V_i)\}$$

is positive definite.

The compact support and density conditions allow uniform-order-statistic spacing bounds to be transferred to V_i by the probability integral transform.

3.1 Oracle ordering

The oracle estimator $\widetilde{\theta}$ is defined in (3). Let

$$s_{n,j} = \Delta_{\pi_0} X_j \Delta_{\pi_0} \varepsilon_j, \qquad S_n = n^{-1/2} \sum_{j=2}^{n} s_{n,j}.$$

For any permutation π, define

$$Q_n(\pi) = \frac{1}{n} \sum_{j=2}^{n} \Delta_\pi X_j \Delta_\pi X_j^\mathsf{T}.$$

We call $Q_n^0 = Q_n(\pi_0)$ the oracle design matrix and $\widehat{Q}_n = Q_n(\widehat{\pi})$ the feasible design matrix. Define

$$\sigma^2(v) = \mathbb{E}(\varepsilon_i^2 \mid V_i = v),$$
$$M(v) = \mathbb{E}(\xi_i \xi_i^\mathsf{T} \varepsilon_i^2 \mid V_i = v), \tag{6}$$
$$\Omega = 4\mathbb{E}\{M(V_i)\} + 2\mathbb{E}\{\Sigma_X(V_i)\sigma^2(V_i)\}. \tag{7}$$

The form of Ω can be seen by fixing $V = v$ and taking three independent draws (ξ_0, ε_0), (ξ_1, ε_1), and (ξ_2, ε_2) from the conditional distribution of (ξ_i, ε_i) given $V_i = v$. Define the contributions from two consecutive pairs by

$$A_1 = (\xi_1 - \xi_0)(\varepsilon_1 - \varepsilon_0), \qquad A_2 = (\xi_2 - \xi_1)(\varepsilon_2 - \varepsilon_1).$$

A direct expansion gives

$$\mathbb{E}(A_1 A_1^{\mathsf{T}} \mid V = v) = 2M(v) + 2\Sigma_X(v)\sigma^2(v).$$

The terms A_1 and A_2 are dependent only because they share the middle draw. Conditioning on (ξ_1, ε_1) gives

$$\mathbb{E}(A_1 A_2^{\mathsf{T}} \mid V = v) = M(v).$$

Contributions from nonoverlapping pairs of observations are conditionally independent. In the variance of their sum, the variance of each term appears once and the covariance between consecutive terms appears twice. This gives

$$2M(v) + 2\Sigma_X(v)\sigma^2(v) + 2M(v) = 4M(v) + 2\Sigma_X(v)\sigma^2(v).$$

Averaging this expression over V_i yields (7).

The following proposition extends the oracle argument in Yatchew (1997) to permit conditional heteroskedasticity. In particular, both $\sigma^2(v) = \mathbb{E}(\varepsilon_i^2 \mid V_i = v)$ and the conditional joint moment $M(v)$ may vary with v.

Proposition 1 (Oracle asymptotics)**.** *Suppose Assumptions 1 and 2 hold. Then*

$$Q_n^0 \xrightarrow{p} Q,$$

and

$$\sqrt{n}(\widetilde{\theta} - \theta_0) = Q^{-1} S_n + o_{\mathbb{P}}(1) \xrightarrow{d} \mathcal{N}(0, Q^{-1}\Omega Q^{-1}).$$

Proof. The proof is given in Appendix A. □

Under homoskedasticity, proposition 1 specializes to the result in Yatchew (1997).

Corollary 1 (Yatchew (1997))**.** *Suppose, in addition, that* $\mathbb{E}(\varepsilon_i^2 \mid X_i, V_i) = \sigma^2$*. Then*

$$M(v) = \sigma^2 \Sigma_X(v), \qquad \Omega = 6\sigma^2 \mathbb{E}\{\Sigma_X(V_i)\},$$

and

$$\sqrt{n}(\widetilde{\theta} - \theta_0) \xrightarrow{d} \mathcal{N}\left(0, \frac{3}{2}\sigma^2 [\mathbb{E}\{\Sigma_X(V_i)\}]^{-1}\right).$$

The semiparametric efficiency bound in this case is $\sigma^2[\mathbb{E}\{\Sigma_X(V_i)\}]^{-1}$ (Chamberlain, 1992). The first-difference estimator therefore has an asymptotic variance that is $3/2$ times the efficiency bound and is less efficient than the kernel-based estimator in Robinson (1988), which attains the bound under suitable regularity conditions. The advantage of the estimator in Yatchew (1997) is that it does not require the researcher to choose a sample-size-dependent smoothing parameter.

The 3/2 efficiency factor is associated with the use of first differences. The variance-estimation literature develops weighted higher-order differences that reduce this loss (Hall, Kay, and Titterington, 1990; Dette, Munk, and Wagner, 1998). As the order increases, the efficiency bound can be approached, although the differencing rule becomes more elaborate. For any chosen order, however, the order can remain fixed as the sample size grows; it need not be selected through a sample-size-dependent tuning sequence. See Dai, Tong, and Zhu (2017) for an overview. We focus on first differences in the main development and formalize the fixed-order extension in corollary 2.

3.2 Ordering by a generated index

The estimated index can change many adjacent pairs. We handle this directly by conditioning on the variables used to construct the first stage and the outcome regressors. The resulting ordering is fixed under the conditional law. The restriction on the first-stage information set is essential: asymptotic linearity alone does not exclude outcome-dependent changes in the ordering at scales smaller than $n^{-1/2}$.

Let A_i collect any additional observed first-stage variables and put $B_i = (X_i, W_i, A_i)$ and $\mathcal{F}_n = \sigma(B_1, \ldots, B_n)$. The augmented observations (Y_i, B_i) are i.i.d. Fix scale and sign by fixing one nonzero index coefficient at a specified value. The parameter set Γ is compact in this affine space, and γ_0 is in its relative interior. We retain the full vectors W_i, γ_0 in all formulas; the corresponding fixed coordinate of the influence function is zero. Its covariance matrix need not be nonsingular.

Assumption 3 (Outcome-free first-stage construction)**.** *The estimator $\widehat{\gamma} \in \Gamma$ is $\mathcal{F}_n$-measurable and $\mathbb{E}(\varepsilon_i \mid B_i) = 0$. It admits the representation*

$$\sqrt{n}(\widehat{\gamma} - \gamma_0) = n^{-1/2} \sum_{i=1}^{n} \psi(B_i) + o_{\mathbb{P}}(1), \tag{8}$$

where $\mathbb{E}\psi(B_i) = 0$ and $\mathbb{E}\|\psi(B_i)\|^2 < \infty$. Write $\Omega_\psi = \mathbb{E}\{\psi(B_i)\psi(B_i)^{\mathsf{T}}\}$.

The conditioning restriction is stronger than simply requiring the influence function to be outcome-free: the finite-sample estimator must use only $\mathcal{F}_n$. In sample selection, $A_i = D_i$ and an index estimator computed from (D_i, W_i) satisfies this measurability requirement. The selected outcome error must be centered conditional on $(X_i, W_i, D_i = 1)$. Asymptotic linearity of the particular first-stage implementation remains a separate requirement. In contrast, the argument of Abadie and Imbens (2016, Section 3.1), building on Andreou and Werker (2012), uses local likelihood ratios and Le Cam's third lemma (van der Vaart, 1998, p. 90), followed by a discretization of the first-stage coefficient estimator. Our conditional argument works directly with the original first-stage estimate and requires neither a likelihood-based change of measure nor this discretization.

For fixed $\delta < \infty$, define

$$\Gamma_n(\delta) \equiv \left\{\gamma \in \Gamma : \sqrt{n}\|\gamma - \gamma_0\| \leq \delta\right\}. \tag{9}$$

The generated ordering need not recover the oracle ordering. A root-n perturbation of the index is large relative to the ordinary n^{-1} spacing between consecutive oracle order statistics. The appropriate stability property is therefore an average-spacing result.

Proposition 2 (Average spacing under an estimated index)**.** *Suppose Assumption 1 holds and* $\sqrt{n}(\widehat{\gamma} - \gamma_0) = O_{\mathbb{P}}(1)$*. Then*

$$\frac{1}{n}\sum_{j=2}^{n}\left(V_{\widehat{\pi}(j)} - V_{\widehat{\pi}(j-1)}\right)^2 = O_{\mathbb{P}}(n^{-1}). \tag{10}$$

Moreover,

$$\sum_{j=2}^{n}(\Delta_{\widehat{\pi}}\widehat{V}_j)^2 = o_{\mathbb{P}}(1), \tag{11}$$

$$\max_{2\le j\le n}|\Delta_{\widehat{\pi}}V_j| = o_{\mathbb{P}}(1). \tag{12}$$

For every fixed $\delta < \infty$*, the first conclusion also holds uniformly as*

$$\sup_{\gamma\in\Gamma_n(\delta)}\frac{1}{n}\sum_{j=2}^{n}\left(V_{\pi_\gamma(j)} - V_{\pi_\gamma(j-1)}\right)^2 = O_{\mathbb{P}}(n^{-1}). \tag{13}$$

Proof. The proof is given in Appendix B. □

The proposition does not assert pointwise recovery of pairs that are consecutive under the true ordering generated by V_i. It shows instead that observations adjacent under the estimated ordering are close in the true index in an average quadratic sense. To see how this result propagates into the subsequent bounds, consider a function $v \mapsto a(v)$ that is Lipschitz on $\mathcal{V}$. For some finite Lipschitz constant L_a, we have

$$\frac{1}{n}\sum_{j=2}^{n}\left\|a(V_{\widehat{\pi}(j)}) - a(V_{\widehat{\pi}(j-1)})\right\|^2 \le \frac{L_a^2}{n}\sum_{j=2}^{n}\left|V_{\widehat{\pi}(j)} - V_{\widehat{\pi}(j-1)}\right|^2 = O_{\mathbb{P}}(n^{-1}). \tag{14}$$

This conclusion applies in particular to m_0, g_0, and the conditional moment functions in (6).

It is useful to compare this result with the oracle argument in Yatchew (1997). For example, suppose the ordered values of the observed V_i are equally spaced on $[0, 1]$. Then

$$|V_{\pi_0(j)} - V_{\pi_0(j-1)}| = \frac{1}{n-1}$$

for every j. Lipschitz continuity consequently gives the pointwise bound

$$\|a(V_{\pi_0(j)}) - a(V_{\pi_0(j-1)})\|^2 \le \frac{L_a^2}{(n-1)^2}$$

for each component inside the sum in (14). This pointwise argument is no longer available when the ordering index is generated. Estimation of γ_0 can cause observations to switch ranks, and the true-index gap for a given pair of consecutive observations under $\widehat{\pi}$ need not have order n^{-1}.

Remark 1 (Pointwise and aggregate behavior)**.** *A related distinction between pointwise and aggregate behavior arises in semiparametric estimation. A nonparametric component may converge more slowly than* $n^{-1/2}$ *at a given point, while an estimator that averages its contributions*

admits a root-n expansion, often through a U-statistic representation. Here, the analogous lesson is that pointwise control of every fitted-neighbor gap is not required. What matters is how these gaps enter the aggregate remainder. Proposition 2 provides the average quadratic bound used to control that remainder in the subsequent analysis. This bound is one ingredient in the root-n result; the central limit argument and design limits require the additional conditions stated below.

To see how proposition 2 becomes relevant, let us represent the centered and scaled feasible estimator as

$$\sqrt{n}(\widehat{\theta} - \theta_0) = \widehat{Q}_n^{-1}\{\widehat{S}_n + R_{g,n}\}, \tag{15}$$

where

$$\widehat{Q}_n \equiv \frac{1}{n}\sum_{j=2}^{n} \Delta_{\widehat{\pi}} X_j \Delta_{\widehat{\pi}} X_j^{\mathsf{T}},$$

$$\widehat{S}_n \equiv \frac{1}{\sqrt{n}}\sum_{j=2}^{n} \Delta_{\widehat{\pi}} X_j \Delta_{\widehat{\pi}} \varepsilon_j,$$

$$R_{g,n} \equiv \frac{1}{\sqrt{n}}\sum_{j=2}^{n} \Delta_{\widehat{\pi}} X_j \Delta_{\widehat{\pi}} g_0(V)_j.$$

Note that the representation in (15) is exact.

Using the definition of ξ_i, write

$$X_i = m_0(V_i) + \xi_i.$$

Equation (14) and Cauchy–Schwarz give

$$\widehat{Q}_n = \frac{1}{n}\sum_{j=2}^{n} \Delta_{\widehat{\pi}} \xi_j \Delta_{\widehat{\pi}} \xi_j^{\mathsf{T}} + o_{\mathbb{P}}(1). \tag{16}$$

The score reduction

$$\widehat{S}_n = \frac{1}{\sqrt{n}}\sum_{j=2}^{n} \Delta_{\widehat{\pi}} \xi_j \Delta_{\widehat{\pi}} \varepsilon_j + o_{\mathbb{P}}(1) \tag{17}$$

follows by conditional Chebyshev's inequality, using the outcome-free ordering and the spacing bound (see Appendix D for the argument).

Equation (14) and Cauchy–Schwarz also give

$$R_{g,n} = \frac{1}{\sqrt{n}}\sum_{j=2}^{n} \Delta_{\widehat{\pi}} \xi_j \Delta_{\widehat{\pi}} g_0(V)_j + o_{\mathbb{P}}(1). \tag{18}$$

Thus proposition 2 removes the smooth m_0 component from all three objects in the presence of the generated index. The resulting expressions resemble their oracle counterparts, but there is an important difference. In the oracle case, differences are taken with respect to the permutation π_0 generated by the observed true index. In (16)–(18), differences are taken with respect to the estimated permutation $\widehat{\pi}$. Because $\widehat{\pi}$ is determined by the generated index, the oracle law of large numbers and central limit theorem cannot be applied directly.

The spacing result controls distances between fitted neighbors. We now give primitive conditions that also imply the required laws of large numbers under the fitted ordering. Let $\mathbf{X} = (X_1, \ldots, X_n)^\mathsf{T}$ be the $n \times d$ matrix whose ith row is X_i^T. Define the $(n-1) \times n$ first-difference matrix $H_{\widehat{\pi}}$, whose row $j-1$ has $+1$ in column $\widehat{\pi}(j)$, -1 in column $\widehat{\pi}(j-1)$, and zeros elsewhere, for $j = 2, \ldots, n$. Thus, for example, the entry in row 1 and column k of $H_{\widehat{\pi}}\mathbf{X}$ is

$$(H_{\widehat{\pi}}\mathbf{X})_{1,k} = X_{\widehat{\pi}(2),k} - X_{\widehat{\pi}(1),k}, \qquad k = 1, \ldots, d.$$

Put

$$c_{n,i} = (H_{\widehat{\pi}}^\mathsf{T} H_{\widehat{\pi}}\mathbf{X})_{i,\cdot}^\mathsf{T}, \qquad v_i = \mathbb{E}(\varepsilon_i^2 \mid B_i).$$

The exact conditional score variance is

$$\mathcal{V}_n = \frac{1}{n}\sum_{i=1}^{n} v_i c_{n,i} c_{n,i}^\mathsf{T}, \qquad \widehat{S}_n = \frac{1}{\sqrt{n}}\sum_i c_{n,i}\varepsilon_i.$$

Also define

$$G_n = \frac{1}{n}\sum_{j=2}^{n} \Delta_{\widehat{\pi}} X_j \, g_0'(V_{\widehat{\pi}(j)}) \Delta_{\widehat{\pi}} W_j^\mathsf{T}.$$

The population limit of G_n is

$$G = 2\mathbb{E}\big[g_0'(V_i)\,\mathrm{Cov}(X_i, W_i \mid V_i)\big].$$

To state primitive conditions for the fitted-order moment limits, let the coefficient fixed by normalization be $\gamma_{0,1} \neq 0$, and write $Z = (W_2, \ldots, W_q)^\mathsf{T}$. Then $W_1 = (V - \gamma_{0,-1}^\mathsf{T} Z)/\gamma_{0,1}$ and every nearby normalized index can be written $U_h = V + h^\mathsf{T} Z$. Write the support of V as $\mathcal{V} = [v_L, v_H]$.

Assumption 4 (Density and conditional moments)**.** *1. The support $\mathcal{Z}$ of Z is compact. The pair (V, Z) has a density $p(v, z)$ relative to $dv\,\mu(dz)$ on $[v_L, v_H] \times \mathcal{Z}$, where μ is a finite measure. The density admits a version satisfying*

$$0 \le p(v, z) \le M_p, \qquad |p(v, z) - p(v', z)| \le L_p |v - v'|$$

for all $v, v' \in [v_L, v_H]$ and $z \in \mathcal{Z}$, where M_p and L_p are finite constants that do not depend on v, v', or z. The marginal density of V is bounded away from zero on $[v_L, v_H]$.

2. $\|X\|$ and $v(B) = \mathbb{E}(\varepsilon^2 \mid B)$ are bounded almost surely. Consider the conditional moments

$$\mathbb{E}(X \mid V, Z), \quad \mathbb{E}(XX^\mathsf{T} \mid V, Z), \quad \mathbb{E}(v(B) \mid V, Z),$$
$$\mathbb{E}(v(B)X \mid V, Z), \quad \mathbb{E}(v(B)XX^\mathsf{T} \mid V, Z)$$

Each displayed conditional moment admits a version $m(v, z)$ satisfying

$$\|m(v, z)\| \le M_m, \qquad \|m(v, z) - m(v', z)\| \le L_m |v - v'|$$

for all $v, v' \in [v_L, v_H]$ and $z \in \mathcal{Z}$. The finite constants M_m and L_m may differ across the five functions but do not depend on v, v', or z.

The reference measure μ permits discrete as well as continuous components of Z. These conditions allow $\mathrm{Cov}(X, W \mid V) \neq 0$ and hence a nonzero first-stage correction G. Bounded regressors and bounded conditional error variance strengthen the moment restrictions in Assumption 1. The assumption concerns the population and does not require independence between the first-stage estimate and the observations being ordered.

Lemma 1 (Fitted-order laws of large numbers)**.** *Suppose Assumptions 1, 2 and 4 hold. Let $Q_n(h)$, $G_n(h)$, and $\mathcal{V}_n(h)$ be the matrices defined above, with observations ordered by $U_{h,i} = V_i + h^{\mathsf{T}} Z_i$ in place of $\widehat{V}_i$. For every deterministic $r_n \downarrow 0$, we have*

$$\sup_{\|h\| \le r_n} \left\{ \|Q_n(h) - Q\| + \|G_n(h) - G\| + \|\mathcal{V}_n(h) - \Omega\| \right\} \xrightarrow{p} 0.$$

Consequently, for the ordering $\widehat{\pi}$ induced by any consistent normalized first-stage estimator $\widehat{\gamma}$, we have

$$\widehat{Q}_n \xrightarrow{p} Q, \qquad G_n \xrightarrow{p} G, \qquad \mathcal{V}_n \xrightarrow{p} \Omega, \tag{19}$$

where Q and Ω are defined in the oracle result.

Proof. See Appendix C. □

The spacing bound in proposition 2 controls the smooth nuisance remainders, while lemma 1 provides the fitted-order moment limits. Combined with the first-stage expansion and conditional centering in Assumption 3, these results yield the following asymptotic distribution for the feasible estimator.

Theorem 1 (Conditional asymptotics under a generated ordering)**.** *Suppose Assumptions 1 to 4 hold. Then*

$$\sqrt{n}(\widehat{\theta} - \theta_0) = Q^{-1} \left\{ \widehat{S}_n - G \frac{1}{\sqrt{n}} \sum_i \psi(B_i) \right\} + o_{\mathbb{P}}(1) \xrightarrow{d} \mathcal{N}(0, V_\theta), \tag{20}$$

where

$$V_\theta = Q^{-1}(\Omega + G\Omega_\psi G^{\mathsf{T}})Q^{-1}. \tag{21}$$

Proof. The proof is given in Appendix D. □

Conditional on $\mathcal{F}_n$, the fitted-order score is a sum of independent centered outcome errors with fixed coefficients. A conditional Lyapunov argument gives a Gaussian limit with deterministic covariance Ω. Its conditional characteristic function converges to a constant, which establishes joint convergence and asymptotic independence from the $\mathcal{F}_n$-measurable first-stage sum. Consequently there are no score–first-stage covariance terms. This argument applies to the unrounded implemented index, subject to Assumption 4.

3.3 Weighted higher-order differences

The first-difference procedures studied earlier can be extended to higher-order differences with fixed weights, as in Rice (1984), Yatchew (2003), and Lewbel and Schennach (2007). Appropriate weights reduce the additional variance due to differencing, including under heteroskedasticity. In particular, the construction in Yatchew (2003) uses higher-order differencing for the partially

linear model. We show that this higher-order construction continues to apply when observations are arranged by a permutation generated from an estimated index. The number of observations entering each difference remains fixed as the sample size grows.

Fix an integer $K \geq 1$ and a nonzero vector $d = (d_0, \ldots, d_K)^\mathsf{T}$ satisfying

$$\sum_{\ell=0}^{K} d_\ell = 0.$$

For any variable A_i and permutation π, define the weighted difference

$$\Delta_{\pi,d} A_j \equiv \sum_{\ell=0}^{K} d_\ell A_{\pi(j-\ell)}, \qquad j = K+1, \ldots, n,$$

and let

$$\widehat{\theta}_d \equiv \left\{ \sum_{j=K+1}^{n} \Delta_{\widehat{\pi},d} X_j \Delta_{\widehat{\pi},d} X_j^\mathsf{T} \right\}^{-1} \sum_{j=K+1}^{n} \Delta_{\widehat{\pi},d} X_j \Delta_{\widehat{\pi},d} Y_j. \tag{22}$$

The original estimator corresponds to $K = 1$ and $d = (1, -1)^\mathsf{T}$, up to an irrelevant reversal of sign. For example, a normalized first-order sequence is

$$d = 2^{-1/2}(1, -1)^\mathsf{T}.$$

For $K = 2$, the optimal sequence in Hall, Kay, and Titterington (1990), as tabulated in Yatchew (2003, Table 4.1), is

$$d = \left(\frac{1+\sqrt{5}}{4}, -\frac{1}{2}, \frac{1-\sqrt{5}}{4} \right)^\mathsf{T}.$$

Both sequences satisfy $\sum_{\ell=0}^{K} d_\ell = 0$ and $\sum_{\ell=0}^{K} d_\ell^2 = 1$. Multiplying every weight by a common nonzero constant does not change $\widehat{\theta}_d$.

To state the limit, define

$$c_h(d) \equiv \sum_{\ell=0}^{K-h} d_\ell d_{\ell+h}, \qquad h = 0, \ldots, K,$$

and put

$$Q_d \equiv c_0(d)\mathbb{E}\{\Sigma_X(V_i)\}, \qquad G_d \equiv c_0(d)\mathbb{E}\left[g_0'(V_i)\,\mathrm{Cov}(X_i, W_i \mid V_i)\right].$$

For each v, let $\{(\xi_r(v), \varepsilon_r(v)) : r \in \mathbb{Z}\}$ be independent draws from the conditional distribution of (ξ_i, ε_i) given $V_i = v$, and define

$$b_{d,r}(v) \equiv \left\{ \sum_{\ell=0}^{K} d_\ell \xi_{r-\ell}(v) \right\} \left\{ \sum_{\ell=0}^{K} d_\ell \varepsilon_{r-\ell}(v) \right\}.$$

Define the conditional long-run covariance

$$\mathcal{B}_d(v) \equiv \sum_{h=-K}^{K} \mathbb{E}\{b_{d,0}(v) b_{d,h}(v)^\mathsf{T}\}.$$

The weighted oracle-score covariance is

$$\Omega_d \equiv \mathbb{E}\{\mathcal{B}_d(V_i)\},$$

and its covariance with the first-stage influence function is zero under Assumption 3.

