

HIGHER-ORDER NEYMAN ORTHOGONALITY IN MOMENT-CONDITION MODELS*

Stéphane Bonhomme[‡]

Koen Jochmans[§]

Whitney K. Newey[¶]

Martin Weidner^{||}

May 11, 2026

Abstract

We construct moment functions that are Neyman-orthogonal to a chosen order in parametric moment condition models. These moment functions reduce sensitivity to nuisance estimation error and, as such, offer a unified and tractable route to higher-order debiasing in a wide range of econometric models. The number of additional nuisance parameters required by our construction, beyond those already present in the original moment conditions, is independent of the order of orthogonalization and can be reduced to a single scalar if desired.

Keywords: Neyman-orthogonality, higher-order bias correction, moment conditions, GMM, U-statistics.

*Funded by the European Union (ERC-NETWORK-101044319), by the French Government and the French National Research Agency under the Investissements d' Avenir program (ANR-17-EURE-0010), and through NSF Grant 2242447. Views and opinions expressed are however those of the authors only and do not necessarily reflect those of the European Union or the European Research Council. Neither the European Union nor the granting authority can be held responsible for them.

[‡]Department of Economics, University of Chicago. Email: sbonhomme@uchicago.edu

[§]Toulouse School of Economics, Université Toulouse Capitole. Email: koen.jochmans@tse-fr.eu

[¶]Department of Economics, MIT, and NBER. Email: wnewey@mit.edu

^{||}Dept. of Economics & Nuffield College, Univ. of Oxford. Email: martin.weidner@economics.ox.ac.uk

1 Introduction

Estimation and inference in the presence of nuisance parameters is a central problem in econometrics and statistics. The generalized method of moments (GMM, Hansen, 1982) provides a unifying framework for a vast range of estimators, from ordinary least squares and instrumental-variable estimation to maximum likelihood and causal-inference procedures. In all these settings, the parameter of interest must be estimated in the presence of additional unknown quantities. The accuracy with which these nuisance parameters are estimated can have a substantial effect on inference for the parameter of interest.

In such a setting, it is helpful to work with estimating equations that are orthogonal to the nuisance parameters in the sense of Neyman (1959). (First-order) orthogonality means that the first derivative of the expected estimating equation with respect to the nuisance parameter vanishes at the true parameter values.¹ First-order Neyman orthogonality, when combined with sample splitting, permits construction of \sqrt{n} -consistent estimators as long as the nuisance parameter is estimated at a rate faster than $n^{-1/4}$, where n is the sample size (Chernozhukov et al., 2018).

The faster-than- $n^{-1/4}$ requirement is binding in many important problems. It fails when the nuisance parameter is high-dimensional relative to the sample size, when its convergence rate is slow, or both. One example is panel data with individual fixed effects, where the incidental parameter problem persists under first-order orthogonality (Neyman and Scott, 1948; Li, Lindsay and Waterman, 2003; Hahn and Newey, 2004; Jochmans and Weidner, 2019; Kline, Saggio and S¸olvsten, 2020). Another example is high-dimensional regression (e.g., Mikusheva and S¸olvsten, 2025), where approximate sparsity conditions are needed for post-double-Lasso inference to be valid (W¸uthrich and Zhu, 2023; Sur and Cand¸es, 2019). Two further examples are instrumental-variable estimation, where machine-learning-based variable selection introduces biases that first-order corrections can fail to offset (Angrist and Frandsen, 2022), and nonparametric models with moderate-dimensional nuisance functions, where minimax rates for estimating those functions fall short of $n^{-1/4}$ (Robins, Li, Tchetgen and van der Vaart, 2008; van der Vaart, 2014).

A natural response is to go beyond first-order orthogonality. Mackey, Syrgkanis and Zadik (2018) formalized the concept of q -th order Neyman orthogonality. The condition requires that all derivatives of the expected estimating equation with respect to the nuisance

¹Neyman-orthogonal estimating equations play a central role in semiparametric estimation (Bickel, Klaassen, Ritov and Wellner, 1993; Newey, 1994). They are also at the heart of the recent literature on debiased inference in high-dimensional models (Belloni, Chernozhukov and Hansen, 2014; Javanmard and Montanari, 2014; van de Geer, B¸uhlmann, Ritov and Dezeure, 2014; Zhang and Zhang, 2014) and on double/debiased machine learning (Chernozhukov et al., 2018).

parameter, up to order q , vanish at the true parameter values. They showed that q -th order orthogonality, again combined with sample splitting, permits valid inference when the nuisance parameter converges at rate $n^{-1/2(q+1)}$; for example, the rate requirement drops from $n^{-1/4}$ to $n^{-1/6}$ in going from first to second order. Constructing moment functions that are orthogonal beyond first order has nevertheless proved difficult. This paper provides general constructions of q -th order Neyman-orthogonal moment functions for parameters that satisfy parametric unconditional moment conditions, with a separate target functional. We derive a closed-form, non-recursive expression for the q -th order orthogonal moment function in a general parametric setting where both the target moment function and the one that identifies the nuisance parameter can be nonlinear.

The orthogonal moment function depends on multiple independent copies of the data, and it features a generalized inverse of a Jacobian matrix as an additional nuisance parameter whose dimension does not grow with the orthogonalization order q . We first provide explicit moment functions in the case where the moment function for the (original) nuisance parameter is affine; this covers linear regression, instrumental-variable estimation with linear moment conditions, and grouped data models with unit-specific intercepts or slopes. We then extend our construction to the case where both moment functions are nonlinear, in which case additional correction terms are needed. Our explicit expression for orthogonal moment functions in the nonlinear case is based on rooted trees. A noteworthy feature of our approach is that the dimension of the additional nuisance parameter can be dramatically reduced: by replacing the moment function with a transformed version constructed from subdeterminants of the Jacobian matrix, the nuisance can be reduced to a single scalar, even under overidentification.

Our paper is related to several recent contributions. [Mackey, Syrgkanis and Zadik \(2018\)](#) provide an explicit second-order orthogonal moment for the partially linear regression model, but only when first-stage errors are non-Gaussian; they prove an impossibility result under Gaussianity. Their constructive results do not extend to general models. Our constructions circumvent this impossibility because we rely on independent copies of the data and on ordinary finite-dimensional derivatives, rather than on functional derivatives. An approach to constructing q -th order orthogonal estimating equations for conditional-likelihood models was provided in [Bonhomme, Jochmans and Weidner \(2025\)](#). In recent independent work, [Chetverikov, Sorensen and Tsyvinski \(2026\)](#) propose a recursive formula for Z-estimation problems and apply it to a “triple-Lasso” estimator with second-order orthogonality. Like our construction, theirs uses independent copies of the data. They also provide numerical evidence that higher-order orthogonality can substantially improve coverage of confidence intervals. Their construction differs from ours in several respects. In particular, they in-

introduce new nuisance parameters at each recursion step, so the nuisance dimension grows with the order of orthogonality. They also restrict attention to exactly-identified moment conditions. In a different direction, [Robins, Li, Tchetgen and van der Vaart \(2008\)](#), [van der Vaart \(2014\)](#), and [Robins, Li, Mukherjee, Tchetgen and van der Vaart \(2017\)](#) develop higher-order influence functions and U-statistic-based estimators for nonparametric models. Those results rely on nonparametric von Mises expansions that hold only approximately. They also do not provide explicit constructions for parameters defined by finite-dimensional moment conditions.

The rest of the paper is organized as follows. [Section 2](#) describes the setup. [Section 3](#) illustrates the main ideas in a heterogeneous-coefficient model for grouped data. [Section 4](#) presents the general construction for the affine case and the nonlinear case, in turn, and also gives a nuisance-dimension reduction technique. [Section 5](#) reports implementation and simulation evidence for the heterogeneous coefficients example. [Section 6](#) establishes the asymptotic distribution of the resulting estimator. Finally, the Appendix contains proofs and additional formulas.

2 Setup

2.1 Model and parameter of interest

We observe a sample $\{W_i : i = 1, \dots, n\}$ from an unknown distribution P_0 . Consider a model defined by two moment conditions,

$$\mathbb{E}[g(W_i, \theta_0, \eta_0)] = 0, \quad \mathbb{E}[m(W_i, \theta_0, \eta_0)] = 0, \quad (1)$$

where $\theta_0 \in \mathbb{R}^{d_\theta}$ is the parameter of interest and $\eta_0 \in \mathbb{R}^{d_\eta}$ is a finite-dimensional nuisance parameter. The functions $g(W_i, \theta, \eta) \in \mathbb{R}^{d_g}$ and $m(W_i, \theta, \eta) \in \mathbb{R}$ are known. All expectations are taken with respect to P_0 .