Let H_d have rows implementing the fitted-order weighted differences and put $c_{n,i,d} = (H_d^\mathsf{T} H_d \mathbf{X})_{i,\cdot}^\mathsf{T}$. For weighted differences, the counterparts of the moment limits in lemma 1 are

$$n^{-1}(H_d\mathbf{X})^\mathsf{T} H_d\mathbf{X} \to_p Q_d, \qquad G_{n,d} \to_p G_d, \qquad n^{-1}\sum_i v_i c_{n,i,d} c_{n,i,d}^\mathsf{T} \to_p \Omega_d,$$

where

$$G_{n,d} = n^{-1}\sum_{j=K+1}^{n} \Delta_{\widehat{\pi},d} X_j \, g_0'(V_{\widehat{\pi}(j)}) \Delta_{\widehat{\pi},d} W_j^\mathsf{T}.$$

Under Assumptions 1, 2 and 4, these limits follow by extending the argument for first differences in lemma 1. The number of observations entering each difference and the weights remain fixed as $n \to \infty$. Consequently, each observation enters only a bounded number of summands, with a bound independent of n. Appendix E gives the extension.

Corollary 2 (Fixed-order weighted differences)**.** *Suppose Assumptions 1 to 4 hold. For any fixed K and fixed weights d satisfying the conditions above,*

$$\sqrt{n}(\widehat{\theta}_d - \theta_0) \xrightarrow{d} \mathcal{N}(0, V_{\theta,d}),$$

where

$$V_{\theta,d} \equiv Q_d^{-1}\left(\Omega_d + G_d \Omega_\psi G_d^\mathsf{T}\right) Q_d^{-1}. \tag{23}$$

The oracle-score covariance admits the expression

$$\Omega_d = c_0(d)^2 \mathbb{E}\{M(V_i)\} + 2\sum_{h=1}^{K} c_h(d)^2 \mathbb{E}\{\sigma^2(V_i)\Sigma_X(V_i)\}. \tag{24}$$

Under conditional homoskedasticity,

$$V_{\theta,d} = \left[1 + 2\sum_{h=1}^{K}\left\{\frac{c_h(d)}{c_0(d)}\right\}^2\right] \sigma^2 [\mathbb{E}\{\Sigma_X(V_i)\}]^{-1} \tag{25}$$

when γ_0 is known. For the optimal order-K difference sequences in Hall, Kay, and Titterington (1990), the factor in brackets equals $1 + 1/(2K)$.

Proof. The proof is given in Appendix E. □

The variance cost of differencing also arises under heteroskedasticity. To make the comparison explicit, suppose the index is known and put

$$A = \mathbb{E}\{\Sigma_X(V_i)\}, \qquad \kappa(d) = 2\sum_{h=1}^{K}\left\{\frac{c_h(d)}{c_0(d)}\right\}^2.$$

When the index is known, the ordinary partialling-out estimator in Robinson (1988) has asymptotic variance $V_{\mathrm{PO}} = A^{-1}\mathbb{E}\{M(V_i)\}A^{-1}$. By (24), the differencing estimator has variance

$$V_{\theta,d} = V_{\mathrm{PO}} + \kappa(d)A^{-1}\mathbb{E}\{\sigma^2(V_i)\Sigma_X(V_i)\}A^{-1}.$$

The additional matrix sandwiched between the two factors of A^{-1} is positive semidefinite. First differences give $\kappa(d) = 1/2$, whereas the optimal order-K sequences give $\kappa(d) = 1/(2K)$. These sequences therefore reduce the additional variance under heteroskedasticity as well. For an estimated index, the first-stage contribution in (23) is common across the fixed rules, since $Q_d^{-1}G_d$ does not depend on d.

If $\mathbb{E}(\varepsilon_i^2 \mid X_i, V_i) = \sigma^2(V_i)$, then $M(v) = \sigma^2(v)\Sigma_X(v)$ and $V_{\theta,d} = \{1 + \kappa(d)\}V_{\mathrm{PO}}$ when the index is known. Thus the 3/2 factor for first differences also holds when the conditional error variance varies only with the index. Under general heteroskedasticity, ordinary partialling-out need not attain the semiparametric efficiency bound characterized in Chamberlain (1992). Attaining this bound generally requires weighting that accounts for the conditional error variance. Higher fixed orders can bring the oracle variance arbitrarily close to the partialling-out variance. Under conditional homoskedasticity, this also approaches the efficiency bound, as in (25). The choice of a fixed order affects efficiency but not root-n consistency. This differs from choosing a bandwidth or a series dimension, whose asymptotic validity generally requires the tuning parameter to change with the sample size. The choice is closer to selecting instruments in a conditional moment model. A fixed finite collection of valid instruments can deliver root-n consistency, while expanding or optimizing that collection involves an efficiency and finite-sample bias–variance tradeoff (Newey, 1990; Donald, Imbens, and Newey, 2009). Likewise, any fixed differencing order satisfying the conditions above preserves the root-n rate, while different orders and weights produce different asymptotic variances. A data-driven choice of the differencing order is left for future work.

3.4 Subsampling inference

In principle, inference could be based directly on the asymptotic variance in theorem 1. Its components, including G and Ω, typically require nonparametric smoothing to estimate. This is unattractive when one of the main appeals of the ordered-difference estimator is that point estimation does not require a sample-size-dependent tuning parameter. We instead propose subsampling. Its validity follows from the preceding asymptotic result.

Subsampling is preferable here to relying on the ordinary nonparametric bootstrap. The map from the empirical distribution to the generated ordering is nonsmooth, and bootstrap resampling can change ranks and consecutive pairs discontinuously. A closely related issue arises for nearest-neighbor matching. The result in Abadie and Imbens (2008) establishes that the ordinary bootstrap is generally invalid for matching estimators with a fixed number of matches, even though those estimators are asymptotically normal. This result does not by itself establish bootstrap failure for the ordered-difference estimator, but it shows why asymptotic normality alone is insufficient to justify the bootstrap for a statistic constructed from estimated neighbors. Subsampling does not require bootstrap consistency and is valid here under the conditions stated below.

Corollary 3 (Validity of subsampling)**.** *Suppose the conditions of theorem 1 hold, and let* $b = b_n$ *satisfy*

$$b \longrightarrow \infty, \qquad \frac{b}{n} \longrightarrow 0.$$

For each subsample $I \subset \{1, \ldots, n\}$ of size b, let $\widehat{\theta}_{b,I}$ denote the estimator obtained by applying the entire estimation procedure, including the first stage, to the observations in I.

For any fixed vector a such that $a^{\mathsf{T}} V_\theta a > 0$, define

$$\widehat{L}_{n,b}(x) \equiv \binom{n}{b}^{-1} \sum_{I:|I|=b} 1\Big\{\sqrt{b}\, a^{\mathsf{T}}(\widehat{\theta}_{b,I} - \widehat{\theta}_n) \le x\Big\}.$$

Then

$$\sup_{x \in \mathbb{R}} \Big|\widehat{L}_{n,b}(x) - \mathbb{P}\Big\{\sqrt{n}\, a^{\mathsf{T}}(\widehat{\theta}_n - \theta_0) \le x\Big\}\Big| \xrightarrow{p} 0.$$

Consequently, subsampling quantiles provide asymptotically valid critical values and confidence intervals for $a^{\mathsf{T}}\theta_0$.

Proof. By theorem 1,

$$\sqrt{n}\, a^{\mathsf{T}}(\widehat{\theta}_n - \theta_0) \xrightarrow{d} \mathcal{N}(0, a^{\mathsf{T}} V_\theta a).$$

The limiting distribution is continuous. The full-sample and subsample normalizing factors are $\sqrt{n}$ and $\sqrt{b}$, respectively, and their ratio satisfies

$$\frac{\sqrt{b}}{\sqrt{n}} = \sqrt{\frac{b}{n}} \longrightarrow 0.$$

Since $b \to \infty$ and $b/n \to 0$, the result follows from Theorem 2.2.1 in Politis, Romano, and Wolf (1999). □

For implementation, practitioners may proceed in two steps.

1. Estimate γ_0 using the full sample and construct $\widehat{V}_i = W_i^{\mathsf{T}}\widehat{\gamma}$. Order the observations by $\widehat{V}_i$ and compute $\widehat{\theta}_n$ from the resulting consecutive differences.

2. Draw a subsample of size b without replacement. Re-estimate the first-stage index using only that subsample, reorder the subsample using the re-estimated index, and compute $\widehat{\theta}_{b,I}$. Repeat this procedure over B subsamples.

We construct an equal-tailed 95 percent confidence interval using the empirical 2.5th and 97.5th percentiles of the centered and scaled subsampling statistics

$$T_{b,r} = \sqrt{b}\, a^{\mathsf{T}}(\widehat{\theta}_{b,I_r} - \widehat{\theta}_n), \qquad r = 1, \ldots, B.$$

Denote these empirical quantiles by $\widehat{q}_{0.025}$ and $\widehat{q}_{0.975}$. The confidence interval for $a^{\mathsf{T}}\theta_0$ is

$$\left[a^{\mathsf{T}}\widehat{\theta}_n - \frac{\widehat{q}_{0.975}}{\sqrt{n}}, \quad a^{\mathsf{T}}\widehat{\theta}_n - \frac{\widehat{q}_{0.025}}{\sqrt{n}}\right]. \tag{26}$$

The quantiles enter in reverse order because the interval is obtained by solving $\widehat{q}_{0.025} \le \sqrt{n}\, a^{\mathsf{T}}(\widehat{\theta}_n - \theta_0) \le \widehat{q}_{0.975}$ for $a^{\mathsf{T}}\theta_0$.

It is important to repeat the complete first-stage estimation and ordering procedure within each subsample. Holding the full-sample ordering fixed would omit the sampling variation due to estimation of the index.

The next section explains how the theorem applies to sample-selection and homogeneous treatment-effect models and states sufficient conditions for its assumptions.

4 Examples

We give two examples covered by theorem 1, two-step estimation of a sample-selection model and two-step estimation of a homogeneous treatment-effect model.

4.1 Semiparametric sample selection

Consider a semiparametric version of the sample selection model considered in Heckman (1979),

$$Y_i^* = X_i^\mathsf{T}\beta_0 + e_i, \tag{27}$$

$$D_i = 1\{W_i^\mathsf{T}\gamma_0 + \eta_i > 0\}. \tag{28}$$

The researcher observes $(D_iY_i^*, D_i, X_i, W_i)$. We impose the semiparametric control-function restriction

$$\mathbb{E}(e_i \mid X_i, W_i, D_i = 1) = \lambda_0(W_i^\mathsf{T}\gamma_0). \tag{29}$$

Equation (29) imposes the conditional-mean index restriction used in semiparametric sample selection models, as in Newey (2009). One sufficient primitive condition is $(e_i, \eta_i) \perp (X_i, W_i)$, which gives $\lambda_0(v) = \mathbb{E}(e_i \mid \eta_i > -v)$. Equation (29) does not require this independence condition or a parametric specification of the disturbance distribution. The additional moment and smoothness conditions needed for our asymptotic results are stated below.

Define

$$V_i \equiv W_i^\mathsf{T}\gamma_0, \qquad \varepsilon_i \equiv e_i - \lambda_0(V_i).$$

Multiplying the outcome equation by D_i gives

$$D_iY_i^* = (D_iX_i)^\mathsf{T}\beta_0 + D_i\lambda_0(V_i) + D_i\varepsilon_i.$$

Among observations with $D_i = 1$, this becomes

$$D_iY_i^* = (D_iX_i)^\mathsf{T}\beta_0 + \lambda_0(V_i) + \varepsilon_i, \qquad \mathbb{E}(\varepsilon_i \mid X_i, W_i, D_i = 1) = 0.$$

The selected-sample model therefore coincides with (1), with Y_i replaced by $D_iY_i^*$, X_i replaced by D_iX_i, $\theta_0 = \beta_0$, $g_0 = \lambda_0$, and sample size $n_1 = \sum_i D_i$.

Let $\widehat{\pi}_1$ order the selected observations by $W_i^\mathsf{T}\widehat{\gamma}$. We write the estimator on the full-sample scale as

$$\widehat{\beta} \equiv \left\{ \frac{1}{n}\sum_{j=2}^{n_1} \Delta_{\widehat{\pi}_1}(DX)_j \Delta_{\widehat{\pi}_1}(DX)_j^\mathsf{T} \right\}^{-1} \frac{1}{n}\sum_{j=2}^{n_1} \Delta_{\widehat{\pi}_1}(DX)_j \Delta_{\widehat{\pi}_1}(DY^*)_j. \tag{30}$$

The factors $1/n$ cancel from the point estimator. We use n, rather than n_1, so that the selected-sample score and the first-stage expansion are both expressed on the full-sample $\sqrt{n}$ scale.

Suppose $p_1 = \Pr(D_i = 1) \in (0, 1)$. Define the selected-population law by

$$P_{\text{sel}}(A) \equiv \Pr\big((Y_i^*, X_i, W_i) \in A \mid D_i = 1\big) = \frac{\Pr\big((Y_i^*, X_i, W_i) \in A,\ D_i = 1\big)}{p_1}$$

for every measurable set A. Expectations under this law satisfy

$$\mathbb{E}_{\text{sel}}\big[a(Y_i^*, X_i, W_i)\big] = \frac{\mathbb{E}\big[D_i\, a(Y_i^*, X_i, W_i)\big]}{p_1}$$

for every integrable function a. To apply theorem 1, we maintain Assumptions 1 and 2 under P_{sel}. In particular, these assumptions impose smoothness of λ_0 and positive definiteness of the selected-population design matrix. We state sufficient conditions for the remaining requirements below and provide the verification in Appendix F.

For the first stage, suppose $\widehat{\gamma}$ uses only (D_i, W_i) and satisfies (8) with a centered, square-integrable influence function $\psi(D_i, W_i)$. Under their respective model and regularity conditions and the index normalization used here, examples include correctly specified probit and logit maximum likelihood estimators and the semiparametric estimators in Powell, Stock, and Stoker (1989), Klein and Spady (1993), and Ichimura (1993). Tuning-free alternatives include the maximum rank correlation estimator in Han (1987), whose asymptotic linearity follows from the proof in Sherman (1993), and the first-stage isotonic score estimator used in Liu and Yu (2022), based on Groeneboom and Hendrickx (2018). These alternatives also require their respective model and regularity conditions.

Sufficient conditions for Assumption 4 are as follows. Suppose $(e_i, \eta_i) \perp (X_i, W_i)$ and the full-population density conditions in Assumption 4(1) hold. Write

$$s(v) = \Pr(\eta_i > -v), \qquad r_S(v) = \operatorname{Var}(e_i \mid \eta_i > -v).$$

Suppose s is Lipschitz and bounded away from zero on $[v_L, v_H]$, and r_S is bounded and Lipschitz on this interval. Also suppose X_i is bounded and its first and second conditional moments given (V_i, Z_i) satisfy Assumption 4(2). Appendix F verifies Assumption 3 and verifies Assumption 4 under P_{sel}. The three laws of large numbers then follow from lemma 1.

Define the full-sample limits

$$\begin{aligned} Q^S &= 2\mathbb{E}[D_i \operatorname{Var}(X_i \mid V_i, D_i = 1)], \\ G^S &= 2\mathbb{E}[D_i \lambda_0'(V_i) \operatorname{Cov}(X_i, W_i \mid V_i, D_i = 1)], \\ \Omega^S &= p_1 \Omega_{\text{sel}}, \end{aligned}$$

where Ω_{sel} is the oracle covariance in (7) evaluated under P_{sel}. Under the maintained assumptions and the sufficient conditions above, theorem 1 gives

$$\sqrt{n}(\widehat{\beta} - \beta_0) \xrightarrow{d} \mathcal{N}\Big(0, (Q^S)^{-1}\{\Omega^S + G^S \Omega_\psi (G^S)^\mathsf{T}\}(Q^S)^{-1}\Big). \tag{31}$$

Our procedure can use the first-stage index estimator in Liu and Yu (2022). In the second stage, that paper estimates the outcome control function λ_0 by isotonic regression under a monotonicity restriction. We instead apply fixed differences to observations ordered by the estimated index. This requires smoothness of λ_0, but not monotonicity. When the first stage requires no sample-size-dependent tuning parameter, neither does our point estimator.

When λ_0 satisfies both the smoothness and monotonicity restrictions, we can compare the asymptotic variances of the two estimators under their respective remaining assumptions. Let $\widehat{\beta}_{\text{OD}}$ and $\widehat{\beta}_{\text{LY}}$ denote the estimators constructed using the same first-stage index estimator. Comparing our result with Theorem 4.2 in Liu and Yu (2022) shows that the correction induced by estimating γ_0 is identical for the two estimators. For simplicity, assume conditional homoskedasticity. Let V^o denote the oracle variance component of the Liu–Yu estimator and let V^γ denote the common generated-index component. Then

$$\operatorname{Avar}\Big\{\sqrt{n}(\widehat{\beta}_{\text{LY}} - \beta_0)\Big\} = V^o + V^\gamma,$$

whereas first differencing gives

$$\mathrm{Avar}\Big\{\sqrt{n}(\widehat{\beta}_{\mathrm{OD}} - \beta_0)\Big\} = \frac{3}{2}V^o + V^\gamma.$$

Thus ordered differencing pays the familiar efficiency cost in its oracle component, while the first-stage contribution is unchanged.

For differencing of order K, using the corresponding optimal weights in Hall, Kay, and Titterington (1990) replaces the factor $3/2$ with $1 + 1/(2K)$. Choosing a larger fixed order therefore reduces the variance penalty. These weights also reduce the additional variance under heteroskedasticity, as shown in the discussion following corollary 2. The first-stage contribution remains unchanged.

Thus, when the monotonicity restriction on λ_0 is correct, the procedure in Liu and Yu (2022) provides a more efficient point estimator without a sample-size-dependent tuning parameter, under the conditions of the comparison above. When monotonicity fails but our smoothness and other maintained assumptions hold, ordered differencing remains consistent with a fixed differencing order.

4.2 Homogeneous Treatment Effects

Let $D_i \in \{0, 1\}$ be a treatment, let $(Y_i(0), Y_i(1))$ be potential outcomes, and let $Y_i = (1 - D_i)Y_i(0) + D_iY_i(1)$. Under the framework in Rosenbaum and Rubin (1983), assume

$$(Y_i(0), Y_i(1)) \perp D_i \mid X_i, \qquad 0 < e(X_i) < 1, \tag{32}$$

where $e(x) = \mathbb{P}(D_i = 1 \mid X_i = x)$.

First suppose that treatment effects are homogeneous,

$$Y_i(1) = Y_i(0) + \beta_0, \qquad Y_i(0) = \mu_0(X_i) + U_i, \qquad \mathbb{E}(U_i \mid X_i) = 0. \tag{33}$$

Suppose

$$e(X_i) = F(X_i^{\mathsf{T}}\alpha_0), \tag{34}$$

where F is strictly increasing but need not be known. The first stage estimates α_0 using treatment indicators and pretreatment covariates, following the separation of propensity-score estimation from outcome analysis discussed in Rubin (2007). The chosen estimator must satisfy the first-stage conditions stated below.

By the balancing property of the propensity score (Rosenbaum and Rubin, 1983) and strict monotonicity of F,

$$\mathbb{E}(Y_i \mid D_i, X_i^{\mathsf{T}}\alpha_0) = \beta_0 D_i + g_0(X_i^{\mathsf{T}}\alpha_0), \tag{35}$$

where $g_0(v) \equiv \mathbb{E}\{Y_i(0) \mid X_i^{\mathsf{T}}\alpha_0 = v\}$.

Equation (35) conditions on treatment and the scalar index. It does not imply the stronger restriction $\mathbb{E}\{Y_i(0) - g_0(V_i) \mid D_i, X_i\} = 0$ used by the general theorem, with X_i here playing the role of the index covariates. To apply that theorem directly, we additionally impose

$$\mathbb{E}\{Y_i(0) \mid X_i\} = g_0(X_i^{\mathsf{T}}\alpha_0). \tag{36}$$

This is an outcome-mean restriction beyond unconfoundedness and a single-index propensity score. Extending the asymptotic-normality argument to unrestricted $\mu_0(X_i)$ requires a separate argument, which we do not pursue here.

Let $\widehat{\pi}$ denote the permutation obtained by ordering the observations according to $X_i^{\mathsf{T}}\widehat{\alpha}$. The feasible ordered-difference estimator is

$$\widehat{\beta}_T \equiv \left\{ \frac{1}{n}\sum_{j=2}^{n} (\Delta_{\widehat{\pi}} D_j)^2 \right\}^{-1} \frac{1}{n}\sum_{j=2}^{n} \Delta_{\widehat{\pi}} D_j \Delta_{\widehat{\pi}} Y_j. \tag{37}$$

Its construction requires the estimated ordering index but does not require estimation of either F or g_0.

To apply theorem 1, we maintain Assumptions 1 and 2 with D_i as the finite-dimensional regressor, X_i as the index covariates, and α_0 as the index coefficient. For Assumption 3, suppose $\widehat{\alpha}$ uses only (D_i, X_i) and satisfies (8), with α replacing γ, for a centered, square-integrable influence function $\psi_T(D_i, X_i)$.

Appendix G gives sufficient density and conditional-variance smoothness conditions and verifies Assumptions 3 and 4 for this model.

In this application, $\mathbb{E}(D_i \mid X_i) = F(V_i)$ implies $\mathrm{Cov}(D_i, X_i \mid V_i) = 0$, so $G = 0$. Also,

$$Q = 2\mathbb{E}[F(V_i)\{1 - F(V_i)\}] > 0$$

under (32). Therefore theorem 1 gives

$$\sqrt{n}(\widehat{\beta}_T - \beta_0) \xrightarrow{d} \mathcal{N}(0, \Omega/Q^2), \tag{38}$$

where Ω is given by (7) with D_i as the regressor and $\varepsilon_i = Y_i(0) - g_0(V_i)$ as the error. This result maintains the outcome-mean restriction (36) and homogeneous treatment effects.

Remark 2 (Heterogeneous treatment effects)**.** *Under heterogeneous effects, suppose* (32) *holds, F is Lipschitz, and the generated-spacing result and the required law of large numbers for consecutive ordered observations hold. Then the ordered-difference estimator is consistent for*

$$\beta_\omega = \frac{\mathbb{E}[e(X_i)\{1 - e(X_i)\}\tau(X_i)]}{\mathbb{E}[e(X_i)\{1 - e(X_i)\}]}, \qquad \tau(x) = \mathbb{E}\{Y_i(1) - Y_i(0) \mid X_i = x\}. \tag{39}$$

This is the overlap-weighted estimand obtained by the propensity-score-residual regression in Lee (2018). The calculation is given in Appendix H. We do not establish asymptotic normality or the validity of subsampling inference for this heterogeneous-effects estimand.

5 Simulation Study

We compare fixed ordered differences with kernel partialling-out in a partially linear model with an artificial first stage. The experiment examines the efficiency cost of differencing, the effect of estimating the index, and subsampling coverage under homoskedasticity and heteroskedasticity.

Following (1)–(2), we generate data from

$$Y_i = X_i\theta_0 + g_0(V_i) + \varepsilon_i, \qquad V_i = W_i^{\mathsf{T}}\gamma_0,$$

with a scalar regressor X_i, $\theta_0 = 1$, $W_i = (V_i, Z_i)^{\mathsf{T}}$, and $\gamma_0 = (1, 0)^{\mathsf{T}}$. The parameter of interest is θ_0. Draw V_i, Z_i, ξ_i independently from Unif$[-1, 1]$, and e_i, ν_i independently from $N(0, 1)$, independently of the covariates. Set

$$X_i = m_0(V_i) + Z_i + \xi_i, \qquad m_0(v) = v + \tfrac{1}{2}\sin(\pi v),$$
$$g_0(v) = 2v + \tfrac{1}{2}\sin(\pi v), \qquad \varepsilon_i = s(Z_i)e_i.$$

The component $m_0(V_i)$ makes X_i correlated with $g_0(V_i)$, while $Z_i + \xi_i$ provides variation in X_i conditional on V_i that identifies θ_0. The shared component Z_i makes first-stage estimation contribute to the limiting variance. Design H has $s^2(z) = 1$ and Design V has $s^2(z) = 1/2+3z^2/2$. The second-stage error satisfies $\mathbb{E}(\varepsilon_i \mid X_i, W_i) = 0$ and $\mathrm{Var}(\varepsilon_i) = 1$ in both designs.

For the artificial first stage, generate the auxiliary response $A_i = V_i + \nu_i/\sqrt{3}$. Normalize the first coordinate of γ_0 to one and estimate its second coordinate by regressing $A_i - V_i$ on Z_i without an intercept. This gives

$$\widehat{\gamma}_2 = \frac{\sum_i Z_i(A_i - V_i)}{\sum_i Z_i^2}, \qquad \widehat{\gamma} = (1, \widehat{\gamma}_2)^\mathsf{T}, \qquad \widehat{V}_i = W_i^\mathsf{T}\widehat{\gamma}.$$

This first-stage estimator satisfies Assumption 3, as required by theorem 1.

We use sample sizes $n \in \{1000, 2000, 4000\}$ and conduct 1,000 replications for each n. We report bias and RMSE for ordered differencing with $K \in \{1, 2, 5\}$, using the corresponding optimal weights in Hall, Kay, and Titterington (1990). We compare these estimates with the partialling-out estimator in Robinson (1988), implemented using Gaussian local-linear regressions. Its bandwidth $\widehat{h}$ is selected by cross-validation. We also use $0.5\widehat{h}$ and $2\widehat{h}$ to examine sensitivity to the bandwidth choice. All methods use common samples. The simulation appendix gives the implementation, known-index comparisons, and variance diagnostics.

Table 1: Bias and RMSE with an estimated index

			Ordered differences			Kernel bandwidth		
Design	n		$K = 1$	$K = 2$	$K = 5$	$0.5\widehat{h}$	$\widehat{h}$	$2\widehat{h}$
H	1000	Bias	-0.0008	-0.0005	-0.0003	-0.0004	0.0002	0.0073
H	1000	RMSE	0.0595	0.0557	0.0528	0.0516	0.0513	0.0518
H	2000	Bias	-0.0003	-0.0006	-0.0009	-0.0007	-0.0004	0.0041
H	2000	RMSE	0.0393	0.0374	0.0356	0.0350	0.0349	0.0350
H	4000	Bias	0.0005	0.0002	0.0002	0.0002	0.0005	0.0032
H	4000	RMSE	0.0295	0.0279	0.0264	0.0260	0.0259	0.0260
V	1000	Bias	-0.0009	-0.0005	0.0000	-0.0001	0.0006	0.0077
V	1000	RMSE	0.0621	0.0585	0.0557	0.0544	0.0542	0.0546
V	2000	Bias	-0.0004	-0.0008	-0.0010	-0.0008	-0.0005	0.0040
V	2000	RMSE	0.0411	0.0393	0.0377	0.0370	0.0369	0.0369
V	4000	Bias	0.0006	0.0003	0.0003	0.0002	0.0004	0.0032
V	4000	RMSE	0.0309	0.0293	0.0279	0.0275	0.0274	0.0276

Notes. All methods use the same fitted index. Ordered differences use $K + 1$ observations per difference. Kernel bandwidth multipliers are fixed before the experiment. The true outcome coefficient is one.

Increasing the differencing order reduces RMSE in both designs and brings it closer to the kernel benchmark. At $n = 4000$, $K = 5$ has RMSE 0.0264 in H and 0.0279 in V, compared with 0.0259 and 0.0274 for kernel partialling-out at the cross-validated bandwidth. Bias is small for all three differencing orders. Doubling the kernel bandwidth raises bias, but RMSE remains similar across the bandwidth multipliers examined here.

The theoretical fitted-index variance is $2.5 + 0.75/K$ in H and $2.8 + 0.75/K$ in V, including a first-stage contribution of one. For $K = 5$ at $n = 4000$, the empirical scaled variances are 2.79 and 3.12, compared with theoretical values 2.65 and 2.95. Gaussian intervals using these population variances cover in 94.6 and 95.2 percent of replications. These are infeasible diagnostics of the variance calculation.

We assess 95 percent subsampling intervals for all three differencing orders. Each replication draws $B = n$ subsets of size $b = \text{round}(\sqrt{n})$ and refits both stages. We report both the equal-tailed interval in equation (26) and a symmetric interval using the empirical 0.95 quantile of the absolute centered subsampling statistic. Both constructions were specified before inspecting these results.

Table 2: Subsampling coverage and interval length

			Coverage (%)		Mean length	
Design	n	K	Equal-tail	Absolute	Equal-tail	Absolute
H	1000	1	91.5	93.7	0.222	0.224
H	1000	2	92.3	93.9	0.205	0.209
H	1000	5	86.7	96.5	0.188	0.215
H	2000	1	94.6	95.8	0.158	0.159
H	2000	2	94.6	95.3	0.147	0.149
H	2000	5	92.8	96.1	0.138	0.146
H	4000	1	94.2	95.4	0.112	0.113
H	4000	2	93.6	94.8	0.105	0.105
H	4000	5	93.6	95.0	0.100	0.102
V	1000	1	91.4	93.5	0.229	0.231
V	1000	2	91.1	93.8	0.214	0.218
V	1000	5	87.1	95.9	0.197	0.224
V	2000	1	94.2	95.5	0.163	0.164
V	2000	2	94.2	95.0	0.153	0.155
V	2000	5	92.4	95.8	0.145	0.153
V	4000	1	93.7	94.2	0.116	0.117
V	4000	2	93.4	94.5	0.110	0.110
V	4000	5	93.5	95.3	0.105	0.107

Notes. $b = \text{round}(\sqrt{n})$ and $B = n$. Coverage is computed over 1,000 replications. Monte Carlo standard error at 95 percent coverage is approximately 0.7 percentage points.

Absolute-quantile intervals generally give coverage closer to the nominal 95 percent level in the smaller samples. In the larger samples, coverage from both constructions is near the target, although equal-tailed intervals retain some undercoverage. Overall, absolute-quantile

intervals perform better in these simulation designs. Coverage is relatively insensitive to the two subsample sizes examined, especially for absolute-quantile intervals and at larger n. The main exception is equal-tailed inference with $K = 5$ at $n = 1000$, where changing b has a more noticeable effect. The appendix reports both subsample sizes in full and retains the additional sample-selection comparisons.

Despite its simplicity, fixed-order differencing performs well relative to kernel partialling-out. Higher-order differences deliver comparable RMSE without estimating auxiliary conditional means or selecting a second-stage bandwidth.

6 Empirical illustration: point estimates

We compare ordered differencing with other semiparametric estimators using the 2024 one-year American Community Survey Public Use Microdata Sample (ACS-PUMS, U.S. Census Bureau, 2024) for California. We restrict the data to married women ages 25–60 who belong to the householder's family and have observed family income. Person and housing records are joined by household identifier, and income amounts are adjusted to 2024 dollars. Wages are observed for women with positive wage income, weeks worked, and usual weekly hours. Hourly earnings are annual wage income divided by weeks times weekly hours, and records with hourly earnings outside \$2–\$500 are excluded. We use all 50,164 eligible records, of whom 35,180 have observed wages. The analysis is unweighted.

Selection means having an observed positive wage and reporting at least 35 usual weekly hours during the past 12 months. There are 27,166 selected women. We estimate the two-step sample-selection model in section 4.1, with log hourly earnings as Y_i. The wage regressors X_i include potential experience divided by ten and its square, indicators for below-high-school education, some college, an associate degree, a bachelor's degree, and an advanced degree, and indicators for Black, Asian, and Hispanic women. A high-school diploma or GED is the reference education category. Potential experience is the maximum of age minus assigned years of schooling minus six and zero.

The selection regressors W_i contain all variables in X_i and four excluded variables. These are the numbers of household children aged 0–2, 3–5, and 6–13, and the inverse hyperbolic sine of other-family income divided by \$50,000. The child counts refer to all children in the household, not necessarily the woman's own children. Other-family income is family income minus the woman's personal income. The specification assumes that the excluded variables affect full-time wage employment but have no direct effect on mean log hourly earnings beyond the wage regressors and selection control function.

We compare ordered differencing with the partialling-out estimator in Robinson (1988), implemented using Gaussian local-linear regressions with cross-validated and rule-of-thumb bandwidths, and the isotonic outcome-regression estimator in Liu and Yu (2022). All three use the same selection index, computed using the tuning-free first-stage procedure in Liu and Yu (2022). Parametric benchmarks are the sample-selection model in Heckman (1979), estimated by full-information maximum likelihood under bivariate normal outcome and selection disturbances, and OLS on the selected observations. Every estimator was computed without numerical errors.

Table 3 reports the point estimates. First differences and fixed $K = 5$ differences give bachelor's-degree coefficients of 0.8593 and 0.8474, compared with 0.8456 for isotonic outcome regression and 0.8345 for kernel partialling-out. Gaussian selection estimation and OLS give 0.7445 and 0.6610. These coefficients measure log-wage differences relative to women with a

Table 3: ACS wage equation: point estimates

	OD $K=1$	OD $K=5$	Isotonic	Kernel CV	Kernel ROT	Heckman	OLS
Experience / 10	0.1363	0.1617	0.1561	0.1533	0.1532	0.2159	0.2340
Experience squared / 100	-0.0248	-0.0310	-0.0294	-0.0285	-0.0284	-0.0385	-0.0405
Below high school	-0.2752	-0.2726	-0.2926	-0.2836	-0.2835	-0.2524	-0.2070
Some college	0.2196	0.2162	0.2183	0.2133	0.2134	0.1937	0.1640
Associate degree	0.3527	0.3480	0.3486	0.3409	0.3409	0.3026	0.2532
Bachelor degree	0.8593	0.8474	0.8456	0.8345	0.8345	0.7445	0.6610
Advanced degree	1.2047	1.1946	1.1973	1.1827	1.1827	1.0214	0.9054
Black	-0.0626	-0.0504	-0.0449	-0.0539	-0.0540	-0.1171	-0.1531
Asian	0.0210	0.0171	0.0170	0.0156	0.0156	-0.0029	-0.0127
Hispanic	-0.1419	-0.1514	-0.1530	-0.1552	-0.1552	-0.1875	-0.2063

Notes: OD denotes ordered differencing with fixed weights. Isotonic outcome regression fits both monotonicity directions and selects the smaller residual sum of squares, giving a decreasing fit here. Kernel CV and ROT use leave-one-out local-linear regressions with a common bandwidth selected by cross-validation or a global-quartic rule of thumb. OLS includes an intercept.

high-school diploma or GED, conditional on the other regressors and the selection correction where applicable.

All methods imply initially increasing, concave experience profiles. Among the semiparametric methods, coefficients on experience divided by ten range from 0.1363 to 0.1617, and those on its square range from -0.0310 to -0.0248.

Figure 1 compares sensitivity to the kernel bandwidth and the differencing order on a common vertical scale. Kernel bandwidths range from 0.1 to 10 times the cross-validated choice of 0.1861 in standardized-index units. The bachelor's-degree coefficient ranges from 0.8343 to 0.8493, a spread of 0.0150. The rule-of-thumb bandwidth is 0.1826, giving a coefficient of 0.8345, almost identical to CV.

For ordered differencing, we consider $K \in \{1, 2, 3, 4, 5, 8, 10\}$ with the corresponding optimal weights in Hall, Kay, and Titterington (1990). The bachelor's-degree coefficient declines from 0.8593 at $K = 1$ to 0.8388 at $K = 10$, a spread of 0.0205. Over these choices, differencing is more sensitive to its order than the kernel estimator is to its bandwidth. The comparison supports the simplicity of fixed-order differencing, but not a claim of greater stability. The order remains a researcher choice, even when it is held fixed as the sample size grows.

We construct the rule-of-thumb bandwidth using global quartic pilot regressions and the local-linear AMISE expression discussed in Ruppert, Sheather, and Wand (1995). The outcome and each regressor are standardized, and a common bandwidth minimizes the sum of their estimated leading integrated squared bias and variance terms.[1]

Table 4 reports computation times. Across the seven orders, ordered differencing takes between 0.0019 and 0.0032 seconds for the second stage. Kernel partialling-out takes 57.01 seconds with CV and 12.14 seconds with ROT, including bandwidth selection. The common first stage takes 4.23 seconds and accounts for almost all of the computation time of the ordered-difference estimator. These timings demonstrate a substantial computational advantage

[1]Let q_k be the quartic pilot for standardized auxiliary target k, let $\widehat{\sigma}_k^2$ be its residual sum of squares divided by $n_1 - 5$, and put $\widehat{\theta}_{22,k} = n_1^{-1} \sum_i \{q_k''(v_i)\}^2$, where v_i is the standardized fitted index among selected observations. For the Gaussian kernel, we use $h_{\mathrm{ROT}} = \{(\max_i v_i - \min_i v_i) \sum_k \widehat{\sigma}_k^2 / (2\sqrt{\pi}\, n_1 \sum_k \widehat{\theta}_{22,k})\}^{1/5}$. This is a global-quartic pilot rule using a constant-variance approximation for each auxiliary regression.

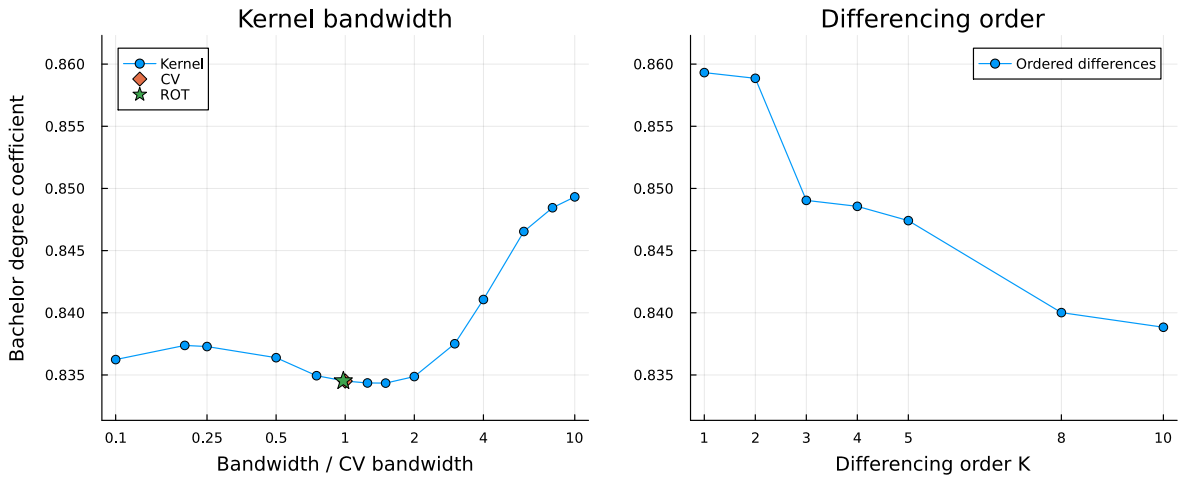


Figure 1: Sensitivity of the bachelor's-degree coefficient to bandwidth and differencing order

Notes: The kernel method uses the same bandwidth rule as in Section 5: a common bandwidth minimizes the average standardized leave-one-out prediction loss for the outcome and regressors over 17 candidates. The left panel scales that bandwidth on a logarithmic horizontal axis and marks CV and the global-quartic rule of thumb (ROT). The right panel varies K using the corresponding optimal weights. Both panels use the same vertical scale, first-stage index, and selected sample. A difference of order K combines $K+1$ consecutive observations and yields $n_1 - K$ differences. At 0.1 times CV, two numerically singular local-linear fits use local-constant fallbacks. No observations are dropped.

for differencing in our implementation.

Ordered differencing produces education and experience estimates similar to the other semiparametric procedures using a fixed differencing rule. It requires neither monotonicity of the outcome control function nor estimation of auxiliary conditional means with a second-stage bandwidth.

7 Conclusion

Ordered differencing provides a simple second stage for partially linear models with an estimated scalar index. The estimator sorts observations, applies fixed differences, and runs least squares without estimating auxiliary conditional-mean regressions. First-stage estimation can change many neighboring pairs. Our spacing result controls the resulting smooth nuisance remainders, and conditional centering given the first-stage information permits a central limit argument for the actual fitted ordering. We verify the required moment limits under bounded primitive conditions and obtain subsampling inference under the stated assumptions.

Sample selection is the main application. A first stage based on selection indicators and covariates is compatible with the maintained selected-outcome conditional mean. Appendix F verifies the first-stage information restriction and fitted-order limits under sufficient population conditions. The outcome control function can be smooth and nonmonotone. Fixed differencing therefore provides an alternative to isotonic outcome regression and to smoothing-based control-function estimation, with an efficiency cost when those procedures exploit valid additional information.

The simulations compare fixed higher-order differences with kernel partialling-out under an

Table 4: Computation time for the ACS application

Method	Seconds
First stage	4.232
OD $K = 1$	0.002
OD $K = 2$	0.002
OD $K = 3$	0.002
OD $K = 4$	0.002
OD $K = 5$	0.002
OD $K = 8$	0.003
OD $K = 10$	0.003
Kernel CV	57.008
Kernel ROT	12.140

Notes: Median wall-clock seconds over three repetitions on an Apple M2 Pro, after compilation. Second-stage timings use the same fitted index and 27,166 selected observations and exclude the common first stage, which uses all 50,164 women. Ordered differencing includes sorting and least squares. Kernel timings include bandwidth selection and the final fit. All methods have the same six-thread allocation, with CV candidates evaluated in parallel. Data loading and plotting are excluded. The kernel implementation evaluates exact leave-one-out local-linear fits without binning.

artificial first stage. Higher-order differences approach kernel accuracy in both variance designs, and the empirical variances are close to the theoretical values. Equal-tailed subsampling intervals can under-cover, especially with small subsamples. Symmetric absolute-quantile intervals give higher coverage and are longer. The results underscore the role of the subsample-size choice in finite-sample inference. The wage-equation illustration compares point estimates under a common fitted index. Extending the primitive verification to unbounded indices and improving finite-sample inference remain useful directions for further work.

References

Aradillas-López, A., Honoré, B. E., and Powell, J. L. (2007). Pairwise difference estimation with nonparametric control variables. *International Economic Review* **48**, 1119–1158.

Ahn, H. and Powell, J. L. (1993). Semiparametric estimation of censored selection models with a nonparametric selection mechanism. *Journal of Econometrics* **58**, 3–29.

Abadie, A. and Imbens, G. W. (2006). Large sample properties of matching estimators for average treatment effects. *Econometrica* **74**, 235–267.

Abadie, A. and Imbens, G. W. (2008). On the failure of the bootstrap for matching estimators. *Econometrica* **76**, 1537–1557.

Abadie, A. and Imbens, G. W. (2016). Matching on the estimated propensity score. *Econometrica* **84**, 781–807.

Andreou, E. and Werker, B. J. M. (2012). An alternative asymptotic analysis of residual-based statistics. *Review of Economics and Statistics* **94**, 88–99.

Billingsley, P. (1995). *Probability and Measure*, 3rd ed. Wiley, New York.

Bulinski, A. V. (2017). Conditional central limit theorem. *Theory of Probability and Its Applications* **61**, 613–631.

Chen, H. (1988). Convergence rates for parametric components in a partly linear model. *The Annals of Statistics* **16**, 136–146.

Chen, L. H. Y. and Shao, Q.-M. (2004). Normal approximation under local dependence. *The Annals of Probability* **32**, 1985–2028.

Chamberlain, G. (1992). Efficiency bounds for semiparametric regression. *Econometrica* **60**, 567–596.

Dai, W., Tong, T., and Zhu, L. (2017). On the choice of difference sequence in a unified framework for variance estimation in nonparametric regression. *Statistical Science* **32**, 455–468.

Dette, H., Munk, A., and Wagner, T. (1998). Estimating the variance in nonparametric regression—what is a reasonable choice? *Journal of the Royal Statistical Society, Series B* **60**, 751–764.

Donald, S. G., Imbens, G. W., and Newey, W. K. (2009). Choosing instrumental variables in conditional moment restriction models. *Journal of Econometrics* **152**, 28–36.

U.S. Census Bureau (2024). American Community Survey one-year Public Use Microdata Sample. Washington, DC: U.S. Department of Commerce.

Groeneboom, P. and Hendrickx, K. (2018). Current status linear regression. *The Annals of Statistics* **46**, 1415–1444.

Hahn, J. (1998). On the role of the propensity score in efficient semiparametric estimation of average treatment effects. *Econometrica* **66**, 315–331.

Hall, P., Kay, J. W., and Titterington, D. M. (1990). Asymptotically optimal difference-based estimation of variance in nonparametric regression. *Biometrika* **77**, 521–528.

Han, A. K. (1987). Non-parametric analysis of a generalized regression model: the maximum rank correlation estimator. *Journal of Econometrics* **35**, 303–316.

Heckman, J. J. (1979). Sample selection bias as a specification error. *Econometrica* **47**, 153–161.

Heckman, J. J. and Vytlacil, E. J. (2007). Econometric evaluation of social programs, Part I: Causal models, structural models, and econometric policy evaluation. In *Handbook of Econometrics*, Vol. 6B, J. J. Heckman and E. E. Leamer, eds., Elsevier, 4779–4874.

Hirano, K., Imbens, G. W., and Ridder, G. (2003). Efficient estimation of average treatment effects using the estimated propensity score. *Econometrica* **71**, 1161–1189.

Honoré, B. E. and Powell, J. L. (2005). Pairwise difference estimators for nonlinear models. In *Identification and Inference for Econometric Models: Essays in Honor of Thomas Rothenberg*, D. W. K. Andrews and J. H. Stock, eds., Cambridge University Press, 520–553.

Ichimura, H. (1993). Semiparametric least squares (SLS) and weighted SLS estimation of single-index models. *Journal of Econometrics* **58**, 71–120.

Imai, K. and Ratkovic, M. (2014). Covariate balancing propensity score. *Journal of the Royal Statistical Society, Series B* **76**, 243–263.

Klein, R. W. and Spady, R. H. (1993). An efficient semiparametric estimator for binary response models. *Econometrica* **61**, 387–421.

Lee, M.-J. (2018). Simple least squares estimator for treatment effects using propensity score residuals. *Biometrika* **105**, 149–164.

Lewbel, A. and Schennach, S. M. (2007). A simple ordered data estimator for inverse density weighted expectations. *Journal of Econometrics* **136**, 189–211.

Liu, R. and Yu, Z. (2022). Sample selection models with monotone control functions. *Journal of Econometrics* **226**, 321–342.

McDiarmid, C. (1989). On the method of bounded differences. In *Surveys in Combinatorics, 1989*, J. Siemons, ed., Cambridge University Press, 148–188.

Newey, W. K. (1990). Efficient instrumental variables estimation of nonlinear models. *Econometrica* **58**, 809–837.

Newey, W. K. (2009). Two-step series estimation of sample selection models. *The Econometrics Journal* **12**, S217–S229.

Newey, W. K. and McFadden, D. (1994). Large sample estimation and hypothesis testing. In *Handbook of Econometrics*, Vol. 4, R. F. Engle and D. L. McFadden, eds., Elsevier, 2111–2245.