Assumption 1 (i.i.d. sample). *The observations W_1, \dots, W_n are independent and identically distributed draws from P_0 .*

The first moment condition identifies η_0 for given θ_0 . We assume $d_g \geq d_\eta$, which allows the nuisance parameter to be either exactly identified or overidentified. The second moment condition is used to identify θ_0 . We take m to be scalar throughout, purely for notational convenience. In practice one will often have $m(W_i, \theta, \eta) \in \mathbb{R}^{d_m}$ with $d_m > 1$, in which case our orthogonalization construction is applied to each component of m separately. Identification of θ_0 in a given application requires $d_m \geq d_\theta$, and the target-moment system in θ may itself

be exactly identified or overidentified. These identification considerations do not affect the orthogonalization construction that we present.

Our goal is to construct a Neyman-orthogonal moment function ψ that replaces the target moment function m . The orthogonal moment function has expected derivatives in η that vanish up to order q at the true parameter values, which makes it less sensitive to estimation error in η . To achieve orthogonality to a given order, our construction relies, in addition to m , on a moment condition for the nuisance parameter, $\mathbb{E}[g(W_i, \theta_0, \eta_0)] = 0$.²

Lastly, in our setup, η_0 is a vector of finite dimension (which may grow with the sample size). The possible extension to infinite-dimensional nuisance parameters, as in the literature on higher-order influence functions (Robins, Li, Tchetgen and van der Vaart, 2008; van der Vaart, 2014), is left for future work.

2.2 Nuisance parameters and the Jacobian inverse

Let ∂_η be the gradient with respect to η . Our construction requires a left inverse of the $d_g \times d_\eta$ Jacobian matrix

$$J_0 := \mathbb{E}[\partial_\eta g(W_i, \theta_0, \eta_0)]. \quad (2)$$

Let Λ_0 be a $d_\eta \times d_g$ matrix satisfying $\Lambda_0 J_0 = I_{d_\eta}$. When $d_g = d_\eta$ and J_0 is invertible, $\Lambda_0 = J_0^{-1}$. When $d_g > d_\eta$ and J_0 has full column rank, we can set $\Lambda_0 = (J_0' \Omega^{-1} J_0)^{-1} J_0' \Omega^{-1}$ for a positive-definite weight matrix Ω .

The matrix Λ_0 enters as an additional nuisance parameter in our construction. We parametrize it as $\Lambda(\lambda)$ for a finite-dimensional parameter $\lambda \in \mathbb{R}^{d_\lambda}$, with true value λ_0 satisfying $\Lambda(\lambda_0) = \Lambda_0$. The simplest choice is $\lambda = \text{vec}(\Lambda)$, giving $d_\lambda = d_\eta \cdot d_g$.

Assumption 2 (Jacobian left inverse). *The Jacobian matrix J_0 has full column rank, so a left inverse Λ_0 with $\Lambda_0 J_0 = I_{d_\eta}$ exists. The map $\lambda \mapsto \Lambda(\lambda)$ satisfies $\Lambda(\lambda_0) = \Lambda_0$.*

While the dimension d_λ does not enter the orthogonalization construction itself, d_λ does matter in applications, because the accuracy with which $\hat{\Lambda}$ can be estimated depends on d_λ , with a smaller d_λ potentially leading to a more accurate preliminary estimate. For this reason, Section 4.3 develops techniques that reduce d_λ , to a single scalar if desired, by replacing the moment function g with a suitably transformed version.

We first briefly illustrate the setup with two cross-sectional examples, and we will study a grouped data model in more detail in the next section.

²In our construction, the parameter θ plays a passive role. It is carried through as an argument of m and g . Only after the orthogonal moment function has been constructed does θ enter the analysis as the parameter to be estimated, through the orthogonal moment equations. Thus, one could drop θ from most of the formulas in Sections 3 through 4 without changing their content, but we retain it throughout the main text to avoid notational confusion.