Newey, W. K., Powell, J. L., and Vella, F. (1999). Nonparametric estimation of triangular simultaneous equations models. *Econometrica* **67**, 565–603.

Pagan, A. (1984). Econometric issues in the analysis of regressions with generated regressors. *International Economic Review* **25**, 221–247.

Politis, D. N., Romano, J. P., and Wolf, M. (1999). *Subsampling*. Springer-Verlag, New York.

Powell, J. L. (2001). Semiparametric estimation of censored selection models. In *Nonlinear Statistical Modeling*, C. Hsiao, J. L. Powell, and K. Morimune, eds., Cambridge University Press, 165–196.

Powell, J. L., Stock, J. H., and Stoker, T. M. (1989). Semiparametric estimation of index coefficients. *Econometrica* **57**, 1403–1430.

Rice, J. (1984). Bandwidth choice for nonparametric regression. *The Annals of Statistics* **12**, 1215–1230.

Robinson, P. M. (1988). Root-N-consistent semiparametric regression. *Econometrica* **56**, 931–954.

Rosenbaum, P. R. and Rubin, D. B. (1983). The central role of the propensity score in observational studies for causal effects. *Biometrika* **70**, 41–55.

Rubin, D. B. (2007). The design versus the analysis of observational studies for causal effects: parallels with the design of randomized trials. *Statistics in Medicine* **26**, 20–36.

Ruppert, D., Sheather, S. J., and Wand, M. P. (1995). An effective bandwidth selector for local least squares regression. *Journal of the American Statistical Association* **90**, 1257–1270.

Sherman, R. P. (1993). The limiting distribution of the maximum rank correlation estimator. *Econometrica* **61**, 123–137.

van der Vaart, A. W. (1991). On differentiable functionals. *The Annals of Statistics* **19**, 178–204.

van der Vaart, A. W. (1998). *Asymptotic Statistics*. Cambridge University Press, Cambridge.

van der Vaart, A. W. and Wellner, J. A. (1996). *Weak Convergence and Empirical Processes: With Applications to Statistics*. Springer, New York.

van de Geer, S. A. (2000). *Empirical Processes in M-Estimation*. Cambridge University Press, Cambridge.

Xu, M. and Otsu, T. (2020). Score estimation of monotone partially linear index models. *Journal of Nonparametric Statistics* **32**, 838–863.

Yatchew, A. (1997). An elementary estimator of the partial linear model. *Economics Letters* **57**, 135–143.

Yatchew, A. (2003). *Semiparametric Regression for the Applied Econometrician*. Cambridge University Press, Cambridge.

A Proof of proposition 1 (the oracle proposition)

Proof of proposition 1. For readability, write (j) for $\pi_0(j)$ and $\Delta a_j = a_{(j)} - a_{(j-1)}$. Conditional on the complete vector of order statistics $\mathcal{O}_n = (V_{(1)}, \ldots, V_{(n)})$, the concomitants $(X_{(j)}, \varepsilon_{(j)})$ are independent. They are not identically distributed because

$$X_{(j)} = m_0(V_{(j)}) + \xi_{(j)},$$

their conditional distributions generally vary with the realized order statistic $V_{(j)}$. By the definition of m_0,

$$\mathbb{E}(\xi_{(j)} \mid V_{(j)}) = 0.$$

Conditional independence across observations then implies $\mathbb{E}(\xi_{(j)} \mid \mathcal{O}_n) = 0$. In addition, (2) and the tower property give

$$\mathbb{E}(\varepsilon_{(j)} \mid X_{(j)}, \mathcal{O}_n) = \mathbb{E}\big[\mathbb{E}(\varepsilon_{(j)} \mid X_{(j)}, W_{(j)}, \mathcal{O}_n) \mid X_{(j)}, \mathcal{O}_n\big] = 0.$$

Before treating the design and score, we record the spacing bound used in both steps. Let F_V be the distribution function of V and put $U_i = F_V(V_i)$. Under Assumption 1, the U_i are independent uniform random variables on $[0,1]$. Write $f_* = \inf_{v \in \mathcal{V}} f_V(v) > 0$. For $v < w$ in $\mathcal{V}$,

$$F_V(w) - F_V(v) = \int_v^w f_V(t)\, dt \geq f_*(w - v).$$

Consequently the inverse distribution function is Lipschitz, and

$$\max_{2 \leq j \leq n} (V_{(j)} - V_{(j-1)}) \leq f_*^{-1} \max_{1 \leq j \leq n+1} (U_{(j)} - U_{(j-1)}),$$

where $U_{(0)} = 0$ and $U_{(n+1)} = 1$. Each uniform spacing has tail probability $(1-t)^n$ for $0 < t < 1$. A union bound therefore gives

$$\Pr\left\{ \max_j (U_{(j)} - U_{(j-1)}) > t \right\} \leq (n+1)(1-t)^n \leq (n+1)e^{-nt}.$$

Taking $t = C \log(n)/n$ with fixed $C > 1$ proves

$$\max_{2 \leq j \leq n} (V_{(j)} - V_{(j-1)}) = O_{\mathbb{P}}\{\log(n)/n\} = O_{\mathbb{P}}(n^{-1+\eta}) \tag{A1}$$

for every fixed $\eta > 0$.

Step 1. The design-matrix limit.

We first establish the limit of Q_n^0. Since $X_i = m_0(V_i) + \xi_i$, we have

$$\Delta X_j = \Delta \xi_j + \Delta m_j, \qquad \Delta m_j = m_0(V_{(j)}) - m_0(V_{(j-1)}).$$

It follows that

$$\Delta X_j \Delta X_j^{\mathsf{T}} = \Delta \xi_j \Delta \xi_j^{\mathsf{T}} + \Delta \xi_j \Delta m_j^{\mathsf{T}} + \Delta m_j \Delta \xi_j^{\mathsf{T}} + \Delta m_j \Delta m_j^{\mathsf{T}}.$$

By Assumption 2, for some finite Lipschitz constant L_m, we have

$$\|m_0(v) - m_0(\widetilde{v})\| \leq L_m |v - \widetilde{v}|.$$

Because the observations are ordered by V, the consecutive gaps are nonnegative and their sum equals $V_{(n)} - V_{(1)}$. Hence

$$\begin{aligned}\sum_{j=2}^{n} \|\Delta m_j\|^2 &\le L_m^2 \sum_{j=2}^{n} (V_{(j)} - V_{(j-1)})^2 \\ &\le L_m^2 \max_{2\le j\le n} (V_{(j)} - V_{(j-1)}) \sum_{j=2}^{n} (V_{(j)} - V_{(j-1)}) \\ &= L_m^2 \max_{2\le j\le n} (V_{(j)} - V_{(j-1)})(V_{(n)} - V_{(1)}) \\ &= O_{\mathbb{P}}(n^{-1+\eta}).\end{aligned} \tag{A2}$$

The last equality in (A2) follows from (A1) and compactness of $\mathcal{V}$.

We next consider the leading term

$$B_n = \frac{1}{n} \sum_{j=2}^{n} \Delta\xi_j \Delta\xi_j^{\mathsf{T}}.$$

For each j, conditional on $\mathcal{O}_n$,

$$\begin{aligned}\mathbb{E}(\Delta\xi_j \Delta\xi_j^{\mathsf{T}} \mid \mathcal{O}_n) = {} & \mathbb{E}(\xi_{(j)}\xi_{(j)}^{\mathsf{T}} \mid \mathcal{O}_n) + \mathbb{E}(\xi_{(j-1)}\xi_{(j-1)}^{\mathsf{T}} \mid \mathcal{O}_n) \\ & - \mathbb{E}(\xi_{(j)}\xi_{(j-1)}^{\mathsf{T}} \mid \mathcal{O}_n) - \mathbb{E}(\xi_{(j-1)}\xi_{(j)}^{\mathsf{T}} \mid \mathcal{O}_n) \\ = {} & \Sigma_X(V_{(j)}) + \Sigma_X(V_{(j-1)}).\end{aligned}$$

The two cross terms are zero because conditional independence gives $\mathbb{E}(\xi_{(j)}\xi_{(j-1)}^{\mathsf{T}} \mid \mathcal{O}_n) = \mathbb{E}(\xi_{(j)} \mid \mathcal{O}_n)\mathbb{E}(\xi_{(j-1)}^{\mathsf{T}} \mid \mathcal{O}_n) = 0$. Therefore,

$$\mathbb{E}(B_n \mid \mathcal{O}_n) = \frac{1}{n} \sum_{j=2}^{n} \{\Sigma_X(V_{(j)}) + \Sigma_X(V_{(j-1)})\}. \tag{A3}$$

Conditional on $\mathcal{O}_n$, the summands defining B_n are one-dependent. We control their conditional variances through an average fourth moment. Put

$$\kappa_4(v) = \mathbb{E}(\|\xi_i\|^4 \mid V_i = v).$$

Since m_0 is bounded on the compact support under Assumption 2, Assumption 1 implies $\mathbb{E}\kappa_4(V_i) = \mathbb{E}\|\xi_i\|^4 < \infty$. For any scalar entry Z_j of $\Delta\xi_j\Delta\xi_j^{\mathsf{T}}$,

$$\mathbb{E}(Z_j^2 \mid \mathcal{O}_n) \le \mathbb{E}(\|\Delta\xi_j\|^4 \mid \mathcal{O}_n) \le C\{\kappa_4(V_{(j)}) + \kappa_4(V_{(j-1)})\}.$$

Only consecutive Z_j can have nonzero conditional covariance. Bounding each covariance by the arithmetic mean of the two conditional second moments therefore gives, for each entry (a, b),

$$\mathbb{V}\{(B_n)_{ab} \mid \mathcal{O}_n\} \le \frac{C}{n^2} \sum_{i=1}^{n} \kappa_4(V_i) = O_{\mathbb{P}}(n^{-1}),$$

where the final equality follows from the ordinary LLN. No uniform bound on $\kappa_4(v)$ is needed. Conditional Chebyshev's inequality consequently gives

$$B_n - \mathbb{E}(B_n \mid \mathcal{O}_n) \xrightarrow{p} 0. \tag{A4}$$

Finally, the right-hand side of (A3) can be written as

$$\frac{2}{n}\sum_{j=1}^{n} \Sigma_X(V_j) - \frac{1}{n}\{\Sigma_X(V_{(1)}) + \Sigma_X(V_{(n)})\}.$$

The first term converges to $2\mathbb{E}\{\Sigma_X(V_i)\}$ by the ordinary LLN, while the second is $o_{\mathbb{P}}(1)$ because $\Sigma_X(v) = \mathbb{E}(X_i X_i^{\mathsf{T}} \mid V_i = v) - m_0(v)m_0(v)^{\mathsf{T}}$ is uniformly bounded under Assumption 1. Hence

$$B_n \xrightarrow{p} 2\mathbb{E}\{\Sigma_X(V_i)\} = Q.$$

It remains to handle the three terms containing Δm_j. Equation (A2), the convergence of B_n, and Cauchy–Schwarz make each of them $o_{\mathbb{P}}(1)$ after division by n. This proves the design-matrix limit.

Step 2. The limiting distribution of S_n.

We first consider the nuisance-function remainder

$$R_{g,n} \equiv n^{-1/2}\sum_{j=2}^{n} \Delta X_j \Delta g_j, \qquad \Delta g_j \equiv g_0(V_{(j)}) - g_0(V_{(j-1)}).$$

The first-order condition for the oracle estimator gives the exact relationship

$$\sqrt{n}(\widetilde{\theta} - \theta_0) = (Q_n^0)^{-1}(S_n + R_{g,n}). \tag{A5}$$

Using $\Delta X_j = \Delta\xi_j + \Delta m_j$, decompose this remainder as

$$R_{g,n} = R_{\xi g,n} + R_{mg,n}, \tag{A6}$$

where

$$R_{\xi g,n} \equiv \frac{1}{\sqrt{n}}\sum_{j=2}^{n} \Delta\xi_j \Delta g_j,$$

$$R_{mg,n} \equiv \frac{1}{\sqrt{n}}\sum_{j=2}^{n} \Delta m_j \Delta g_j.$$

As in (A2), Lipschitz continuity of g_0, compactness of the support of V_i, and (A1) give

$$\sum_{j=2}^{n} |\Delta g_j|^2 = O_{\mathbb{P}}(n^{-1+\eta})$$

for every fixed $\eta > 0$. We first bound $R_{\xi g,n}$. Conditional on $\mathcal{O}_n$, the quantities Δg_j are fixed and $\mathbb{E}(\Delta\xi_j \mid \mathcal{O}_n) = 0$. Therefore,

$$\mathbb{E}(R_{\xi g,n} \mid \mathcal{O}_n) = 0.$$

The summands are M-dependent with $M = 1$ conditional on $\mathcal{O}_n$. Uniform boundedness of their conditional second moments and $2|\Delta g_j \Delta g_{j+1}| \leq |\Delta g_j|^2 + |\Delta g_{j+1}|^2$ give, for some finite constant C,

$$\|\mathbb{V}(R_{\xi g,n} \mid \mathcal{O}_n)\| \leq \frac{C}{n} \sum_{j=2}^{n} |\Delta g_j|^2 = O_{\mathbb{P}}(n^{-2+\eta}).$$

Conditional Chebyshev's inequality gives

$$R_{\xi g,n} = O_{\mathbb{P}}(n^{-1+\eta/2}). \tag{A7}$$

We next bound $R_{mg,n}$. Cauchy–Schwarz gives

$$\begin{aligned} \|R_{mg,n}\| &\leq \frac{1}{\sqrt{n}} \left(\sum_{j=2}^{n} \|\Delta m_j\|^2 \right)^{1/2} \left(\sum_{j=2}^{n} |\Delta g_j|^2 \right)^{1/2} \\ &= O_{\mathbb{P}}(n^{-3/2+\eta}). \end{aligned}$$

Combining this bound with (A7) in (A6) gives

$$R_{g,n} = O_{\mathbb{P}}(n^{-1+\eta/2}) \tag{A8}$$

for any fixed $\eta \in (0, 1)$. In particular, $R_{g,n} = o_{\mathbb{P}}(1)$. Combining this result with (A5) gives

$$\sqrt{n}(\widetilde{\theta} - \theta_0) = (Q_n^0)^{-1} S_n + o_{\mathbb{P}}(1). \tag{A9}$$

The same argument used to obtain (A7), now using $\mathbb{E}(\varepsilon_{(j)} \mid X_{(j)}, \mathcal{O}_n) = 0$, shows that

$$n^{-1/2} \sum_{j=2}^{n} \Delta m_j \Delta \varepsilon_j = O_{\mathbb{P}}(n^{-1+\eta/2}). \tag{A10}$$

Consequently,

$$S_n = n^{-1/2} \sum_{j=2}^{n} \Delta \xi_j \Delta \varepsilon_j + o_{\mathbb{P}}(1). \tag{A11}$$

Let $a_{n,j} \equiv \Delta \xi_j \Delta \varepsilon_j$. Conditional on $\mathcal{O}_n$, the array $\{a_{n,j} \mid j = 2, \ldots, n\}$ is M-dependent with $M = 1$ and mean zero. Write

$$\Sigma_j \equiv \Sigma_X(V_{(j)}), \qquad \sigma_j^2 \equiv \sigma^2(V_{(j)}), \qquad M_j \equiv M(V_{(j)}).$$

Expanding $a_{n,j} a_{n,j}^{\mathsf{T}}$ and using conditional independence and the conditional mean-zero restrictions gives

$$\mathbb{E}(a_{n,j} a_{n,j}^{\mathsf{T}} \mid \mathcal{O}_n) = M_j + M_{j-1} + \Sigma_{j-1} \sigma_j^2 + \Sigma_j \sigma_{j-1}^2. \tag{A12}$$

The terms $a_{n,j}$ and $a_{n,j+1}$ both depend on $(\xi_{(j)}, \varepsilon_{(j)})$. Conditioning first on this common observation gives

$$\mathbb{E}(a_{n,j} a_{n,j+1}^{\mathsf{T}} \mid \mathcal{O}_n) = M_j. \tag{A13}$$

Terms constructed from nonoverlapping pairs of observations are conditionally independent.

The spacing bound in (A1) also controls the variation in the neighboring conditional moments. By Assumption 1, there is a finite constant L that bounds the Lipschitz constants of $M(v)$, $\Sigma_X(v)$, and $\sigma^2(v)$. It follows that

$$\begin{aligned}\max_{2\le j\le n}\|M_j - M_{j-1}\| &\le L\max_{2\le j\le n}|V_{(j)} - V_{(j-1)}| = O_{\mathbb{P}}(n^{-1+\eta}),\\ \max_{2\le j\le n}\|\Sigma_j - \Sigma_{j-1}\| &= O_{\mathbb{P}}(n^{-1+\eta}),\\ \max_{2\le j\le n}|\sigma_j^2 - \sigma_{j-1}^2| &= O_{\mathbb{P}}(n^{-1+\eta}).\end{aligned}$$

For example, uniform boundedness of $\Sigma_X(v)$ gives

$$\begin{aligned}&\left\|\frac{1}{n}\sum_{j=2}^{n}\Sigma_{j-1}(\sigma_j^2 - \sigma_{j-1}^2)\right\|\\ &\qquad\le \sup_{v\in\mathcal{V}}\|\Sigma_X(v)\|\max_{2\le j\le n}|\sigma_j^2 - \sigma_{j-1}^2| = O_{\mathbb{P}}(n^{-1+\eta}).\end{aligned}$$

Hence the average of $\Sigma_{j-1}\sigma_j^2$ has the same probability limit as the average of $\Sigma_{j-1}\sigma_{j-1}^2$. The same argument applies to $\Sigma_j\sigma_{j-1}^2$. The ordinary LLN then gives

$$\frac{1}{n}\sum_{j=2}^{n}(M_j + M_{j-1}) \xrightarrow{p} 2\mathbb{E}\{M(V_i)\}$$

and

$$\frac{1}{n}\sum_{j=2}^{n}(\Sigma_{j-1}\sigma_j^2 + \Sigma_j\sigma_{j-1}^2) \xrightarrow{p} 2\mathbb{E}\{\Sigma_X(V_i)\sigma^2(V_i)\}.$$

Averaging (A12) therefore gives

$$\frac{1}{n}\sum_{j=2}^{n}\mathbb{E}(a_{n,j}a_{n,j}^{\mathsf{T}} \mid \mathcal{O}_n) \xrightarrow{p} 2\mathbb{E}\{M(V_i) + \Sigma_X(V_i)\sigma^2(V_i)\}. \tag{A14}$$

Averaging (A13) gives

$$\frac{1}{n}\sum_{j=2}^{n-1}\mathbb{E}(a_{n,j}a_{n,j+1}^{\mathsf{T}} \mid \mathcal{O}_n) \xrightarrow{p} \mathbb{E}\{M(V_i)\}. \tag{A15}$$

Since the array is M-dependent with $M = 1$ and has conditional mean zero, its conditional variance is

$$\begin{aligned}&\mathbb{V}\left(\frac{1}{\sqrt{n}}\sum_{j=2}^{n}a_{n,j} \mid \mathcal{O}_n\right)\\ &\quad= \frac{1}{n}\sum_{j=2}^{n}\mathbb{E}(a_{n,j}a_{n,j}^{\mathsf{T}} \mid \mathcal{O}_n)\\ &\qquad+ \frac{1}{n}\sum_{j=2}^{n-1}\left\{\mathbb{E}(a_{n,j}a_{n,j+1}^{\mathsf{T}} \mid \mathcal{O}_n) + \mathbb{E}(a_{n,j+1}a_{n,j}^{\mathsf{T}} \mid \mathcal{O}_n)\right\}.\end{aligned} \tag{A16}$$

The two covariance terms have the same symmetric probability limit. Substituting (A14) and (A15) into (A16) gives Ω.

Put $p = 2 + \min(\eta_s/2, 1) \in (2, 3]$, and define

$$u_j = \mathbb{E}(\|\xi_{(j)}\|^{2p} \mid \mathcal{O}_n), \qquad w_j = \mathbb{E}(|\varepsilon_{(j)}|^{2p} \mid \mathcal{O}_n).$$

Conditional Cauchy–Schwarz gives

$$\mathbb{E}(\|a_{n,j}\|^p \mid \mathcal{O}_n) \le C(u_j + u_{j-1})^{1/2}(w_j + w_{j-1})^{1/2}.$$

Summing and applying Cauchy–Schwarz again yields

$$\frac{1}{n}\sum_{j=2}^{n} \mathbb{E}(\|a_{n,j}\|^p \mid \mathcal{O}_n) \le C\left(\frac{1}{n}\sum_{j=1}^{n} u_j\right)^{1/2}\left(\frac{1}{n}\sum_{j=1}^{n} w_j\right)^{1/2} = O_{\mathbb{P}}(1).$$

Both averages are $O_{\mathbb{P}}(1)$ by the LLN for their conditional-moment functions evaluated at the unordered V_i. Their expectations are finite under Assumptions 1 and 2, since $2p \le 4 + \eta_s$.

For a fixed vector t with $\sigma_t^2 = t^{\mathsf{T}}\Omega t > 0$, put

$$\sigma_{n,t}^2 = \mathbb{V}\left(n^{-1/2}\sum_{j=2}^{n} t^{\mathsf{T}} a_{n,j} \mid \mathcal{O}_n\right) \xrightarrow{p} \sigma_t^2.$$

On the event $\sigma_{n,t}^2 > \sigma_t^2/2$, normalize each scalar summand by $\sqrt{n}\,\sigma_{n,t}$. Their sum has conditional variance one and is one-dependent. Theorem 2.6 in Chen and Shao (2004), applied under the conditional law, bounds its distance from the standard normal distribution by

$$C\frac{n^{1-p/2}\|t\|^p}{\sigma_{n,t}^p}\left\{\frac{1}{n}\sum_{j=2}^{n}\mathbb{E}(\|a_{n,j}\|^p \mid \mathcal{O}_n)\right\} = o_{\mathbb{P}}(1).$$

The complementary event has probability tending to zero. When $\sigma_t^2 = 0$, conditional Chebyshev's inequality gives convergence of the projection to zero. Thus every fixed projection has the required conditional Gaussian limit. Averaging the bounded conditional characteristic functions and applying the Cramér–Wold device gives

$$n^{-1/2}\sum_{j} a_{n,j} \xrightarrow{d} \mathcal{N}(0, \Omega).$$

The conclusion follows from (A8)–(A11), the design-matrix limit, and Slutsky's theorem. □

B Proof of proposition 2 (the generated-index spacing proposition)

Put $h_n \equiv \widehat{\gamma} - \gamma_0$. Since $\widehat{V}_i = V_i + W_i^{\mathsf{T}} h_n$, we have

$$\Delta_{\widehat{\pi}} V_j = \Delta_{\widehat{\pi}} \widehat{V}_j - \Delta_{\widehat{\pi}} W_j^{\mathsf{T}} h_n.$$

It follows that

$$\frac{1}{n}\sum_{j=2}^{n}(\Delta_{\widehat{\pi}}V_j)^2 \le \frac{2}{n}\sum_{j=2}^{n}(\Delta_{\widehat{\pi}}\widehat{V}_j)^2 + 2\|h_n\|^2\frac{1}{n}\sum_{j=2}^{n}\|\Delta_{\widehat{\pi}}W_j\|^2. \tag{B1}$$

By construction, $\Delta_{\widehat{\pi}}\widehat{V}_j \ge 0$. Therefore

$$\begin{aligned}\sum_{j=2}^{n}(\Delta_{\widehat{\pi}}\widehat{V}_j)^2 &\le \left(\sum_{j=2}^{n}\Delta_{\widehat{\pi}}\widehat{V}_j\right)^2 \\ &= \left(\max_{1\le i\le n}\widehat{V}_i - \min_{1\le i\le n}\widehat{V}_i\right)^2. \end{aligned}\tag{B2}$$

The range of the estimated index satisfies

$$\max_i \widehat{V}_i - \min_i \widehat{V}_i \le \overline{v} - \underline{v} + 2\|h_n\|\max_i\|W_i\|.$$

Under Assumption 1, $\max_i\|W_i\| = o_{\mathbb{P}}(n^{1/(4+\eta_s)})$. To see this, put $p = 4+\eta_s$. For every $c>0$,

$$\begin{aligned}\mathbb{P}\left(n^{-1/p}\max_i\|W_i\| > c\right) &\le n\mathbb{P}(\|W_i\| > cn^{1/p}) \\ &= c^{-p}(cn^{1/p})^p\mathbb{P}(\|W_i\| > cn^{1/p}) \longrightarrow 0,\end{aligned}$$

where the convergence follows from

$$t^p\mathbb{P}(\|W_i\| > t) \le \mathbb{E}[\|W_i\|^p 1\{\|W_i\| > t\}] \longrightarrow 0.$$

Since $\|h_n\| = O_{\mathbb{P}}(n^{-1/2})$ and $1/(4+\eta_s) < 1/2$, the estimated-index range is $O_{\mathbb{P}}(1)$. Equation (B2) consequently gives

$$\frac{1}{n}\sum_{j=2}^{n}(\Delta_{\widehat{\pi}}\widehat{V}_j)^2 = O_{\mathbb{P}}(n^{-1}). \tag{B3}$$

The same calculation holds uniformly over the shrinking neighborhood. For $\gamma\in\Gamma_n(\delta)$, put $h = \gamma - \gamma_0$ and $U_i(\gamma) = V_i + W_i^{\mathsf{T}}h$. Replacing $\widehat{\gamma}$ by γ in (B1) gives

$$\frac{1}{n}\sum_{j=2}^{n}(\Delta_{\pi_\gamma}V_j)^2 \le \frac{2}{n}\sum_{j=2}^{n}(\Delta_{\pi_\gamma}U(\gamma)_j)^2 + 2\|h\|^2\frac{1}{n}\sum_{j=2}^{n}\|\Delta_{\pi_\gamma}W_j\|^2.$$

Uniformly over $\gamma\in\Gamma_n(\delta)$, the first term is $O_{\mathbb{P}}(n^{-1})$ because the ordered increments of $U(\gamma)$ are nonnegative and its range is $O_{\mathbb{P}}(1)$. The second term is $O_{\mathbb{P}}(n^{-1})$ because $\sup_{\gamma\in\Gamma_n(\delta)}\|h\|^2 \le \delta^2/n$ by construction of $\Gamma_n(\delta)$, and permutation invariance bounds its remaining average by $4n^{-1}\sum_i\|W_i\|^2 = O_{\mathbb{P}}(1)$. This proves (13).

We also need a maximal-spacing bound. Put

$$e_n(\delta) = \delta n^{-1/2}\max_i\|W_i\| = o_{\mathbb{P}}(1), \qquad M_n^V = \max_{2\le j\le n}\{V_{\pi_0(j)} - V_{\pi_0(j-1)}\}.$$

Under Assumption 1, $M_n^V = o_{\mathbb{P}}(1)$. Let π_u and π_v be permutations that sort $(u_1, \ldots, u_n)$ and $(v_1, \ldots, v_n)$, respectively, in nondecreasing order. Sorting is nonexpansive in the sup norm, meaning that

$$\max_{1 \le i \le n} |u_i - v_i| \le e \quad \Longrightarrow \quad \max_{1 \le j \le n} |u_{\pi_u(j)} - v_{\pi_v(j)}| \le e.$$

Applying this fact to $U_i(\gamma)$ and V_i, whose sorting permutations are π_γ and π_0, respectively, gives

$$\sup_{\gamma \in \Gamma_n(\delta)} \max_j \Delta_{\pi_\gamma} U(\gamma)_j \le M_n^V + 2e_n(\delta) = o_{\mathbb{P}}(1). \tag{B4}$$

Since the estimated-index range is $O_{\mathbb{P}}(1)$,

$$\sum_{j=2}^{n} (\Delta_{\widehat{\pi}} \widehat{V}_j)^2 \le \max_{2 \le j \le n} \Delta_{\widehat{\pi}} \widehat{V}_j \left(\widehat{V}_{\widehat{\pi}(n)} - \widehat{V}_{\widehat{\pi}(1)} \right) = o_{\mathbb{P}}(1),$$

which proves (11). Finally,

$$\max_{2 \le j \le n} |\Delta_{\widehat{\pi}} V_j| \le \max_{2 \le j \le n} \Delta_{\widehat{\pi}} \widehat{V}_j + 2\|h_n\| \max_i \|W_i\| = o_{\mathbb{P}}(1),$$

which proves (12).

For the second term in (B1),

$$\|\Delta_{\widehat{\pi}} W_j\|^2 \le 2\|W_{\widehat{\pi}(j)}\|^2 + 2\|W_{\widehat{\pi}(j-1)}\|^2.$$

Because $\widehat{\pi}$ is a permutation, the law of large numbers yields

$$\frac{1}{n} \sum_{j=2}^{n} \|\Delta_{\widehat{\pi}} W_j\|^2 \le \frac{4}{n} \sum_{i=1}^{n} \|W_i\|^2 = O_{\mathbb{P}}(1).$$

Together with $\|h_n\|^2 = O_{\mathbb{P}}(n^{-1})$, this shows that the second term in (B1) is $O_{\mathbb{P}}(n^{-1})$. Combining this bound with (B3) proves (10).

C Primitive verification of the fitted-order limits

We prove lemma 1. Throughout this proof, we use C and c to denote generic finite positive constants, reusing the notation for constants whose values may change from one occurrence to the next. These constants may depend on the fixed dimensions and population bounds, but not on n or on h. Write $T_n(h)$ for any scalar entry of the matrices $Q_n(h)$, $G_n(h)$, or $\mathcal{V}_n(h)$, formed by ordering $U_{h,i} = V_i + h^\mathsf{T} Z_i$. Here h is a deterministic (nonrandom) perturbation of the free index coefficients, changing observation i's true index V_i by $h^\mathsf{T} Z_i$. We first establish concentration for each deterministic h, then extend it uniformly over nearby coefficients to permit evaluation at the random first-stage estimate. In $G_n(h)$ the derivative remains $g_0'(V_{\pi_h(j)})$, as in the main text. Each scalar entry of the summands defining $Q_n(h)$, $G_n(h)$, and $\mathcal{V}_n(h)$ is uniformly bounded over observations and nearby coefficients h, under Assumptions 2 and 4.

Concentration at a fixed coefficient. Our first goal is to show that, for each fixed deterministic h,

$$T_n(h) - \mathbb{E}T_n(h) \xrightarrow{p} 0.$$

We apply McDiarmid's bounded-differences inequality by showing that replacing one observation changes $T_n(h)$ by at most C/n. First note that, for deterministic h, replacing one observation leaves the relative order of all other observations unchanged. Removing the original observation and inserting its replacement therefore changes only a bounded number of consecutive pairs, since the differencing order is fixed. For example, for $\mathcal{V}_n(h)$, we have

$$\mathcal{V}_n(h) = \frac{1}{n}\sum_{i=1}^{n} v_i c_{n,i}(h) c_{n,i}(h)^{\mathsf{T}}, \qquad c_{n,i}(h) = 2X_i - X_{i^-} - X_{i^+}$$

for an interior observation i, where i^- and i^+ denote its immediate predecessor and successor in the ordering induced by h. At an endpoint, the coefficient is X_i minus its sole neighbor's regressor vector. To see the effect of replacement, consider five consecutive observations a, b, i, c, d. Removing i makes b and c consecutive, so their coefficients change as follows.

$$\begin{aligned} c_{n,b}(h) = 2X_b - X_a - X_i \quad &\longrightarrow \quad 2X_b - X_a - X_c = c_{n,b}(h) + X_i - X_c, \\ c_{n,c}(h) = 2X_c - X_i - X_d \quad &\longrightarrow \quad 2X_c - X_b - X_d = c_{n,c}(h) + X_i - X_b. \end{aligned}$$

For instance, with $\delta_b = X_i - X_c$, the variance summand for observation b changes by

$$v_b\big\{(c_{n,b}(h) + \delta_b)(c_{n,b}(h) + \delta_b)^{\mathsf{T}} - c_{n,b}(h)c_{n,b}(h)^{\mathsf{T}}\big\}.$$

The summand for i is removed, and all other summands except those for b and c remain unchanged at this deletion step. Inserting the replacement observation similarly changes only the coefficients of its new neighbors and adds its own summand, using its new regressor vector and conditional variance. Endpoint deletions and insertions affect no more observations. Thus replacement changes only a bounded number of summands in $\mathcal{V}_n(h)$, each uniformly bounded because X_i and v_i are bounded under Assumption 4(2). Division by n therefore bounds the change in each matrix entry by C/n. Consequently every $T_n(h)$ has bounded differences C/n as a function of the independent observations. The inequality in McDiarmid (1989) gives

$$\Pr\{|T_n(h) - \mathbb{E}T_n(h)| > t\} \le 2\exp(-cnt^2).$$

This is an unconditional concentration argument at a deterministic perturbation h. It does not condition on a data-dependent ordering cell.

A deterministic grid and nearly tied observations. We next extend the concentration result uniformly over a fixed neighborhood of the true first-stage coefficient γ_0, corresponding to $h = 0$. To do so, we approximate this neighborhood in the perturbation coordinates h by a deterministic grid $\mathcal{H}_n(\gamma_0)$ with spacing $\delta_n = n^{-3}$. Its cardinality is polynomial in n, since γ_0 belongs to a parameter space of fixed finite dimension. With $q-1$ free coefficients, the grid has $O(n^{3(q-1)})$ points. For each h, choose a nearest grid point $\bar{h} \in \mathcal{H}_n(\gamma_0)$, so that $\|h - \bar{h}\| \le C_q\delta_n$, where C_q depends only on the fixed dimension. Compactness of the support of Z_i, imposed in Assumption 4(1), gives $\|Z_i\| \le M_Z$. Consequently,

$$|U_{h,i} - U_{\bar{h},i}| = |(h - \bar{h})^{\mathsf{T}} Z_i| \le \|h - \bar{h}\|\,\|Z_i\| \le C_q M_Z \delta_n.$$

Taking the maximum over observations yields $\max_i |U_{h,i} - U_{\bar{h},i}| \le K\delta_n$, with $K = C_q M_Z$ independent of n and h. For a pair of observations i, j whose projected gap at $\bar{h}$ exceeds $2K\delta_n$, moving from $\bar{h}$ to h cannot change their relative order. This follows because each index moves by at most $K\delta_n$. Thus only pairs whose projected gap at $\bar{h}$ is at most $2K\delta_n$ can change order. To control how many observations can be affected by the grid approximation, collect the observations belonging to at least one such pair and denote their number by

$$N_n(\bar{h}) = \#\{i : \ |U_{\bar{h},i} - U_{\bar{h},k}| \le 2K\delta_n \text{ for some } k \ne i\}.$$

Recall that $p(v, z)$ denotes the joint density of (V, Z) relative to $dv\,\mu(dz)$ introduced in Assumption 4(1). The density of the perturbed scalar index $U_{h,i} = V_i + h^\mathsf{T} Z_i$ is

$$f_h(u) = \int p(u - h^\mathsf{T} z, z) 1\{v_L \le u - h^\mathsf{T} z \le v_H\}\,\mu(dz),$$

which is uniformly bounded because p is bounded and μ is a finite measure under Assumption 4(1). Let M_f be a uniform upper bound for $f_h(u)$ over u and the coefficient neighborhood. For a fixed grid point $\bar{h}$ and distinct observations i, k, independence across observations from Assumption 1 gives

$$\begin{aligned}\Pr\big\{|U_{\bar{h},i} - U_{\bar{h},k}| \le 2K\delta_n \mid U_{\bar{h},i} = u\big\} \\ = \int_{u-2K\delta_n}^{u+2K\delta_n} f_{\bar{h}}(s)\,ds \le 4KM_f\delta_n.\end{aligned}$$

The same bound holds unconditionally. Observation i is counted in $N_n(\bar{h})$ if at least one other observation $k \ne i$ has its perturbed index $U_{\bar{h},k}$ in the interval

$$[U_{\bar{h},i} - 2K\delta_n,\ U_{\bar{h},i} + 2K\delta_n].$$

A union bound therefore yields

$$\begin{aligned}\mathbb{E}N_n(\bar{h}) &= \sum_{i=1}^n \Pr\big\{\exists k \ne i : |U_{\bar{h},i} - U_{\bar{h},k}| \le 2K\delta_n\big\} \\ &\le \sum_{i=1}^n \sum_{k\ne i} \Pr\big\{|U_{\bar{h},i} - U_{\bar{h},k}| \le 2K\delta_n\big\} \\ &\le 4KM_f n(n-1)\delta_n \le Cn^2\delta_n.\end{aligned}$$

The bound $Cn^2\delta_n$ does not depend on $\bar{h}$, since M_f and K are uniform over the coefficient neighborhood. Hence

$$\sup_{\bar{h}\in\mathcal{H}_n(\gamma_0)} \mathbb{E}N_n(\bar{h}) \le Cn^2\delta_n.$$

Uniform control of the count over the grid. Our next goal is to show that

$$\max_{\bar{h}\in\mathcal{H}_n(\gamma_0)} \frac{N_n(\bar{h})}{n} \xrightarrow{p} 0.$$

The preceding step bounds the expectation at every grid point $\bar{h} \in \mathcal{H}_n(\gamma_0)$. We now control the realized counts simultaneously over the growing grid, using McDiarmid's inequality and a union

bound. We show that the count $N_n(\bar{h})$ itself has a bounded-differences constant independent of n. Let $r = 2K\delta_n$. An observation is counted in $N_n(\bar{h})$ when another observation's index lies within distance r of its own index. Suppose a new observation with index u is inserted. An existing observation can become newly counted only if its index lies in $[u-r, u+r]$ and it previously had no other index within distance r. There can be at most two such observations. To see this, suppose there were three, with indices $x_1 < x_2 < x_3$. Since they were previously uncounted, their gaps satisfy

$$x_2 - x_1 > r, \qquad x_3 - x_2 > r, \qquad \text{and hence} \qquad x_3 - x_1 > 2r.$$

This is impossible because all three indices lie in $[u-r, u+r]$, an interval of length $2r$. Insertion therefore changes the count by at most three, allowing for the inserted observation itself to be counted. Deletion is the reverse operation and also changes the count by at most three. Replacing one observation consists of a deletion followed by an insertion, so the absolute change in $N_n(\bar{h})$ is at most six, independently of n and $\bar{h}$. For any fixed $\epsilon > 0$, the bound $\sup_{\bar{h}} \mathbb{E}N_n(\bar{h}) \le Cn^2\delta_n = o(n)$ implies that these expectations are at most $n\epsilon/2$ for all sufficiently large n. The replacement bound six and the same inequality in McDiarmid (1989) then give

$$\Pr\left\{\max_{\bar{h}\in\mathcal{H}_n(\gamma_0)} N_n(\bar{h}) > n\epsilon\right\} \le 2|\mathcal{H}_n(\gamma_0)| \exp(-n\epsilon^2/72) \longrightarrow 0.$$

The exponential bound dominates the polynomial grid size. Hence

$$\max_{\bar{h}\in\mathcal{H}_n(\gamma_0)} N_n(\bar{h})/n = o_p(1).$$

Uniform concentration of the matrix entries over the grid. We separately apply the concentration bound for $T_n(h)$ established in the fixed-coefficient step. For each deterministic grid point $\bar{h}$,

$$\Pr\big\{|T_n(\bar{h}) - \mathbb{E}T_n(\bar{h})| > \epsilon\big\} \le 2\exp(-cn\epsilon^2),$$

where $c > 0$ does not depend on the grid point, because the replacement bound C/n is uniform over the coefficient neighborhood. For every fixed $\epsilon > 0$, a union bound gives

$$\begin{aligned}
&\Pr\left\{\max_{\bar{h}\in\mathcal{H}_n(\gamma_0)} |T_n(\bar{h}) - \mathbb{E}T_n(\bar{h})| > \epsilon\right\} \\
&\le \sum_{\bar{h}\in\mathcal{H}_n(\gamma_0)} \Pr\big\{|T_n(\bar{h}) - \mathbb{E}T_n(\bar{h})| > \epsilon\big\} \\
&\le 2|\mathcal{H}_n(\gamma_0)| \exp(-cn\epsilon^2) \\
&\le Cn^{3(q-1)} \exp(-cn\epsilon^2) \longrightarrow 0.
\end{aligned}$$

The exponential decay dominates the polynomial grid size. Consequently,

$$\max_{\bar{h}\in\mathcal{H}_n(\gamma_0)} |T_n(\bar{h}) - \mathbb{E}T_n(\bar{h})| = o_p(1).$$

Extension from the grid to the coefficient neighborhood. We now use the uniform grid bounds to control $T_n(h)$ at every coefficient in the neighborhood, including coefficients between grid points. If two observations reverse order between h and $\bar{h}$, the gap between their scalar index values at $\bar{h}$ is at most $2K\delta_n$. Hence every observation outside the set counted by $N_n(\bar{h})$ has the same rank in both orderings. Only a constant multiple of $N_n(\bar{h})$ edge or neighbor summands can change, including the summands of the conditional variance. Deterministic tie breaking causes no problem because any tie at h also satisfies this gap bound. It follows that

$$|T_n(h) - T_n(\bar{h})| \leq CN_n(\bar{h})/n.$$

Taking expectations also gives a uniform bound $Cn\delta_n = o(1)$ for $|\mathbb{E}T_n(h) - \mathbb{E}T_n(\bar{h})|$. Combining these statements proves

$$\sup_h |T_n(h) - \mathbb{E}T_n(h)| = o_p(1) \tag{C1}$$

on the fixed cube. In particular, the supremum covers a coefficient estimated from these same observations.

Convergence of the expectations. In this step, we show that, for every deterministic sequence $h_n \to 0$,

$$\mathbb{E}Q_n(h_n) \to Q, \qquad \mathbb{E}G_n(h_n) \to G, \qquad \mathbb{E}\mathcal{V}_n(h_n) \to \Omega.$$

The expectations involve products of conditional moments for neighboring observations, evaluated at different index values. For example, write $m_n(u) = \mathbb{E}(X \mid U_{h_n} = u)$ and recall that $m_0(u) = \mathbb{E}(X \mid V = u)$. For two neighboring observations with perturbed index values u, u' in the interior of the true index support, conditional independence under Assumption 1, justified below, expresses their regressor cross-product expectation conditional on all index values as $m_n(u)m_n(u')^\mathsf{T}$. The corresponding expression at a common true-index value is $m_0(u)m_0(u)^\mathsf{T}$. To show that $m_n(u)m_n(u')^\mathsf{T}$ converges to $m_0(u)m_0(u)^\mathsf{T}$ for neighboring index values, we control the differences between the corresponding factors. In particular, by the triangle inequality,

$$\|m_n(u') - m_0(u)\| \leq \underbrace{\|m_n(u') - m_0(u')\|}_{\text{change of conditioning index}} + \underbrace{\|m_0(u') - m_0(u)\|}_{\text{different neighboring index values}}. \tag{C2}$$

In Step 1 below, we control the first term uniformly over an interior interval I_ϵ, defined at the start of that step. This also controls $\|m_n(u) - m_0(u)\|$ for the other factor in the product. In Step 2, we control $|u' - u|$ using the sorting argument from proposition 2. Lipschitz continuity then bounds the second term by $L|u' - u|$. The same two controls apply to the other conditional moments entering the matrix expectations. In Step 3, we combine these results and handle observations outside I_ϵ by letting $\epsilon \downarrow 0$.

Step 1. Uniform control of the change of conditioning index. Fix a deterministic sequence $h_n \to 0$ and an interior interval $I_\epsilon = [v_L + \epsilon, v_H - \epsilon]$, with $0 < \epsilon < (v_H - v_L)/2$ fixed. Since $h_n \to 0$ and $\|z\| \leq M_Z$ on the support of Z under Assumption 4(1), for all sufficiently large n we have $|h_n^\mathsf{T} z| \leq \|h_n\| M_Z < \epsilon$ uniformly over z. Thus, for every $u \in I_\epsilon$,

$$v_L \leq u - h_n^\mathsf{T} z \leq v_H,$$

so the indicator in the formula for f_{h_n} equals one throughout I_ϵ. The density and conditional-moment restrictions in Assumption 4(1)–(2) therefore imply that f_{h_n} is bounded away from zero on I_ϵ, and that the conditional moments of $X, XX^\mathsf{T}, v, vX, vXX^\mathsf{T}$ given $U_{h_n} = u$ are uniformly Lipschitz on I_ϵ. As $n \to \infty$, these conditional moments given $U_{h_n} = u$ converge uniformly over $u \in I_\epsilon$ to the corresponding conditional moments given $V = u$. For instance, let $m_X(v, z) = \mathbb{E}(X \mid V = v, Z = z)$ and define

$$A_h(u) = \int m_X(u - h^\mathsf{T} z, z)\, p(u - h^\mathsf{T} z, z)\, \mu(dz).$$

For $u \in I_\epsilon$ and all sufficiently large n, the support indicator equals one, so

$$\mathbb{E}(X \mid U_{h_n} = u) = \frac{A_{h_n}(u)}{f_{h_n}(u)}, \qquad \mathbb{E}(X \mid V = u) = \frac{A_0(u)}{f_0(u)}.$$

Under Assumption 4(1)–(2), both $m_X(v, z)$ and $p(v, z)$ are bounded and uniformly Lipschitz in v. Their product is therefore uniformly Lipschitz in v as well. Since μ is finite and $\|z\| \le M_Z$, we obtain

$$\sup_{u \in I_\epsilon} \|A_{h_n}(u) - A_0(u)\| \le C \int |h_n^\mathsf{T} z|\, \mu(dz) \le C\|h_n\|,$$
$$\sup_{u \in I_\epsilon} |f_{h_n}(u) - f_0(u)| \le C \int |h_n^\mathsf{T} z|\, \mu(dz) \le C\|h_n\|.$$

The lower bound for f_0 in Assumption 4(1) and the second inequality imply that f_{h_n} is also bounded away from zero on I_ϵ for all sufficiently large n. Consequently,

$$\begin{aligned}
&\sup_{u \in I_\epsilon} \|\mathbb{E}(X \mid U_{h_n} = u) - \mathbb{E}(X \mid V = u)\| \\
&\quad \le \sup_{u \in I_\epsilon} \left\{ \frac{\|A_{h_n}(u) - A_0(u)\|}{f_{h_n}(u)} + \frac{\|A_0(u)\|\, |f_{h_n}(u) - f_0(u)|}{f_{h_n}(u) f_0(u)} \right\} \\
&\quad \le C\|h_n\| \longrightarrow 0.
\end{aligned}$$

The same calculation applies to XX^T, v, vX, and vXX^T using their conditional-moment restrictions in Assumption 4(2). The same statements hold for the needed moments involving W, because $W = ((V - \gamma_{0,-1}^\mathsf{T} Z)/\gamma_{0,1}, Z^\mathsf{T})^\mathsf{T}$. Thus, for each

$$R_i \in \{X_i,\ X_i X_i^\mathsf{T},\ v_i,\ v_i X_i,\ v_i X_i X_i^\mathsf{T}\},$$

the gap between the conditional moment given the perturbed index and the corresponding conditional moment given the true index satisfies

$$\sup_{u \in I_\epsilon} \|\mathbb{E}(R_i \mid U_{h_n,i} = u) - \mathbb{E}(R_i \mid V_i = u)\| \le C_R\|h_n\| \longrightarrow 0. \tag{C3}$$

The same bound holds for the needed moments involving W_i, including $R_i = W_i$ and $R_i = X_i W_i^\mathsf{T}$, under Assumption 4(1)–(2). These bounds allow us to replace conditional moments given the perturbed index by their counterparts given the true index when calculating the matrix expectation limits. Establishing those limits also requires controlling the gaps between neighboring index values, as we do next.