Example 1 (Linear instrumental variables, including linear regression as a special case). Let $W_i = (Y_i, D_i, Z_i, X_i)$ with $D_i \in \mathbb{R}$ a regressor of interest, $Z_i \in \mathbb{R}$ an instrument satisfying $\mathbb{E}[Z_i D_i] \neq 0$, and $X_i \in \mathbb{R}^K$ exogenous controls. Consider

$$Y_i = D_i \theta_0 + X_i' \eta_0 + U_i, \quad \mathbb{E}[X_i U_i] = 0, \quad \mathbb{E}[Z_i U_i] = 0.$$

The parameter of interest is $\theta_0 \in \mathbb{R}$. The framework (1) applies with

$$g(W_i, \theta, \eta) = X_i(Y_i - D_i \theta - X_i' \eta), \quad m(W_i, \theta, \eta) = Z_i(Y_i - D_i \theta - X_i' \eta).$$

Both moment functions are affine in η , and the Jacobian is $J_0 = -\mathbb{E}[X_i X_i']$, so $\Lambda_0 = -(\mathbb{E}[X_i X_i'])^{-1}$ is minus the inverse of the population second-moment matrix of the controls. The special case $Z_i = D_i$ corresponds to linear regression with high-dimensional controls, in which θ_0 is the coefficient on D_i in the projection of Y_i on (D_i, X_i) and the additional moment condition $\mathbb{E}[D_i U_i] = 0$ is needed for identification.

Example 2 (Mean of a nonlinear function of a generated regressor). As an instance of the two-step generated-regressor setting studied by Cattaneo, Jansson and Ma (2019), let $W_i = (R_i, X_i)$ with $R_i \in \mathbb{R}$ and $X_i \in \mathbb{R}^{d_\eta}$, and let f be a known nonlinear function. Define $\eta_0 = (\mathbb{E}[X_i X_i'])^{-1} \mathbb{E}[X_i R_i]$ as the coefficient of the linear projection of R_i on X_i , so that $\mu_i := X_i' \eta_0$ is the population fitted value. The parameter of interest is

$$\theta_0 = \mathbb{E}[f(\mu_i)] = \mathbb{E}[f(X_i' \eta_0)].$$

The framework applies with

$$g(W_i, \theta, \eta) = X_i(R_i - X_i' \eta), \quad m(W_i, \theta, \eta) = f(X_i' \eta) - \theta.$$

Here g is affine in η , but m is nonlinear in η through f .

Remark 1 (Λ as a by-product of estimating η). Although Assumption 2 formally introduces $\Lambda(\lambda)$ as an additional nuisance parameter, in practice $\widehat{\Lambda}$ is often a by-product of estimating η rather than a separate quantity. This is most transparent in the affine case. In Example 1, computing the least-squares estimator of η requires inverting the sample second-moment matrix $n^{-1} \sum_i X_i X_i'$, and that inversion already delivers an estimator for Λ at no additional cost. The same is true in nonlinear settings, where the Jacobian inverse at the solution $\widehat{\eta}$ is computed as part of the numerical optimization.

2.3 Higher-order Neyman orthogonality

Let $\psi(W_1, \dots, W_L; \theta, \eta, \lambda) \in \mathbb{R}$ be a moment function that depends on $L \geq 1$ independent copies of W_i , the parameter of interest θ , the nuisance parameter η , and the Jacobian-inverse parameter λ . Define the population moment

$$\Psi(\theta, \eta, \lambda) := \mathbb{E}[\psi(W_1, \dots, W_L; \theta, \eta, \lambda)]. \quad (3)$$

Definition 1 (Higher-order Neyman orthogonality). *For $q \in \{1, 2, 3, \dots\}$, the moment function ψ is q -th order Neyman-orthogonal if*

$$\begin{aligned} \Psi(\theta_0, \eta_0, \lambda_0) &= 0, \\ \partial_\eta^\alpha \partial_\lambda^\beta \Psi(\theta_0, \eta_0, \lambda_0) &= 0 \quad \text{for all multi-indices } (\alpha, \beta) \text{ with } 1 \leq |\alpha| + |\beta| \leq q, \end{aligned} \quad (4)$$

where $\alpha = (\alpha_1, \dots, \alpha_{d_\eta})$ and $\beta = (\beta_1, \dots, \beta_{d_\lambda})$ are vectors of non-negative integers, $|\alpha| = \sum_j \alpha_j$, and $\partial_\eta^\alpha \partial_\lambda^\beta = \partial^{|\alpha|+|\beta|} / (\partial \eta_1^{\alpha_1} \dots \partial \eta_{d_\eta}^{\alpha_{d_\eta}} \partial \lambda_1^{\beta_1} \dots \partial \lambda_{d_\lambda}^{\beta_{d_\lambda}})$.