Step 2. Control of the distances between neighboring index values. For deterministic h_n, independence across observations under Assumption 1 implies that, conditional on $U_{h_n,1}, \ldots, U_{h_n,n}$, the vectors (X_i, W_i, v_i) are independent across i. Each vector has its conditional distribution given its own index value $U_{h_n,i}$. Conditioning preserves independence because each index $U_{h_n,i}$ is a function of observation i alone. To control the distances between consecutive perturbed index values, we reuse the sorting argument in the proof of proposition 2. Let $U_{h_n,(j)}$ and $V_{(j)}$ denote the jth order statistics of the perturbed and true index values, respectively. Boundedness of Z_i under Assumption 4(1) gives

$$\max_i |U_{h_n,i} - V_i| \leq M_Z \|h_n\| \longrightarrow 0.$$

Sorting does not increase the maximum coordinatewise difference, so

$$\max_j |U_{h_n,(j)} - V_{(j)}| \leq M_Z \|h_n\|.$$

Consequently,

$$\begin{aligned}\max_{2\leq j\leq n} \{U_{h_n,(j)} - U_{h_n,(j-1)}\} &\leq \max_{2\leq j\leq n} \{V_{(j)} - V_{(j-1)}\} + 2M_Z\|h_n\| \\ &\xrightarrow{p} 0.\end{aligned}$$

The maximal spacing of the true index tends to zero under Assumption 1, as shown in the proof of proposition 2. We use that proof's sorting argument rather than its root-n spacing conclusion. Because Z_i has bounded support, $h_n \to 0$ suffices to ensure that

$$\max_{2\leq j\leq n} \left\{U_{h_n,\pi_{h_n}(j)} - U_{h_n,\pi_{h_n}(j-1)}\right\} \xrightarrow{p} 0. \tag{C4}$$

No root-n rate for h_n is required. For example, consider two consecutive perturbed index values $u = U_{h_n,(j-1)}$ and $u' = U_{h_n,(j)}$ in I_ϵ. Lipschitz continuity of m_0 under Assumption 2 gives

$$\|m_0(u') - m_0(u)\| \leq L|u' - u| \xrightarrow{p} 0, \tag{C5}$$

uniformly over consecutive pairs with both perturbed index values in I_ϵ, by (C4). Equation (C5) controls the second term in (C2). Combining it with (C3) for $R_i = X_i$ gives

$$\|m_n(u') - m_0(u)\| \leq \|m_n(u') - m_0(u')\| + \|m_0(u') - m_0(u)\| \leq C\|h_n\| + L|u' - u| \xrightarrow{p} 0.$$

Thus the conditional mean for the neighboring observation can be replaced by $m_0(u)$ with an error tending to zero, uniformly over neighboring pairs in I_ϵ. The same argument applies to the other conditional moments under Assumption 4(1)–(2).

Step 3. Combining the controls to obtain the expectation limits. We now calculate the limiting conditional expectations of the summands using Steps 1 and 2, then average these expressions to obtain Q, G, and Ω. We also control the contribution of observations whose perturbed index values lie outside I_ϵ.

First consider the design matrix. Write $s_n(u) = \mathbb{E}(XX^\mathsf{T} \mid U_{h_n} = u)$ and $s_0(u) = \mathbb{E}(XX^\mathsf{T} \mid V = u)$, and let $\mathcal{U}_n = \sigma(U_{h_n,1}, \ldots, U_{h_n,n})$. For consecutive observations $i = \pi_{h_n}(j)$ and

$k = \pi_{h_n}(j-1)$ with index values $u, u' \in I_\epsilon$, conditional independence from Step 2 gives

$$\begin{aligned}
&\mathbb{E}\big[(X_i - X_k)(X_i - X_k)^\mathsf{T} \mid \mathcal{U}_n\big] \\
&\quad = s_n(u) + s_n(u') - m_n(u)m_n(u')^\mathsf{T} - m_n(u')m_n(u)^\mathsf{T} \\
&\quad = 2s_0(u) - 2m_0(u)m_0(u)^\mathsf{T} + r_{n,j} \\
&\quad = 2\Sigma_X(u) + r_{n,j}.
\end{aligned}$$

Here the remainder is defined by

$$r_{n,j} = s_n(u) + s_n(u') - m_n(u)m_n(u')^\mathsf{T} - m_n(u')m_n(u)^\mathsf{T} - 2\Sigma_X(u),$$

where $u = U_{h_n,\pi_{h_n}(j)}$ and $u' = U_{h_n,\pi_{h_n}(j-1)}$. Define

$$R_{n,\epsilon} = \max_{\substack{2\le j\le n \\ U_{h_n,(j)},\, U_{h_n,(j-1)} \in I_\epsilon}} \|r_{n,j}\|,$$

with $R_{n,\epsilon} = 0$ if no pair satisfies the displayed conditions. Steps 1 and 2 imply $R_{n,\epsilon} \xrightarrow{p} 0$. Boundedness of X_i under Assumption 4(2) implies $0 \le R_{n,\epsilon} \le C$, where C does not depend on n. For every $\eta > 0$,

$$\mathbb{E}R_{n,\epsilon} \le \eta + C\Pr(R_{n,\epsilon} > \eta).$$

Taking $n \to \infty$, followed by $\eta \downarrow 0$, gives $\mathbb{E}R_{n,\epsilon} \to 0$.

We next control the omitted consecutive pairs (u, u'), for which at least one of the perturbed index values u, u' lies outside I_ϵ. Let

$$B_{n,\epsilon} = \sum_{i=1}^{n} 1\{U_{h_n,i} \notin I_\epsilon\}$$

count the observations whose perturbed index values lie outside I_ϵ. Put $d_n = M_Z\|h_n\|$. The bound $|U_{h_n,i} - V_i| \le d_n$ implies

$$\{U_{h_n,i} \notin I_\epsilon\} \subseteq \{V_i < v_L + \epsilon + d_n\} \cup \{V_i > v_H - \epsilon - d_n\}.$$

Thus an observation whose perturbed index lies outside I_ϵ must have its true index within $\epsilon + d_n$ of a support endpoint. Let M_V bound the density of V_i under Assumption 1. Since $d_n \to 0$ and $\epsilon < (v_H - v_L)/2$, for all sufficiently large n,

$$\begin{aligned}
\Pr\{U_{h_n,i} \notin I_\epsilon\} &\le \Pr\{V_i < v_L + \epsilon + d_n\} + \Pr\{V_i > v_H - \epsilon - d_n\} \\
&\le M_V(\epsilon + d_n) + M_V(\epsilon + d_n) \\
&= 2M_V(\epsilon + M_Z\|h_n\|) \le C(\epsilon + \|h_n\|).
\end{aligned}$$

By definition, $I_\epsilon = [v_L + \epsilon, v_H - \epsilon]$ excludes a strip of width ϵ at each endpoint of the true index support $[v_L, v_H]$. The term d_n allows for the additional displacement caused by the index perturbation. Identical distribution across observations under Assumption 1 therefore gives

$$\frac{\mathbb{E}B_{n,\epsilon}}{n} = \Pr\{U_{h_n,1} \notin I_\epsilon\} \le C(\epsilon + \|h_n\|).$$

Each observation counted in $B_{n,\epsilon}$ belongs to at most two consecutive pairs. Therefore, at most $2B_{n,\epsilon}$ consecutive pairs have at least one perturbed index value outside I_ϵ. Boundedness of $\|X_i\| \le M_X$ under Assumption 4(2) gives

$$\left\|(X_{\pi_{h_n}(j)} - X_{\pi_{h_n}(j-1)})(X_{\pi_{h_n}(j)} - X_{\pi_{h_n}(j-1)})^\intercal\right\| \le 4M_X^2.$$

Since at most $2B_{n,\epsilon}$ pairs are omitted, their total contribution to $Q_n(h_n)$ has norm at most $8M_X^2 B_{n,\epsilon}/n$. Taking expectations therefore bounds the expected norm of this contribution by

$$8M_X^2 \frac{\mathbb{E}B_{n,\epsilon}}{n} \le C(\epsilon + \|h_n\|).$$

For each consecutive pair with both perturbed index values in I_ϵ, the later observation $i = \pi_{h_n}(j)$ satisfies $|U_{h_n,i} - V_i| \le M_Z\|h_n\|$. Lipschitz continuity of Σ_X under Assumption 1 therefore permits replacing $\Sigma_X(U_{h_n,i})$ by $\Sigma_X(V_i)$ with error at most $C\|h_n\|$. As j runs from 2 to n, $\pi_{h_n}(j)$ runs through all observations except $\pi_{h_n}(1)$. Therefore,

$$\sum_{j=2}^{n} \Sigma_X(V_{\pi_{h_n}(j)}) = \sum_{i=1}^{n} \Sigma_X(V_i) - \Sigma_X(V_{\pi_{h_n}(1)}).$$

Let

$$J_{n,\epsilon} = \{j \in \{2,\dots,n\} :\ U_{h_n,(j)}, U_{h_n,(j-1)} \in I_\epsilon\}.$$

The calculation for pairs with both perturbed index values in I_ϵ gives the sum

$$\frac{2}{n} \sum_{j \in J_{n,\epsilon}} \Sigma_X(V_{\pi_{h_n}(j)}).$$

To complete the full-sample sum, we add the terms indexed by $j \in \{2,\dots,n\} \setminus J_{n,\epsilon}$ and the term for the first observation in the ordering. Explicitly,

$$\begin{aligned}
&\frac{2}{n}\sum_{i=1}^{n} \Sigma_X(V_i) - \frac{2}{n} \sum_{j \in J_{n,\epsilon}} \Sigma_X(V_{\pi_{h_n}(j)}) \\
&\quad = \frac{2}{n} \sum_{j \in \{2,\dots,n\} \setminus J_{n,\epsilon}} \Sigma_X(V_{\pi_{h_n}(j)}) + \frac{2}{n}\Sigma_X(V_{\pi_{h_n}(1)}).
\end{aligned}$$

There are at most $2B_{n,\epsilon}$ terms in the sum on the right. Boundedness of Σ_X under Assumption 1 therefore gives

$$\left\|\frac{2}{n}\sum_{i=1}^{n} \Sigma_X(V_i) - \frac{2}{n} \sum_{j \in J_{n,\epsilon}} \Sigma_X(V_{\pi_{h_n}(j)})\right\| \le C\frac{B_{n,\epsilon} + 1}{n}.$$

Taking expectations and using the bound for $\mathbb{E}B_{n,\epsilon}/n$ yields

$$\mathbb{E}\left\|\frac{2}{n}\sum_{i=1}^{n} \Sigma_X(V_i) - \frac{2}{n} \sum_{j \in J_{n,\epsilon}} \Sigma_X(V_{\pi_{h_n}(j)})\right\| \le C\frac{\mathbb{E}B_{n,\epsilon} + 1}{n} \le C(\epsilon + \|h_n\| + n^{-1}).$$

The n^{-1} term accounts for adding the first observation's contribution. Taking expectations of the conditional design-summand calculation yields

$$\left\| \mathbb{E}Q_n(h_n) - \mathbb{E}\left\{ \frac{2}{n} \sum_{i=1}^{n} \Sigma_X(V_i) \right\} \right\| \leq C\epsilon + o(1)$$

for fixed ϵ. The second expectation equals $2\mathbb{E}\{\Sigma_X(V_i)\} = Q$. Hence

$$\limsup_{n\to\infty} \|\mathbb{E}Q_n(h_n) - Q\| \leq C\epsilon.$$

Letting $\epsilon \downarrow 0$ proves $\mathbb{E}Q_n(h_n) \to Q$.

The derivative-matrix expectation. We next show $\mathbb{E}G_n(h_n) \to G$. Define

$$w_n(u) = \mathbb{E}(W \mid U_{h_n} = u), \qquad t_n(u) = \mathbb{E}(XW^{\mathsf{T}} \mid U_{h_n} = u),$$

and use subscripts zero for the corresponding moments given $V = u$. These moments satisfy the uniform convergence and Lipschitz bounds from Step 1. Indeed, conditional on $(V, Z) = (v, z)$, the vector W equals $w(v, z) = ((v - \gamma_{0,-1}^{\mathsf{T}} z)/\gamma_{0,1}, z^{\mathsf{T}})^{\mathsf{T}}$. The conditional moments entering the numerator integrals are therefore $w(v, z)$ and $m_X(v, z)w(v, z)^{\mathsf{T}}$. Both are bounded and uniformly Lipschitz in v under Assumption 4(1)–(2), so the numerator and denominator calculation in Step 1 applies.

Write $i = \pi_{h_n}(j)$, $k = \pi_{h_n}(j-1)$, $u = U_{h_n,i}$, and $u' = U_{h_n,k}$. Define

$$D_{n,j} = (X_i - X_k)g_0'(V_i)(W_i - W_k)^{\mathsf{T}}, \qquad G_n(h_n) = \frac{1}{n} \sum_{j=2}^{n} D_{n,j}.$$

Let $\widetilde{D}_{n,j}$ replace $g_0'(V_i)$ in this expression by $g_0'(u)$. Boundedness of g_0'' under Assumption 2 gives

$$|g_0'(V_i) - g_0'(u)| \leq C|V_i - u| \leq C\|h_n\|.$$

For large n, all perturbed index values lie in the neighborhood on which Assumption 2 imposes these derivative bounds. Boundedness of X_i, W_i under Assumption 4(1)–(2) therefore implies $\max_j \|D_{n,j} - \widetilde{D}_{n,j}\| \leq C\|h_n\|$. Conditional independence from Step 2 yields

$$\mathbb{E}(\widetilde{D}_{n,j} \mid \mathcal{U}_n) = g_0'(u)\{t_n(u) + t_n(u') - m_n(u)w_n(u')^{\mathsf{T}} - m_n(u')w_n(u)^{\mathsf{T}}\}.$$

For $j \in J_{n,\epsilon}$, Steps 1 and 2 replace the moments in this formula by their true-index counterparts at the common value u. Put

$$d(u) = 2g_0'(u)\{t_0(u) - m_0(u)w_0(u)^{\mathsf{T}}\} = 2g_0'(u)\operatorname{Cov}(X, W \mid V = u)$$

and define

$$R_{n,\epsilon}^G = \max_{j \in J_{n,\epsilon}} \|\mathbb{E}(\widetilde{D}_{n,j} \mid \mathcal{U}_n) - d(U_{h_n,(j)})\|,$$

assigning zero to an empty maximum. Uniform moment convergence and the maximal-gap bound give $R_{n,\epsilon}^G \to_p 0$. The boundedness conditions above give $0 \leq R_{n,\epsilon}^G \leq C$. For each $\eta > 0$,

$$\mathbb{E}R_{n,\epsilon}^G \leq \eta + C\Pr(R_{n,\epsilon}^G > \eta),$$

so taking $n \to \infty$ and then $\eta \downarrow 0$ proves $\mathbb{E}R^G_{n,\epsilon} \to 0$.

Choose bounds $\|X_i\| \le M_X$, $\|W_i\| \le M_W$, and $|g_0'| \le M_g$ as supplied by Assumption 2 and Assumption 4(1)–(2). Then

$$\|D_{n,j}\| \le 4M_X M_W M_g.$$

At most $2B_{n,\epsilon}$ pairs have an index value outside I_ϵ. Their contribution to $G_n(h_n)$ has expected norm at most $8M_X M_W M_g \mathbb{E}B_{n,\epsilon}/n$. The function d is bounded and Lipschitz on $[v_L, v_H]$ by the same assumptions and the conditional-moment calculation above. For retained pairs, replacing $d(U_{h_n,i})$ by $d(V_i)$ therefore costs at most $C\|h_n\|$ per term. Completing the sum gives the exact identity

$$\begin{aligned}\frac{1}{n}\sum_{i=1}^n d(V_i) - \frac{1}{n}\sum_{j\in J_{n,\epsilon}} d(V_{\pi_{h_n}(j)}) \\ = \frac{1}{n}\sum_{j\in\{2,\dots,n\}\setminus J_{n,\epsilon}} d(V_{\pi_{h_n}(j)}) + \frac{1}{n}d(V_{\pi_{h_n}(1)}).\end{aligned}$$

Its expected norm is at most $C(\mathbb{E}B_{n,\epsilon}+1)/n$. Combining these bounds yields, for fixed ϵ,

$$\begin{aligned}\left\|\mathbb{E}G_n(h_n) - \mathbb{E}\left\{\frac{1}{n}\sum_i d(V_i)\right\}\right\| &\le C\|h_n\| + \mathbb{E}R^G_{n,\epsilon} + C\frac{\mathbb{E}B_{n,\epsilon}+1}{n} \\ &\le C\epsilon + o(1).\end{aligned}$$

Since $\mathbb{E}d(V_i) = G$, this proves $\limsup_n \|\mathbb{E}G_n(h_n) - G\| \le C\epsilon$, with C independent of ϵ. Letting $\epsilon \downarrow 0$ gives $\mathbb{E}G_n(h_n) \to G$.

The conditional-variance expectation. We now show $\mathbb{E}\mathcal{V}_n(h_n) \to \Omega$. Define

$$a_n(u) = \mathbb{E}(v_i \mid U_{h_n,i} = u), \qquad b_n(u) = \mathbb{E}(v_i X_i \mid U_{h_n,i} = u),$$
$$C_n(u) = \mathbb{E}(v_i X_i X_i^\mathsf{T} \mid U_{h_n,i} = u).$$

Use subscripts zero for the corresponding moments conditional on $V_i = u$. Step 1 controls these moments under Assumption 4(2). For $2 \le j \le n-1$, write $i = \pi_{h_n}(j)$, $i^- = \pi_{h_n}(j-1)$, and $i^+ = \pi_{h_n}(j+1)$. Denote their perturbed index values by u, u_-, u_+, respectively. Then

$$c_{n,i}(h_n) = 2X_i - X_{i^-} - X_{i^+}, \qquad Z_{n,j} = v_i c_{n,i}(h_n) c_{n,i}(h_n)^\mathsf{T}.$$

At $j = 1, n$, define $Z_{n,j}$ using the endpoint coefficient X_i minus the regressor vector of its sole neighbor. Thus $\mathcal{V}_n(h_n) = n^{-1}\sum_{j=1}^n Z_{n,j}$. Expanding the product and using conditional independence from Step 2 gives

$$\begin{aligned}\mathbb{E}(Z_{n,j} \mid \mathcal{U}_n) = {} & 4C_n(u) - 2b_n(u)\{m_n(u_-) + m_n(u_+)\}^\mathsf{T} \\ & - 2\{m_n(u_-) + m_n(u_+)\}b_n(u)^\mathsf{T} \\ & + a_n(u)\{s_n(u_-) + s_n(u_+) + m_n(u_-)m_n(u_+)^\mathsf{T} + m_n(u_+)m_n(u_-)^\mathsf{T}\}.\end{aligned}$$

This factorization allows dependence between v_i and X_i. Independence is used only between different observations.

Retain the positions in

$$K_{n,\epsilon} = \{j \in \{2, \ldots, n-1\} : \; U_{h_n,(j-1)}, U_{h_n,(j)}, U_{h_n,(j+1)} \in I_\epsilon\}.$$

For these positions, both neighbors' index values differ from u by at most the maximal adjacent gap. Steps 1 and 2 replace every moment in the expansion by its true-index counterpart at u, giving

$$\begin{aligned}\omega_0(u) = 4C_0(u) - 4b_0(u)m_0(u)^\mathsf{T} - 4m_0(u)b_0(u)^\mathsf{T} \\ + 2a_0(u)s_0(u) + 2a_0(u)m_0(u)m_0(u)^\mathsf{T}.\end{aligned}$$

To identify this expression, put $\xi = X - m_0(u)$ when conditioning on $V = u$. Direct expansion yields

$$\begin{aligned}\mathbb{E}(v\xi\xi^\mathsf{T} \mid V = u) = C_0(u) - b_0(u)m_0(u)^\mathsf{T} - m_0(u)b_0(u)^\mathsf{T} \\ + a_0(u)m_0(u)m_0(u)^\mathsf{T}.\end{aligned}$$

Using $\Sigma_X(u) = s_0(u) - m_0(u)m_0(u)^\mathsf{T}$, we obtain

$$\omega_0(u) = 4\mathbb{E}(v\xi\xi^\mathsf{T} \mid V = u) + 2a_0(u)\Sigma_X(u).$$

The definition $v_i = \mathbb{E}(\varepsilon_i^2 \mid B_i)$ and the fact that B_i contains X_i, W_i in Assumption 3 imply

$$\begin{aligned}a_0(u) = \mathbb{E}\{\mathbb{E}(\varepsilon_i^2 \mid B_i) \mid V_i = u\} = \sigma^2(u), \\ \mathbb{E}(v_i\xi_i\xi_i^\mathsf{T} \mid V_i = u) = \mathbb{E}\{\mathbb{E}(\varepsilon_i^2\xi_i\xi_i^\mathsf{T} \mid B_i) \mid V_i = u\} = M(u).\end{aligned}$$

The second equality uses that ξ_i and V_i are measurable functions of B_i. Thus $\omega_0(u) = 4M(u) + 2\sigma^2(u)\Sigma_X(u)$ and $\mathbb{E}\omega_0(V_i) = \Omega$ by (7).

Define

$$R_{n,\epsilon}^V = \max_{j \in K_{n,\epsilon}} \|\mathbb{E}(Z_{n,j} \mid \mathcal{U}_n) - \omega_0(U_{h_n,(j)})\|,$$

with zero assigned to an empty maximum. The expanded conditional expectation and Steps 1 and 2 imply $R_{n,\epsilon}^V \to_p 0$. Under Assumption 4(2), choose $v_i \le M_v$ and $\|X_i\| \le M_X$. Then

$$\|Z_{n,j}\| \le 16M_vM_X^2$$

for every position, including the endpoints. The function ω_0 is bounded as well, so $R_{n,\epsilon}^V \le C$ and

$$\mathbb{E}R_{n,\epsilon}^V \le \eta + C\Pr(R_{n,\epsilon}^V > \eta)$$

for each $\eta > 0$. Taking limits gives $\mathbb{E}R_{n,\epsilon}^V \to 0$.

Each observation whose perturbed index lies outside I_ϵ belongs to at most three consecutive triples. Including the two endpoint positions, at most $3B_{n,\epsilon} + 2$ positions lie outside $K_{n,\epsilon}$. Hence

$$\mathbb{E}\left\|\frac{1}{n}\sum_{j \in \{1,\ldots,n\} \setminus K_{n,\epsilon}} Z_{n,j}\right\| \le 16M_vM_X^2\frac{3\mathbb{E}B_{n,\epsilon} + 2}{n}.$$

The function ω_0 is bounded and Lipschitz under Assumption 1. For retained positions, replacing $\omega_0(U_{h_n,(j)})$ by $\omega_0(V_{\pi_{h_n}(j)})$ therefore costs at most $C\|h_n\|$ per term. Completing the sum over all observations adds

$$\frac{1}{n}\sum_{j\in\{1,\dots,n\}\setminus K_{n,\epsilon}} \omega_0(V_{\pi_{h_n}(j)}),$$

whose expected norm is at most $C(3\mathbb{E}B_{n,\epsilon}+2)/n$. Combining these bounds gives

$$\begin{aligned}\left\|\mathbb{E}\mathcal{V}_n(h_n) - \mathbb{E}\left\{\frac{1}{n}\sum_i \omega_0(V_i)\right\}\right\| &\le \mathbb{E}R^V_{n,\epsilon} + C\|h_n\| + C\frac{3\mathbb{E}B_{n,\epsilon}+2}{n}\\ &\le C\epsilon + o(1).\end{aligned}$$

Since $\mathbb{E}\omega_0(V_i)=\Omega$, for fixed ϵ this implies $\limsup_n \|\mathbb{E}\mathcal{V}_n(h_n)-\Omega\| \le C\epsilon$, with C independent of ϵ. Letting $\epsilon \downarrow 0$ proves $\mathbb{E}\mathcal{V}_n(h_n)\to\Omega$.

Uniform limits and evaluation at the estimated coefficient. We have proved all three expectation limits for every deterministic sequence $h_n\to 0$. Let $a_n(h)=\mathbb{E}T_n(h)$ and let a be the corresponding scalar entry of Q, G, or Ω. Fix any deterministic $r_n\downarrow 0$. If $\sup_{\|h\|\le r_n}|a_n(h)-a|$ did not tend to zero, there would be an $\eta>0$, a subsequence n_k, and deterministic points $\|h_{n_k}\|\le r_{n_k}$ such that

$$|a_{n_k}(h_{n_k})-a|>\eta.$$

Set $h_n=0$ outside this subsequence. This deterministic sequence tends to zero, contradicting the expectation limit already proved. Therefore,

$$\sup_{\|h\|\le r_n}|\mathbb{E}T_n(h)-a|\longrightarrow 0.$$

Together with (C1), this yields

$$\begin{aligned}\sup_{\|h\|\le r_n}|T_n(h)-a| &\le \sup_{\|h\|\le r_n}|T_n(h)-\mathbb{E}T_n(h)| + \sup_{\|h\|\le r_n}|\mathbb{E}T_n(h)-a|\\ &\xrightarrow{p} 0.\end{aligned}$$

There are finitely many matrix entries, so the three matrix limits hold uniformly over $\|h\|\le r_n$.

Finally, put $\widehat{h}_n=\widehat{\gamma}_{-1}-\gamma_{0,-1}$. Consistency guarantees a deterministic sequence $r_n\downarrow 0$ with $\Pr(\|\widehat{h}_n\|>r_n)\to 0$. To construct one, choose strictly increasing integers N_k such that $\Pr(\|\widehat{h}_n\|>1/k)\le 1/k$ for all $n\ge N_k$, and put $r_n=1/k$ for $N_k\le n<N_{k+1}$, with any positive initial values. For every fixed $\eta>0$,

$$\begin{aligned}\Pr\{|T_n(\widehat{h}_n)-a|>\eta\} &\le \Pr(\|\widehat{h}_n\|>r_n) + \Pr\left\{\sup_{\|h\|\le r_n}|T_n(h)-a|>\eta\right\}\\ &\longrightarrow 0.\end{aligned}$$

For a consistent normalized first-stage estimator, the ordering at $\widehat{h}_n$ is the fitted ordering $\widehat{\pi}$. Applying this probability bound to each matrix entry proves (19). □

D Proof of theorem 1

Throughout the proof, C denotes a generic finite positive constant whose value may change between occurrences but does not depend on n or the realized fitted ordering $\widehat{\pi}$.

Our goal is to establish the expansion and limiting distribution in (20). Under Assumptions 1, 2 and 4, lemma 1 and the consistency supplied by Assumption 3 give the three moment limits in (19).

Write $\Psi_n = n^{-1/2} \sum_i \psi(B_i)$ for the first-stage influence sum. We show that

$$\sqrt{n}(\widehat{\theta} - \theta_0) = Q^{-1}(\widehat{S}_n - G\Psi_n) + o_{\mathbb{P}}(1)$$

and that

$$(\widehat{S}_n, \Psi_n) \xrightarrow{d} \mathcal{N}\left(0, \begin{pmatrix} \Omega & 0 \\ 0 & \Omega_\psi \end{pmatrix}\right).$$

The key step is the conditional Gaussian limit of $\widehat{S}_n$ given $\mathcal{F}_n$. We obtain this limit from the conditional Lindeberg–Feller theorem in Bulinski (2017, Theorem 1), by verifying conditional centering, independence, variance convergence, and the Lindeberg condition. We then combine this conditional limit with the ordinary i.i.d. central limit theorem for Ψ_n to establish the joint limit above. Finally, we bound the nuisance remainder to obtain the expansion and apply Slutsky's theorem to conclude asymptotic normality with variance (21).

Conditional score limit. Conditional on $\mathcal{F}_n$, the errors are independent across observations, centered, and the coefficient vectors $c_{n,i}$ are fixed. For $p = 2 + \min(\eta_s/2, 1) > 2$, define the conditional $2p$-th absolute moment of the outcome error by

$$M_{\varepsilon,2p}(B_i) \equiv \mathbb{E}(|\varepsilon_i|^{2p} \mid B_i).$$

From the definition of $c_{n,i}$, an observation in position j satisfies

$$c_{n,\widehat{\pi}(j)} = \begin{cases} X_{\widehat{\pi}(1)} - X_{\widehat{\pi}(2)}, & j = 1, \\ 2X_{\widehat{\pi}(j)} - X_{\widehat{\pi}(j-1)} - X_{\widehat{\pi}(j+1)}, & 2 \le j \le n-1, \\ X_{\widehat{\pi}(n)} - X_{\widehat{\pi}(n-1)}, & j = n. \end{cases}$$

Consequently, $c_{n,i}$ depends only on observation i and its immediate neighbors in the fitted ordering $\widehat{\pi}$. Each $c_{n,i}$ is a sum of at most three regressor vectors, each multiplied by 1, 2, or -1. These numerical multipliers have absolute value at most two. If N_i denotes the set containing observation i and these neighbors, then $\|c_{n,i}\|^p \le C \sum_{k \in N_i} \|X_k\|^p$. Independence across observations under Assumption 1 and conditional Cauchy–Schwarz give

$$\mathbb{E}(|\varepsilon_i|^p \mid \mathcal{F}_n) = \mathbb{E}(|\varepsilon_i|^p \mid B_i) \le \{\mathbb{E}(|\varepsilon_i|^{2p} \mid B_i)\}^{1/2} = M_{\varepsilon,2p}(B_i)^{1/2}.$$

Since $c_{n,i}$ is fixed conditional on $\mathcal{F}_n$, it follows that

$$\mathbb{E}(\|c_{n,i}\varepsilon_i\|^p \mid \mathcal{F}_n) \le C \sum_{k \in N_i} \|X_k\|^p M_{\varepsilon,2p}(B_i)^{1/2}.$$

Each observation belongs to at most three sets N_i, namely its own set and those of its immediate neighbors in the fitted ordering $\widehat{\pi}$. This bound does not depend on n. Cauchy–Schwarz over the incidences (i, k) therefore yields

$$\frac{1}{n}\sum_i \mathbb{E}(\|c_{n,i}\varepsilon_i\|^p \mid \mathcal{F}_n) \le C\left(\frac{1}{n}\sum_i \|X_i\|^{2p}\right)^{1/2}\left(\frac{1}{n}\sum_i M_{\varepsilon,2p}(B_i)\right)^{1/2} = O_{\mathbb{P}}(1).$$

Here $2p \le 4 + \eta_s$, and the law of large numbers applies to both averages. Dividing each score summand by $\sqrt{n}$, we obtain

$$\begin{aligned}
&\sum_{i=1}^n \mathbb{E}\left(\left\|\frac{c_{n,i}\varepsilon_i}{\sqrt{n}}\right\|^p \,\middle|\, \mathcal{F}_n\right) \\
&\quad = n^{-p/2}\sum_{i=1}^n \mathbb{E}(\|c_{n,i}\varepsilon_i\|^p \mid \mathcal{F}_n) \\
&\quad = n^{1-p/2}\underbrace{\frac{1}{n}\sum_{i=1}^n \mathbb{E}(\|c_{n,i}\varepsilon_i\|^p \mid \mathcal{F}_n)}_{O_{\mathbb{P}}(1)} = o_{\mathbb{P}}(1),
\end{aligned}$$

where the final $o_{\mathbb{P}}(1)$ follows because $p > 2$ implies $n^{1-p/2} \to 0$. Thus the sum of the conditional p-th moments of the normalized score summands tends to zero in probability. Together with the variance convergence $\mathcal{V}_n \to_p \Omega$ established by lemma 1, this proves, for each fixed $t \in \mathbb{R}^d$,

$$\mathbb{E}\{\exp(it^{\mathsf{T}}\widehat{S}_n) \mid \mathcal{F}_n\} \xrightarrow{p} \exp(-t^{\mathsf{T}}\Omega t/2),$$

by an application of the conditional Lindeberg–Feller theorem in Bulinski (2017, Theorem 1) when $t^{\mathsf{T}}\Omega t > 0$. For each fixed t, we apply this scalar conditional central limit theorem to

$$t^{\mathsf{T}}\widehat{S}_n = \frac{1}{\sqrt{n}}\sum_{i=1}^n t^{\mathsf{T}}c_{n,i}\varepsilon_i.$$

Considering these scalar linear combinations for every fixed t is the reduction used in the Cramér–Wold device. To verify the conditional Lindeberg condition explicitly, put $Z_{n,i} = n^{-1/2}t^{\mathsf{T}}c_{n,i}\varepsilon_i$. For each fixed $\epsilon > 0$,

$$\sum_i \mathbb{E}\{Z_{n,i}^2 1\{|Z_{n,i}| > \epsilon\} \mid \mathcal{F}_n\} \le \epsilon^{2-p}\sum_i \mathbb{E}(|Z_{n,i}|^p \mid \mathcal{F}_n) = o_{\mathbb{P}}(1).$$

The last equality follows from the preceding Lyapunov bound and $|t^{\mathsf{T}}c_{n,i}\varepsilon_i| \le \|t\|\|c_{n,i}\varepsilon_i\|$. Moreover, the conditional variance is $s_{n,t}^2 = t^{\mathsf{T}}\mathcal{V}_n t \to_p t^{\mathsf{T}}\Omega t$. For a positive limiting variance it is bounded away from zero with probability tending to one, so the same bound verifies Lindeberg after normalization by $s_{n,t}$. When $t^{\mathsf{T}}\Omega t = 0$, conditional Chebyshev gives the stated conditional characteristic-function limit.

Joint limit with the first-stage sum. To establish joint convergence of $\widehat{S}_n$ and Ψ_n, we use L^1 convergence of the conditional characteristic functions. Their absolute values are at most one, so the convergence in probability established in the preceding step implies

$$\mathbb{E}\left|\mathbb{E}\{\exp(it^{\mathsf{T}}\widehat{S}_n) \mid \mathcal{F}_n\} - \exp(-t^{\mathsf{T}}\Omega t/2)\right| \longrightarrow 0.$$

Since Ψ_n is $\mathcal{F}_n$-measurable, for each fixed b, multiply the conditional characteristic function of $\widehat{S}_n$ by $\exp(ib^{\mathsf{T}}\Psi_n)$ and take expectations. The L^1 convergence and the ordinary i.i.d. CLT give

$$(\widehat{S}_n, \Psi_n) \xrightarrow{d} \mathcal{N}\left(0, \begin{pmatrix} \Omega & 0 \\ 0 & \Omega_\psi \end{pmatrix}\right).$$

Thus, the fitted-order score $\widehat{S}_n$ and the first-stage sum Ψ_n have a joint Gaussian limit, with independent components.

Removal of the smooth regressor mean. In this step, we show that subtracting $m_0(V_i)$ from X_i changes the score by $o_{\mathbb{P}}(1)$, establishing (17). Let f be the matrix whose i-th row is $m_0(V_i)^{\mathsf{T}}$, and put

$$a_{n,i} = (H_{\widehat{\pi}}^{\mathsf{T}} H_{\widehat{\pi}} f)_{i,\cdot}^{\mathsf{T}}.$$

Define

$$d_j = m_0(V_{\widehat{\pi}(j)}) - m_0(V_{\widehat{\pi}(j-1)}), \qquad j = 2, \ldots, n,$$

and put $d_1 = d_{n+1} = 0$. The explicit formula for the first-difference matrix gives

$$a_{n,\widehat{\pi}(j)} = d_j - d_{j+1}, \qquad j = 1, \ldots, n.$$

Using $\|a - b\|^2 \le 2\|a\|^2 + 2\|b\|^2$, we obtain

$$\begin{aligned} \sum_{i=1}^{n} \|a_{n,i}\|^2 &= \sum_{j=1}^{n} \|d_j - d_{j+1}\|^2 \\ &\le 2 \sum_{j=1}^{n} \left(\|d_j\|^2 + \|d_{j+1}\|^2\right) \\ &= 4 \sum_{j=2}^{n} \|d_j\|^2. \end{aligned}$$

The bounded derivative of m_0 under Assumption 2 and the spacing bound in proposition 2 imply

$$\sum_{j=2}^{n} \|d_j\|^2 \le C \sum_{j=2}^{n} |V_{\widehat{\pi}(j)} - V_{\widehat{\pi}(j-1)}|^2 = O_{\mathbb{P}}(1).$$

Hence $\sum_{i=1}^{n} \|a_{n,i}\|^2 = O_{\mathbb{P}}(1)$. Since $v_i = \mathbb{E}(\varepsilon_i^2 \mid B_i)$ is integrable, $\max_i v_i / n = o_{\mathbb{P}}(1)$. The omitted score has conditional mean zero and conditional variance bounded in trace by

$$\frac{\max_i v_i}{n} \sum_i \|a_{n,i}\|^2 = o_{\mathbb{P}}(1).$$

Conditional Chebyshev's inequality proves (17).

Nuisance remainder. Put $h_n = \widehat{\gamma} - \gamma_0$. Taylor's theorem and $\Delta V_j = \Delta \widehat{V}_j - \Delta W_j^{\mathsf{T}} h_n$ give

$$\Delta_{\widehat{\pi}} g_0(V)_j = g_0'(V_{\widehat{\pi}(j)}) \Delta_{\widehat{\pi}} \widehat{V}_j - g_0'(V_{\widehat{\pi}(j)}) \Delta_{\widehat{\pi}} W_j^{\mathsf{T}} h_n + r_{n,j}, \quad |r_{n,j}| \le C |\Delta_{\widehat{\pi}} V_j|^2.$$

Thus $R_{g,n} = A_{1n} - G_n \sqrt{n} h_n + A_{3n}$, with

$$\|A_{1n}\| \le C \left(\frac{1}{n} \sum_j \|\Delta_{\widehat{\pi}} X_j\|^2 \right)^{1/2} \left(\sum_j |\Delta_{\widehat{\pi}} \widehat{V}_j|^2 \right)^{1/2} = o_{\mathbb{P}}(1),$$

$$\|A_{3n}\| \le C \left(\frac{1}{n} \sum_j \|\Delta_{\widehat{\pi}} X_j\|^2 \right)^{1/2} \max_j |\Delta_{\widehat{\pi}} V_j| \left(\sum_j |\Delta_{\widehat{\pi}} V_j|^2 \right)^{1/2} = o_{\mathbb{P}}(1).$$

The permutation-invariant moment bound and proposition 2 justify both conclusions. The limit $G_n \to_p G$ established by lemma 1, combined with (8), gives $R_{g,n} = -G\Psi_n + o_{\mathbb{P}}(1)$. Finally, lemma 1 gives $\widehat{Q}_n \to_p Q$, and Q is positive definite under Assumption 2. The exact decomposition (15) and the joint CLT above therefore prove the theorem.

E Proof of corollary 2 (fixed-order weighted differences)

Proof of corollary 2. We first extend the uniform laws of large numbers in lemma 1 to the fixed weights d. For deterministic h, let $T_{n,d}(h)$ be any scalar entry of the weighted design matrix, derivative matrix, or conditional score variance, with observations ordered by $U_{h,i} = V_i + h^{\mathsf{T}} Z_i$. Each weighted difference involves $K+1$ observations, and each vector $c_{n,i,d}$ involves observations within K positions of i in that ordering. Since K and d are fixed, Assumptions 2 and 4 bounds every scalar summand by a constant depending only on K, d and the population bounds. Replacing one observation changes only a bounded number of these summands, so $T_{n,d}(h)$ has bounded differences $C_{K,d}/n$. McDiarmid's inequality therefore gives the same exponential concentration bound as in Appendix C, with a constant that may depend on K and d.

The deterministic grid and the count of observations with nearly equal index values used in that appendix depend only on the scalar indices. They can therefore be used without modification. A change in the ordering of these observations affects only summands involving observations within K positions of them. Thus the bound on the difference between a statistic at h and at its nearest grid point is multiplied by a constant depending only on K, d. The concentration and grid arguments give uniform convergence of each $T_{n,d}(h)$ to its expectation over the same shrinking neighborhoods.

To identify the expectations, fix a deterministic sequence $h_n \to 0$ and an interior interval $I_\epsilon = [v_L + \epsilon, v_H - \epsilon]$. Conditional on the perturbed index values, observations are independent. The conditional-moment convergence and spacing arguments in Appendix C imply that all observations in a fixed block can be evaluated at a common index value, with an error tending to zero when the block lies in I_ϵ. Indeed, a gap across at most $2K$ positions is bounded by $2K$ times the largest consecutive gap. At a common index value v, the zero-sum weights give the conditional expectations

$$c_0(d) \Sigma_X(v), \qquad c_0(d) g_0'(v) \operatorname{Cov}(X, W \mid V = v)$$

for the design and derivative summands. For the variance summand, the conditional expectation is

$$c_0(d)^2 M(v) + 2\sum_{h=1}^{K} c_h(d)^2 \sigma^2(v)\Sigma_X(v),$$

as shown by the conditional-draw calculation below. Each observation outside I_ϵ enters only a bounded number of summands. The expected contribution of these summands is therefore at most $C_{K,d}(\epsilon + \|h_n\|)$. The finitely many endpoint summands contribute at most $C_{K,d}/n$. Integrating the conditional expectations, letting $n \to \infty$, and then letting $\epsilon \downarrow 0$ gives the limits Q_d, G_d, Ω_d. The sequential argument in Appendix C then makes these expectation limits uniform over shrinking neighborhoods. Combined with the uniform concentration bound, this proves the three weighted laws of large numbers. Consistency under Assumption 3 permits evaluation at the fitted coefficient $\widehat{\gamma}$.

We next control the smooth nuisance remainders and establish the conditional limit. The condition $\sum_{\ell=0}^{K} d_\ell = 0$ gives, for any function a,

$$\Delta_{\pi_\gamma,d} a(V)_j = \sum_{\ell=0}^{K} d_\ell \{a(V_{\pi_\gamma(j-\ell)}) - a(V_{\pi_\gamma(j)})\}.$$

Because K is fixed, Lipschitz continuity and repeated applications of Cauchy–Schwarz give a finite constant $C_{d,K}$ such that

$$\frac{1}{n}\sum_{j=K+1}^{n} \|\Delta_{\pi_\gamma,d} a(V)_j\|^2 \leq C_{d,K}\frac{1}{n}\sum_{j=2}^{n} |V_{\pi_\gamma(j)} - V_{\pi_\gamma(j-1)}|^2.$$

Indeed, for each $1 \leq k \leq K$,

$$|V_{\pi_\gamma(j)} - V_{\pi_\gamma(j-k)}|^2 \leq k\sum_{r=0}^{k-1} |V_{\pi_\gamma(j-r)} - V_{\pi_\gamma(j-r-1)}|^2.$$

Every ordinary spacing occurs in at most K of the resulting sums. Hence proposition 2 implies

$$\sup_{\gamma\in\Gamma_n(\delta)} \frac{1}{n}\sum_{j=K+1}^{n} \|\Delta_{\pi_\gamma,d} a(V)_j\|^2 = O_{\mathbb{P}}(n^{-1}). \tag{E1}$$

The bound applies to m_0, g_0, and the conditional moment functions used in the proof of theorem 1.

With the weighted moment limits established, conditional on $\mathcal{F}_n$, the weighted score equals $n^{-1/2}\sum_i c_{n,i,d}\varepsilon_i$. Each coefficient is a bounded linear combination of regressors within K positions on either side. The incidence bound in Appendix D therefore applies with constants depending only on K, d. It gives the same conditional Lyapunov bound and joint asymptotic independence from Ψ_n.

For the nuisance expansion, expand each $g_0(V_{\widehat{\pi}(j-\ell)})$ around $V_{\widehat{\pi}(j)}$. The zero-sum weights remove the constant term. The linear term is $g_0'(V_{\widehat{\pi}(j)})\Delta_{\widehat{\pi},d}V_j$, and the remainder is bounded by $C\sum_{\ell=0}^{K}|V_{\widehat{\pi}(j-\ell)}-V_{\widehat{\pi}(j)}|^2$. Denote this remainder by $\rho_{n,j}$ and put

$$M_n=\max_{2\le j\le n}|\Delta_{\widehat{\pi}}V_j|,\qquad J_n=\sum_{j=2}^{n}|\Delta_{\widehat{\pi}}V_j|^2.$$

Since K is fixed, the preceding block bounds give

$$\max_j|\rho_{n,j}|\le C_{K,d}M_n^2,\qquad\sum_j|\rho_{n,j}|\le C_{K,d}J_n.$$

It follows from proposition 2 that

$$\sum_j\rho_{n,j}^2\le\max_j|\rho_{n,j}|\sum_j|\rho_{n,j}|\le C_{K,d}M_n^2J_n=o_{\mathbb{P}}(1).$$

Applying the same fixed-block inequality to the fitted indices gives

$$\sum_j|\Delta_{\widehat{\pi},d}\widehat{V}_j|^2\le C_{K,d}\sum_{j=2}^{n}|\Delta_{\widehat{\pi}}\widehat{V}_j|^2=o_{\mathbb{P}}(1)$$

by (11). Finally, the incidence of each observation in at most $K+1$ differences gives

$$\frac{1}{n}\sum_j\|\Delta_{\widehat{\pi},d}X_j\|^2\le\frac{C_{K,d}}{n}\sum_i\|X_i\|^2=O_{\mathbb{P}}(1).$$

Cauchy–Schwarz therefore makes the score contributions from both $\rho_{n,j}$ and $g_0'(V_{\widehat{\pi}(j)})\Delta_{\widehat{\pi},d}\widehat{V}_j$ $o_{\mathbb{P}}(1)$. Together with $G_{n,d}\to_p G_d$ and (8), this yields

$$n^{-1/2}\sum_j\Delta_{\widehat{\pi},d}X_j\Delta_{\widehat{\pi},d}g_0(V)_j=-G_d\Psi_n+o_{\mathbb{P}}(1).$$

Combining the score limit, this expansion, and the established design limit proves (23).

To obtain (24), fix v and use the independent conditional draws defining $\mathcal{B}_d(v)$. Grouping the terms multiplying each error in the sum of weighted scores gives the interior vector

$$a_{d,0}(v)=c_0(d)\xi_0(v)+\sum_{h=1}^{K}c_h(d)\{\xi_{-h}(v)+\xi_h(v)\}.$$

The zero-sum restriction on d gives $c_0(d)+2\sum_{h=1}^{K}c_h(d)=0$, so subtracting the common conditional regressor mean does not change this vector. Conditional centering under Assumption 3 and independence across the draws imply that different errors contribute zero cross-covariance after conditioning on all regressors. The finitely many endpoint terms have vanishing contribution to the variance per observation. Consequently,

$$\mathcal{B}_d(v)=\mathbb{E}\{\varepsilon_0(v)^2a_{d,0}(v)a_{d,0}(v)^{\mathsf{T}}\}.$$

Because $\mathbb{E}\{\xi_h(v)\} = 0$ and the neighboring draws are independent of the central draw, expansion gives

$$\mathcal{B}_d(v) = c_0(d)^2 M(v) + 2\sum_{h=1}^{K} c_h(d)^2 \sigma^2(v)\Sigma_X(v).$$

Taking expectations over V_i proves (24).