When $q = 1$, Definition 1 reduces to the standard Neyman orthogonality condition. When $q = 0$, no orthogonality is required and $\psi(W_i; \theta, \eta) = m(W_i, \theta, \eta)$ suffices (with $L = 1$). The condition (4) imposes orthogonality in η and λ jointly. Beyond the pure η -derivatives and pure λ -derivatives of Ψ , it also requires all mixed derivatives $\partial_\eta^\alpha \partial_\lambda^\beta \Psi$ with $|\alpha| \geq 1$, $|\beta| \geq 1$, and $|\alpha| + |\beta| \leq q$ to vanish at the true values. The pure η -conditions control bias from estimating the original nuisance parameter, the pure λ -conditions control bias from estimating the Jacobian inverse, and the mixed conditions control bias arising from the interaction of the two estimation errors in the Taylor expansion of Ψ . All three kinds of condition are necessary for q -th order orthogonality.

Estimation. Given an i.i.d. sample $\{W_i : i = 1, \dots, n\}$, the natural sample analogue of the population moment $\Psi(\theta, \eta, \lambda)$ is the U-statistic

$$\widehat{\Psi}_n(\theta, \eta, \lambda) := \frac{1}{n(n-1) \cdots (n-L+1)} \sum_{\substack{i_1, \dots, i_L=1 \\ \text{all distinct}}}^n \psi(W_{i_1}, \dots, W_{i_L}; \theta, \eta, \lambda).$$

An estimator $\widehat{\theta}$ of θ_0 is obtained by solving $\widehat{\Psi}_n(\widehat{\theta}, \widehat{\eta}, \widehat{\lambda}) = 0$, where $\widehat{\eta}$ and $\widehat{\lambda}$ are preliminary estimators of the nuisance parameters. To ensure that the estimation error in $(\widehat{\eta}, \widehat{\lambda})$ does not contaminate the U-statistic, the sample is split into two independent parts. One part is used to construct $\widehat{\eta}$ and $\widehat{\lambda}$, and the other is used to evaluate $\widehat{\Psi}_n$. Efficiency can be improved by cross-fitting, that is by swapping the roles of the two sample parts and averaging, as in

Chernozhukov et al. (2018).

The key consequence of q -th order orthogonality is that, under appropriate regularity conditions, the estimator satisfies $\widehat{\theta} - \theta_0 = O_p(\|\widehat{\eta} - \eta_0\|^{q+1} + \|\widehat{\lambda} - \lambda_0\|^{q+1} + n^{-1/2})$, since all lower-order terms in the Taylor expansion of Ψ around $(\theta_0, \eta_0, \lambda_0)$ are of smaller order because of the orthogonality conditions. Valid inference is therefore possible even when the nuisance parameters converge at rates much slower than $n^{-1/4}$, provided q is chosen large enough. We provide more details on the asymptotic theory in Section 6.

3 Illustration: a heterogeneous coefficients model

Before presenting the general construction of higher-order orthogonal moment functions, we illustrate the main ideas in a concrete example.

Consider a grouped data model with heterogeneous coefficients,

$$Y_{it} = X'_{it}\eta_{i0} + U_{it}, \quad i = 1, \dots, N, \quad t = 1, \dots, T, \quad (5)$$

for a scalar outcome Y_{it} and a d_η -dimensional covariate vector X_{it} , under the uncorrelatedness condition

$$\mathbb{E}[X_{it}U_{it}] = 0, \quad i = 1, \dots, N, \quad t = 1, \dots, T. \quad (6)$$

Such models have been studied extensively under the mean independence condition $\mathbb{E}[U_{it} | X_{i1}, \dots, X_{iT}] = 0$ (e.g., Chamberlain, 1992). Condition (6) is weaker, since it does not restrict the conditional mean of Y_{it} given X_{it} to be linear. The coefficients η_{i0} are the best linear predictors,

$$\eta_{i0} = \left(\mathbb{E} \left[\sum_{t=1}^T X_{it}X'_{it} \right] \right)^{-1} \mathbb{E} \left[\sum_{t=1}^T X_{it}Y_{it} \right].$$

We focus on the target parameter

$$\theta_0 = \frac{1}{N} \sum_{i=1}^N m(\eta_{i0}),$$

where m is a smooth scalar function. The mean and second moment of the components of η_{i0} are natural choices, for example to quantify the level and dispersion of the coefficients across units. As an application, Kline, Rose and Walters (2022) study employment discrimination by sending randomized fictitious job applications to firms and measuring callbacks Y_{it} . There, η_{i0} captures how each firm i responds to applicant characteristics X_{it} such as race, gender, and age, and moments of η_{i0} quantify the extent and variation of discrimination across firms.

This model fits the framework of Section 2. For each unit i , write the observation at t as

$W_{it} = (Y_{it}, X_{it})$. The nuisance-identifying moment function is

$$g(W_{it}, \theta, \eta_i) = X_{it}(Y_{it} - X_{it}'\eta_i),$$

and condition (6) is equivalent to $\mathbb{E}[g(W_{it}, \theta_0, \eta_{i0})] = 0$ for every t . The Jacobian-inverse nuisance parameter is $\Lambda_{i0} = -(\mathbb{E}[X_{it}X_{it}'])^{-1}$, and the target-moment function for a given unit i is $m_i(W_{it}, \theta, \eta_i) = m(\eta_i) - \theta$.

The observations W_{it} (i.e., job applications) are independent across t for each i (i.e., within each firm). The moment conditions for η_{i0} are affine in η_i , so this example falls within the affine case of Section 4.1 below. The ordinary least-squares estimators

$$\hat{\eta}_i^{\text{OLS}} = \left(\sum_{t=1}^T X_{it}X_{it}' \right)^{-1} \sum_{t=1}^T X_{it}Y_{it}, \quad i = 1, \dots, N,$$

are noisy when T is small, and the plug-in estimator $\hat{\theta}^{\text{plug-in}} = (1/N) \sum_{i=1}^N m(\hat{\eta}_i^{\text{OLS}})$ is biased. Our approach constructs alternative estimators with reduced bias.

To apply the framework from Section 2 to this setting, let us fix a single unit i and consider a sample of observations for $t = 1, \dots, T$. Each observation W_{it} plays the role of W_i in Section 2, the pair $(\eta_{i0}, \Lambda_{i0})$ plays the role of (η_0, Λ_0) , and the orthogonal moment function ψ becomes a function of several observations for the same unit. The U-statistic constructions that follow are therefore applied across t within a unit. The unit dimension $i = 1, \dots, N$ plays no role at the orthogonalization stage. It enters only through the final step, in which the unit-level orthogonal moment functions are averaged over i to form the estimator of θ_0 .³

3.1 Main ideas

Since the problem stratifies across units, we focus on a single unit i and describe the construction at the unit level. It is useful to introduce the reparameterization $A_{i0} = -\Lambda_{i0}^{-1}$ and $b_{i0} = -\Lambda_{i0}^{-1}\eta_{i0}$, with $a_{i0} = \text{vech}(A_{i0})$ denoting the vector that contains the elements of A_{i0} .

The construction relies on three observations.

The *first observation* is that A_{i0} and b_{i0} admit simple unbiased estimators:

$$\mathbb{E}[X_{it}X_{it}'] = A_{i0}, \quad \mathbb{E}[X_{it}Y_{it}] = b_{i0}. \quad (7)$$

³An equivalent approach is to stack $(\eta_{10}, \dots, \eta_{N0})$ into a single nuisance vector of dimension Nd_η and treat the full data set of $n = NT$ observations as one sample to which the Section 2 construction is applied directly. The two routes produce the same orthogonal moment functions. We adopt the unit-by-unit view because it keeps the effective nuisance dimension bounded as N grows and because it is conceptually simpler.

The *second observation* is that any polynomial $P(a_{i0}, b_{i0})$ admits an unbiased estimator. Each element of (a_{i0}, b_{i0}) has an unbiased estimator by (7). Since observations are independent across t , one can combine estimators from different t observations to produce unbiased estimators of powers and products. For example, when $d_\eta = 1$, an unbiased estimator of $a_{i0}^2 = (\mathbb{E}[X_{it}^2])^2$ is $X_{i1}^2 X_{i2}^2$, since $\mathbb{E}[X_{i1}^2 X_{i2}^2] = \mathbb{E}[X_{i1}^2] \mathbb{E}[X_{i2}^2] = a_{i0}^2$.