It remains to specialize the oracle variance under homoskedasticity. Two weighted score summands separated by h positions share observations only when $|h| \le K$. Direct expansion using $\mathbb{E}(\varepsilon_i^2 \mid X_i, V_i) = \sigma^2$ gives

$$\mathrm{Cov}\{b_{d,0}(v), b_{d,h}(v) \mid V_i = v\} = c_h(d)^2 \sigma^2 \Sigma_X(v), \qquad h = 0, \ldots, K.$$

Adding the lag-zero term and twice the positive-lag terms yields

$$\Omega_d = \left\{ c_0(d)^2 + 2\sum_{h=1}^{K} c_h(d)^2 \right\} \sigma^2 \mathbb{E}\{\Sigma_X(V_i)\}.$$

Premultiplication and postmultiplication by $Q_d^{-1} = c_0(d)^{-1}[\mathbb{E}\{\Sigma_X(V_i)\}]^{-1}$ proves (25). The optimal-sequence calculation in Hall, Kay, and Titterington (1990) gives the factor $1 + 1/(2K)$. □

F Verification for the sample selection application

We verify Assumptions 3 and 4 under the sufficient conditions stated in section 4, maintaining Assumptions 1 and 2 under the selected-population law P_{sel}. The verification has two steps.

Step 1. Verification of Assumption 3. Suppose $\widehat{\gamma}$ is computed only from (D_i, W_i) and admits the full-sample expansion

$$\sqrt{n}(\widehat{\gamma} - \gamma_0) = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \psi(D_i, W_i) + o_{\mathbb{P}}(1), \qquad \mathbb{E}\psi(D_i, W_i) = 0, \quad \mathbb{E}\|\psi(D_i, W_i)\|^2 < \infty.$$

This expansion is a requirement on the chosen first-stage estimator. It is satisfied by correctly specified probit and logit maximum likelihood estimators under standard regularity conditions, and by semiparametric estimators such as those in Powell, Stock, and Stoker (1989), Klein and Spady (1993), and Ichimura (1993) under their respective model and regularity conditions, after imposing the index normalization used here. The maximum rank correlation estimator in Han (1987) is a tuning-free example. The proof of Theorem 2 in Sherman (1993), combined with the empirical-average representation in the proof of Theorem 4 in that paper, yields the required asymptotically linear expansion under the assumptions of that paper. Another tuning-free example is the first-stage isotonic score estimator used in Liu and Yu (2022), based on Groeneboom and Hendrickx (2018). Its first-stage asymptotic result is stated in Lemma 4.1 in Liu and Yu (2022), and its linear representation is used in the proof of Theorem 4.2. Both estimators use only (D_i, W_i) when applied to the selection equation. Take $\mathcal{F}_n^S = \sigma\{(X_i, W_i, D_i) : 1 \le i \le n\}$. Both $\widehat{\gamma}$ and its influence-function sum are measurable with respect to $\mathcal{F}_n^S$. Moreover, (29) gives

$$\mathbb{E}(\varepsilon_i \mid X_i, W_i, D_i = 1) = \mathbb{E}(e_i \mid X_i, W_i, D_i = 1) - \lambda_0(V_i) = 0.$$

Independence across observations under Assumption 1 implies that, conditional on $\mathcal{F}_n^S$, the errors $\{\varepsilon_i : D_i = 1\}$ are mutually independent. Their conditional distributions may differ across observations. Thus the first-stage expansion and (29) supply the requirements of Assumption 3 for the selected-sample argument.

Step 2. Verification of Assumption 4 and the resulting moment limits. We apply lemma 1 under the law conditional on $D_i = 1$. The following sufficient conditions let us check the density and conditional-moment restrictions in Assumption 4 explicitly. Suppose $(e_i, \eta_i) \perp (X_i, W_i)$ and write

$$s(v) = \Pr(\eta_i > -v), \qquad r_S(v) = \operatorname{Var}(e_i \mid \eta_i > -v).$$

Suppose the full-population density $p(v, z)$ of (V, Z) satisfies Assumption 4(1). Suppose also that s is Lipschitz and bounded away from zero on $[v_L, v_H]$, and that r_S is bounded and Lipschitz on this interval. Finally, suppose X_i is bounded and the conditional moments of X_i and $X_i X_i^\mathsf{T}$ given (V_i, Z_i) satisfy Assumption 4(2).

First, independence of the disturbances and covariates gives $\Pr(D_i = 1 \mid V_i, Z_i) = s(V_i)$. Hence the selected-population density is

$$p_S(v, z) = \frac{s(v)p(v, z)}{p_1}, \qquad f_{V|D=1}(v) = \frac{s(v)f_V(v)}{p_1}.$$

The imposed bounds and Lipschitz conditions on s and p imply that p_S is bounded and uniformly Lipschitz in v, uniformly in z. The lower bounds on s and f_V imply that $f_{V|D=1}$ is bounded away from zero on $[v_L, v_H]$. Selection preserves the compact support of Z. This verifies Assumption 4(1) under the selected-population law.

Next, define

$$m_X(v, z) = \mathbb{E}(X_i \mid V_i = v, Z_i = z), \qquad S_X(v, z) = \mathbb{E}(X_i X_i^\mathsf{T} \mid V_i = v, Z_i = z),$$

and put $v_i^S = \mathbb{E}(\varepsilon_i^2 \mid X_i, W_i, D_i = 1)$. Under disturbance–covariate independence, selection leaves the conditional law of X_i given (V_i, Z_i) unchanged and $v_i^S = r_S(V_i)$. The five conditional moments required by Assumption 4(2) are therefore

$$\begin{aligned}
\mathbb{E}(X_i \mid V_i = v, Z_i = z, D_i = 1) &= m_X(v, z),\\
\mathbb{E}(X_i X_i^\mathsf{T} \mid V_i = v, Z_i = z, D_i = 1) &= S_X(v, z),\\
\mathbb{E}(v_i^S \mid V_i = v, Z_i = z, D_i = 1) &= r_S(v),\\
\mathbb{E}(v_i^S X_i \mid V_i = v, Z_i = z, D_i = 1) &= r_S(v)m_X(v, z),\\
\mathbb{E}(v_i^S X_i X_i^\mathsf{T} \mid V_i = v, Z_i = z, D_i = 1) &= r_S(v)S_X(v, z).
\end{aligned}$$

Each expression is bounded and uniformly Lipschitz in v, uniformly in z, by the conditions imposed above. Also, X_i and v_i^S are bounded. This verifies Assumption 4(2) under the selected-population law.

It remains to express the resulting limits on the full-sample scale. Let Q_{sel}, G_{sel}, and Ω_{sel} denote the limits in lemma 1 evaluated under the selected-population law, with $g_0 = \lambda_0$. Define

$$\begin{aligned}
Q^S &= p_1 Q_{\text{sel}} = 2\mathbb{E}[D_i \operatorname{Var}(X_i \mid V_i, D_i = 1)],\\
G^S &= p_1 G_{\text{sel}} = 2\mathbb{E}[D_i \lambda_0'(V_i) \operatorname{Cov}(X_i, W_i \mid V_i, D_i = 1)],\\
\Omega^S &= p_1 \Omega_{\text{sel}}.
\end{aligned}$$

Write $\widehat{Q}_n^S$, G_n^S, and $\mathcal{V}_n^S$ for the fitted-order design matrix, derivative matrix, and conditional score variance formed from selected observations with denominator n. In particular,

$$\mathcal{V}_n^S = \frac{1}{n} \sum_{i:D_i=1} v_i^S c_{n,i}^S c_{n,i}^{S\mathsf{T}},$$

where $c_{n,i}^S$ is the coefficient of ε_i in the selected ordered score before division by $\sqrt{n}$, formed using the first-difference matrix for the selected observations.

Conditional on the selection indicators, the selected observations are independent draws from the selected-population law. Lemma 1 shows that, as $n_1 \to \infty$, the design matrix, derivative matrix, and conditional score variance computed from the selected observations converge to their population limits uniformly over coefficients in a shrinking neighborhood of γ_0. This uniform convergence allows us to evaluate the three matrices at the consistent estimator $\widehat{\gamma}$ from Step 1, even though it is estimated using the full sample. Consequently,

$$\frac{n}{n_1}\widehat{Q}_n^S \to_p Q_{\text{sel}}, \qquad \frac{n}{n_1}G_n^S \to_p G_{\text{sel}}, \qquad \frac{n}{n_1}\mathcal{V}_n^S \to_p \Omega_{\text{sel}}.$$

Since $n_1/n \to_p p_1$ by Assumption 1, multiplication gives

$$\widehat{Q}_n^S \to_p Q^S, \qquad G_n^S \to_p G^S, \qquad \mathcal{V}_n^S \to_p \Omega^S.$$

These are the selected-sample counterparts of the three laws of large numbers in lemma 1, expressed on the full-sample scale.

For this application, the conditional argument in Appendix D uses $\mathcal{F}_n^S$. Given $\mathcal{F}_n^S$, the selected set and its fitted order are fixed, and the selected errors are independent and centered by Step 1. The spacing and conditional Lyapunov bounds apply to the selected sample under the maintained Assumptions 1 and 2 and the first-stage expansion. Since $n_1/n \to_p p_1 > 0$, these bounds also hold on the full-sample $\sqrt{n}$ scale. Step 2 supplies the conditional variance limit. The joint-limit argument then applies to the selected score and the $\mathcal{F}_n^S$-measurable full-sample first-stage sum.

This establishes the limiting distribution in (31).

G Verification for the treatment-effects application

We verify Assumptions 3 and 4 under the sufficient conditions stated in section 4.2 and below. We maintain (32), (33), (34), and (36), together with Assumptions 1 and 2 for the treatment-effects model. Here D_i replaces the generic regressor, X_i supplies the index covariates, and α_0 replaces the generic index coefficient. The verification has two steps.

Sufficient conditions for Assumption 4 are as follows. Normalize a nonzero coefficient of α_0, put $V_i = X_i^{\mathsf{T}}\alpha_0$, and let Z_i contain the remaining pretreatment covariates, using the normalization preceding Assumption 4. Suppose (V_i, Z_i) satisfies Assumption 4(1). Define

$$r_T(v, z) = \text{Var}\{Y_i(0) \mid V_i = v, Z_i = z\}.$$

Suppose r_T is bounded and uniformly Lipschitz in v, uniformly in z, and F is Lipschitz on $[v_L, v_H]$.

Step 1. Verification of Assumption 3. Under homogeneous treatment effects in (33),

$$Y_i = \beta_0 D_i + g_0(V_i) + \varepsilon_i, \qquad V_i = X_i^{\mathsf{T}}\alpha_0, \quad \varepsilon_i = Y_i(0) - g_0(V_i).$$

Take $B_i = (D_i, X_i)$ and $\mathcal{F}_n^T = \sigma\{(D_i, X_i) : 1 \le i \le n\}$. Unconfoundedness in (32) and the outcome-mean restriction (36) give

$$\mathbb{E}(\varepsilon_i \mid D_i, X_i) = \mathbb{E}\{Y_i(0) \mid D_i, X_i\} - g_0(V_i) = \mathbb{E}\{Y_i(0) \mid X_i\} - g_0(V_i) = 0.$$

Independence across observations under Assumption 1 implies that, conditional on $\mathcal{F}_n^T$, the errors $\varepsilon_1, \ldots, \varepsilon_n$ are mutually independent. Their conditional distributions may differ across observations.

Suppose the first-stage estimator uses only (D_i, X_i) and satisfies

$$\sqrt{n}(\widehat{\alpha} - \alpha_0) = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \psi_T(D_i, X_i) + o_{\mathbb{P}}(1), \qquad \mathbb{E}\psi_T(D_i, X_i) = 0, \quad \mathbb{E}\|\psi_T(D_i, X_i)\|^2 < \infty.$$

Both $\widehat{\alpha}$ and its influence-function sum are then $\mathcal{F}_n^T$-measurable. Together with the conditional-mean identity above, this verifies Assumption 3. The expansion remains a requirement on the chosen propensity-index estimator and does not follow from unconfoundedness alone.

Step 2. Verification of Assumption 4 and the resulting moment limits. We apply lemma 1 with D_i as the regressor. The density and support requirements in Assumption 4(1) are imposed directly on (V_i, Z_i) in the application. It remains to verify the boundedness and conditional-moment requirements in Assumption 4(2).

Because a nonzero coefficient of α_0 is fixed, (V_i, Z_i) determines X_i. Unconfoundedness and (36) therefore imply

$$\begin{aligned} v_i^T &\equiv \mathbb{E}(\varepsilon_i^2 \mid D_i, X_i) \\ &= \mathrm{Var}\{Y_i(0) \mid D_i, X_i\} \\ &= \mathrm{Var}\{Y_i(0) \mid X_i\} = r_T(V_i, Z_i). \end{aligned}$$

Thus v_i^T is bounded by the imposed bound on r_T, and D_i is bounded because it is binary. Moreover, (34) gives

$$\mathbb{E}(D_i \mid V_i = v, Z_i = z) = F(v).$$

Using $D_i^2 = D_i$, the five conditional moments in Assumption 4(2) are

$$\begin{aligned} \mathbb{E}(D_i \mid V_i = v, Z_i = z) &= F(v), \\ \mathbb{E}(D_i^2 \mid V_i = v, Z_i = z) &= F(v), \\ \mathbb{E}(v_i^T \mid V_i = v, Z_i = z) &= r_T(v, z), \\ \mathbb{E}(v_i^T D_i \mid V_i = v, Z_i = z) &= r_T(v, z)F(v), \\ \mathbb{E}(v_i^T D_i^2 \mid V_i = v, Z_i = z) &= r_T(v, z)F(v). \end{aligned}$$

Each expression is bounded and uniformly Lipschitz in v, uniformly in z, by the imposed conditions on F and r_T. This verifies Assumption 4(2).

Consequently, lemma 1 shows that the design matrix, derivative matrix, and conditional score variance converge to their population limits uniformly over coefficients in a shrinking neighborhood of α_0. This uniform convergence permits evaluation at the consistent $\widehat{\alpha}$ from Step 1, even though it uses the same treatment indicators and covariates as the second stage. This establishes the fitted-order moment limits in lemma 1 for this application.

To identify the limits, (34) and iterated expectation give

$$\mathbb{E}(D_i X_i^{\mathsf{T}} \mid V_i) = \mathbb{E}\{X_i^{\mathsf{T}} \mathbb{E}(D_i \mid X_i) \mid V_i\} = F(V_i)\mathbb{E}(X_i^{\mathsf{T}} \mid V_i).$$

Also $\mathbb{E}(D_i \mid V_i) = F(V_i)$, so $\operatorname{Cov}(D_i, X_i \mid V_i) = 0$ and

$$G = 2\mathbb{E}[g_0'(V_i) \operatorname{Cov}(D_i, X_i \mid V_i)] = 0.$$

Since D_i is binary,

$$Q = 2\mathbb{E}\{\operatorname{Var}(D_i \mid V_i)\} = 2\mathbb{E}[F(V_i)\{1 - F(V_i)\}] > 0$$

by overlap in (32). Applying theorem 1 with $G = 0$ establishes (38).

H Consistency calculation for Remark 1

Write $e_i \equiv e(X_i) = F(V_i)$ and $D_i = e_i + \xi_i$. Define the denominator and numerator of the ordered-difference slope as

$$A_n \equiv \frac{1}{n}\sum_{j=2}^{n}(\Delta_{\widehat{\pi}} D_j)^2,$$

$$B_n \equiv \frac{1}{n}\sum_{j=2}^{n}\Delta_{\widehat{\pi}} D_j \Delta_{\widehat{\pi}} Y_j.$$

Step 1. Remove the smooth conditional-mean component.

For the denominator,

$$A_n = \frac{1}{n}\sum_{j=2}^{n}(\Delta_{\widehat{\pi}}\xi_j)^2 + \frac{2}{n}\sum_{j=2}^{n}\Delta_{\widehat{\pi}}\xi_j \Delta_{\widehat{\pi}} e_j + \frac{1}{n}\sum_{j=2}^{n}(\Delta_{\widehat{\pi}} e_j)^2.$$

Since F is Lipschitz, proposition 2 gives

$$\frac{1}{n}\sum_{j=2}^{n}(\Delta_{\widehat{\pi}} e_j)^2 = O_{\mathbb{P}}(n^{-1}).$$

Moreover, $|\xi_i| \le 1$, so

$$\frac{1}{n}\sum_{j=2}^{n}(\Delta_{\widehat{\pi}}\xi_j)^2 = O_{\mathbb{P}}(1).$$

Cauchy–Schwarz therefore gives

$$\left|\frac{1}{n}\sum_{j=2}^{n}\Delta_{\widehat{\pi}}\xi_j \Delta_{\widehat{\pi}} e_j\right| = O_{\mathbb{P}}(n^{-1/2}).$$

Consequently,

$$A_n = \frac{1}{n}\sum_{j=2}^{n}(\Delta_{\widehat{\pi}}\xi_j)^2 + o_{\mathbb{P}}(1). \tag{H1}$$

For the numerator,

$$B_n = \frac{1}{n}\sum_{j=2}^{n}\Delta_{\widehat{\pi}}\xi_j\Delta_{\widehat{\pi}}Y_j + \frac{1}{n}\sum_{j=2}^{n}\Delta_{\widehat{\pi}}e_j\Delta_{\widehat{\pi}}Y_j.$$

Under the maintained second-moment condition,

$$\frac{1}{n}\sum_{j=2}^{n}(\Delta_{\widehat{\pi}}Y_j)^2 = O_{\mathbb{P}}(1).$$

Another application of Cauchy–Schwarz yields

$$\left|\frac{1}{n}\sum_{j=2}^{n}\Delta_{\widehat{\pi}}e_j\Delta_{\widehat{\pi}}Y_j\right| = O_{\mathbb{P}}(n^{-1/2}),$$

and hence

$$B_n = \frac{1}{n}\sum_{j=2}^{n}\Delta_{\widehat{\pi}}\xi_j\Delta_{\widehat{\pi}}Y_j + o_{\mathbb{P}}(1). \tag{H2}$$

Step 2. Limit of the denominator.

By (H1), it remains to study

$$\frac{1}{n}\sum_{j=2}^{n}(\Delta_{\widehat{\pi}}\xi_j)^2.$$

Expanding the square gives

$$\frac{1}{n}\sum_{j=2}^{n}(\Delta_{\widehat{\pi}}\xi_j)^2 = \frac{1}{n}\sum_{j=2}^{n}\xi^2_{\widehat{\pi}(j)} + \frac{1}{n}\sum_{j=2}^{n}\xi^2_{\widehat{\pi}(j-1)} - \frac{2}{n}\sum_{j=2}^{n}\xi_{\widehat{\pi}(j)}\xi_{\widehat{\pi}(j-1)}.$$

The first sum contains every observation except $\widehat{\pi}(1)$, while the second contains every observation except $\widehat{\pi}(n)$. Since $|\xi_i| \le 1$, these two omitted endpoint terms are $o_{\mathbb{P}}(1)$. The ordinary law of large numbers then gives

$$\frac{1}{n}\sum_{j=2}^{n}\xi^2_{\widehat{\pi}(j)} \xrightarrow{p} \mathbb{E}(\xi_i^2), \qquad \frac{1}{n}\sum_{j=2}^{n}\xi^2_{\widehat{\pi}(j-1)} \xrightarrow{p} \mathbb{E}(\xi_i^2).$$

Moreover,

$$\mathbb{E}(\xi_i^2 \mid X_i) = \mathrm{Var}(D_i \mid X_i) = e(X_i)\{1 - e(X_i)\},$$

so $\mathbb{E}(\xi_i^2) = \mathbb{E}[e(X_i)\{1 - e(X_i)\}]$.

It remains to consider the product involving consecutive observations. For a deterministic local ordering, distinct observations are independent and

$$\mathbb{E}(\xi_i \xi_k \mid X_i, X_k) = \mathbb{E}(\xi_i \mid X_i)\mathbb{E}(\xi_k \mid X_k) = 0, \qquad i \neq k.$$

The consecutive-product law of large numbers used in the proof of theorem 1, uniformly over the shrinking neighborhood of α_0, therefore gives

$$\frac{1}{n}\sum_{j=2}^{n} \xi_{\widehat{\pi}(j)} \xi_{\widehat{\pi}(j-1)} = o_{\mathbb{P}}(1).$$

Combining the two diagonal limits and the vanishing product term yields

$$A_n \xrightarrow{p} 2\mathbb{E}[e(X_i)\{1 - e(X_i)\}].$$

Step 3. Limit of the numerator.

By (H2), expand the remaining term as

$$\begin{aligned}\frac{1}{n}\sum_{j=2}^{n} \Delta_{\widehat{\pi}} \xi_j \Delta_{\widehat{\pi}} Y_j = \frac{1}{n}\sum_{j=2}^{n} \xi_{\widehat{\pi}(j)} Y_{\widehat{\pi}(j)} + \frac{1}{n}\sum_{j=2}^{n} \xi_{\widehat{\pi}(j-1)} Y_{\widehat{\pi}(j-1)} \\ - \frac{1}{n}\sum_{j=2}^{n} \xi_{\widehat{\pi}(j)} Y_{\widehat{\pi}(j-1)} - \frac{1}{n}\sum_{j=2}^{n} \xi_{\widehat{\pi}(j-1)} Y_{\widehat{\pi}(j)}.\end{aligned}$$

The first two sums differ from $n^{-1}\sum_i \xi_i Y_i$ only through one endpoint each. The maintained moment condition makes the endpoint contributions $o_{\mathbb{P}}(1)$, so both sums converge to $\mathbb{E}(\xi_i Y_i)$.

The two products involving consecutive observations vanish. Indeed, for $i \neq k$,

$$\mathbb{E}(\xi_i Y_k \mid X_i, X_k) = \mathbb{E}(\xi_i \mid X_i)\mathbb{E}(Y_k \mid X_k) = 0.$$

The same uniform consecutive-product law of large numbers used above therefore gives an $o_{\mathbb{P}}(1)$ contribution from each product.

It remains to calculate the diagonal expectation. Let $\mu_d(X_i) \equiv \mathbb{E}\{Y_i(d) \mid X_i\}$. Since $Y_i = D_i Y_i(1) + (1 - D_i)Y_i(0)$, unconfoundedness gives

$$\begin{aligned}\mathbb{E}(\xi_i Y_i \mid X_i) &= \mathbb{E}[(D_i - e(X_i))D_i Y_i(1) \mid X_i] \\ &\quad + \mathbb{E}[(D_i - e(X_i))(1 - D_i)Y_i(0) \mid X_i] \\ &= e(X_i)\{1 - e(X_i)\}\mu_1(X_i) \\ &\quad - e(X_i)\{1 - e(X_i)\}\mu_0(X_i) \\ &= e(X_i)\{1 - e(X_i)\}\tau(X_i).\end{aligned}$$

Combining the two diagonal limits and the two vanishing products yields

$$B_n \xrightarrow{p} 2\mathbb{E}[e(X_i)\{1 - e(X_i)\}\tau(X_i)].$$

Step 4. Form the ratio.

Since $\widehat{\beta}_T = B_n / A_n$, the continuous mapping theorem and the preceding two limits give (39).