The *third observation* is that $m(\eta_{i0})$ can be approximated by a polynomial in (a_{i0}, b_{i0}) . Writing $m(\eta_{i0}) = \varphi(a_{i0}, b_{i0})$ with $\varphi(a, b) = m(A^{-1}b)$, the function φ has a singularity at $\det A = 0$ and so cannot be globally represented as a polynomial. However, given preliminary estimators $\widehat{\Lambda}_i$ and $\widehat{\eta}_i$ that are independent of the remaining data, one can expand locally: setting $\widehat{a}_i = -\text{vec}(\widehat{\Lambda}_i^{-1})$ and $\widehat{b}_i = -\widehat{\Lambda}_i^{-1} \widehat{\eta}_i$,

$$m(\eta_{i0}) = \sum_{|\alpha|+|\beta|\leq q} \frac{1}{\alpha! \beta!} \partial_a^\alpha \partial_b^\beta \varphi(\widehat{a}_i, \widehat{b}_i) (a_{i0} - \widehat{a}_i)^\alpha (b_{i0} - \widehat{b}_i)^\beta + R_{q,i}, \quad (8)$$

where the remainder $R_{q,i}$ depends on powers of order $q + 1$ in $a_{i0} - \widehat{a}_i$ and $b_{i0} - \widehat{b}_i$. Each term $(a_{i0} - \widehat{a}_i)^\alpha (b_{i0} - \widehat{b}_i)^\beta$ admits an unbiased estimator built from independent observations. This yields an approximately unbiased estimator of $m(\eta_{i0})$ based on a q -th order orthogonal moment function.

In practice, the preliminary estimators $\widehat{\Lambda}_i$ and $\widehat{\eta}_i$ are obtained by holding out a set of t observations. Efficiency can be improved by cross-fitting. A feature of this construction is that the nuisance parameters $\widehat{\Lambda}_i$ and $\widehat{\eta}_i$ do not change with the order of orthogonalization, q .

3.2 First example: the normal-means model without normality

Consider first the special case without covariates,

$$Y_{it} = \eta_{i0} + U_{it}, \quad \mathbb{E}[U_{it}] = 0, \quad i = 1, \dots, N, \quad t = 1, \dots, T, \quad (9)$$

where U_{it} are i.i.d. but not necessarily normally distributed. This is the [Neyman and Scott \(1948\)](#) model without the normality assumption. Here $d_\eta = 1$ and $\Lambda_{i0} = -1$ is known, so η_{i0} is the only nuisance parameter.

A q -th order Taylor expansion of $m(\eta_{i0})$ around a preliminary estimator $\widehat{\eta}_i$ gives

$$m(\eta_{i0}) = \sum_{k=0}^q \frac{\partial_\eta^k m(\widehat{\eta}_i)}{k!} (\eta_{i0} - \widehat{\eta}_i)^k + R_{q,i}.$$

Expanding $(\eta_{i0} - \hat{\eta}_i)^k = \sum_{j=0}^k \binom{k}{j} \eta_{i0}^j (-\hat{\eta}_i)^{k-j}$ and interchanging the order of summation,

$$m(\eta_{i0}) = \sum_{j=0}^q \left(\sum_{k=j}^q \frac{\partial^k m(\hat{\eta}_i)}{k!} \binom{k}{j} (-\hat{\eta}_i)^{k-j} \right) \eta_{i0}^j + R_{q,i}.$$

Since, by independence and $\mathbb{E}[U_{it}] = 0$,

$$\mathbb{E} \left[\prod_{s=1}^j Y_{is} \right] = \eta_{i0}^j, \quad (10)$$

an approximately unbiased estimator of $m(\eta_{i0})$ is

$$\sum_{j=0}^q \left(\sum_{k=j}^q \frac{\partial^k m(\hat{\eta}_i)}{k!} \binom{k}{j} (-\hat{\eta}_i)^{k-j} \right) \prod_{s=1}^j Y_{is}. \quad (11)$$

A lower-variance estimator is obtained by averaging (11) over all subsets of q observations, yielding a U-statistic.

Orthogonality. Define the moment function

$$\psi(W_i; \theta, \eta_i) = \sum_{j=0}^q \left(\sum_{k=j}^q \frac{\partial^k m(\eta_i)}{k!} \binom{k}{j} (-\eta_i)^{k-j} \right) \prod_{s=1}^j Y_{is} - \theta, \quad (12)$$

where $W_i = (Y_{i1}, \dots, Y_{iT})$. Using (10), reversing the order of summation, and applying the binomial theorem, we have

$$\mathbb{E}[\psi(W_i; \theta, \eta_i)] = \sum_{k=0}^q \frac{\partial^k m(\eta_i)}{k!} (\eta_{i0} - \eta_i)^k - \theta. \quad (13)$$

At (θ_{i0}, η_{i0}) with $\theta_{i0} = m(\eta_{i0})$, the right-hand side equals $m(\eta_{i0}) - \theta_{i0} = 0$. Write $f(\eta_i) = \sum_{k=0}^q \frac{\partial^k m(\eta_i)}{k!} (\eta_{i0} - \eta_i)^k$. A direct computation using the product rule shows that the derivative telescopes:

$$\partial_\eta f(\eta_i) = \frac{\partial_\eta^{q+1} m(\eta_i)}{q!} (\eta_{i0} - \eta_i)^q, \quad (14)$$

which vanishes at $\eta_i = \eta_{i0}$. By Leibniz' rule, the j -th derivative of f at $\eta_i = \eta_{i0}$ also vanishes for all $1 \leq j \leq q$, since every term contains a factor $(\eta_{i0} - \eta_i)^{q+1-j+\ell}$ for some $\ell \geq 0$. The moment function ψ is, therefore, q -th order Neyman-orthogonal in the sense of Definition 1.

3.3 Second example: a scalar covariate at second order

We now consider the model (5) with a scalar covariate ($d_\eta = 1$) and construct a second-order orthogonal moment function ($q = 2$). Unlike the previous example, the Jacobian inverse $\Lambda_{i0} = -(\mathbb{E}[X_{it}^2])^{-1}$ is now unknown and enters as an additional nuisance parameter.

A second-order Taylor expansion of $m(\eta_{i0})$ around $\hat{\eta}_i$ gives

$$m(\eta_{i0}) \approx m(\hat{\eta}_i) + \partial_\eta m(\hat{\eta}_i) (\eta_{i0} - \hat{\eta}_i) + \frac{1}{2} \partial_\eta^2 m(\hat{\eta}_i) (\eta_{i0} - \hat{\eta}_i)^2. \quad (15)$$

To express $\eta_{i0} - \hat{\eta}_i$ in terms of quantities with unbiased estimators, write $a_{i0} = -\Lambda_{i0}^{-1} = \mathbb{E}[X_{it}^2]$ and $b_{i0} = -\Lambda_{i0}^{-1} \eta_{i0} = \mathbb{E}[X_{it} Y_{it}]$. Expanding $\eta_{i0} = b_{i0}/a_{i0}$ around $(\hat{a}_i, \hat{b}_i) = (-\hat{\Lambda}_i^{-1}, -\hat{\Lambda}_i^{-1} \hat{\eta}_i)$ to second order,

$$\eta_{i0} - \hat{\eta}_i \approx -\hat{\Lambda}_i (b_{i0} - \hat{\eta}_i a_{i0}) - \hat{\Lambda}_i (1 + \hat{\Lambda}_i a_{i0}) (b_{i0} - \hat{\eta}_i a_{i0}).$$

Substituting into (15) and replacing a_{i0} and b_{i0} by their unbiased estimators (X_{it}^2 and $X_{it} Y_{it}$, respectively), we obtain the approximately unbiased estimator of $m(\eta_{i0})$

$$\begin{aligned} m(\hat{\eta}_i) - \partial_\eta m(\hat{\eta}_i) & \left(\hat{\Lambda}_i (X_{i1} Y_{i1} - \hat{\eta}_i X_{i1}^2) + \hat{\Lambda}_i (1 + \hat{\Lambda}_i X_{i2}^2) (X_{i1} Y_{i1} - \hat{\eta}_i X_{i1}^2) \right) \\ & + \frac{1}{2} \partial_\eta^2 m(\hat{\eta}_i) (\hat{\Lambda}_i (X_{i1} Y_{i1} - \hat{\eta}_i X_{i1}^2)) (\hat{\Lambda}_i (X_{i2} Y_{i2} - \hat{\eta}_i X_{i2}^2)). \end{aligned}$$

The resulting moment function is

$$\begin{aligned} \psi(W_i; \theta, \eta_i, \Lambda_i) & = m(\eta_i) - \partial_\eta m(\eta_i) \Lambda_i (X_{i1} Y_{i1} - \eta_i X_{i1}^2) \\ & \quad - \partial_\eta m(\eta_i) \Lambda_i (1 + \Lambda_i X_{i2}^2) (X_{i1} Y_{i1} - \eta_i X_{i1}^2) \\ & \quad + \frac{1}{2} \partial_\eta^2 m(\eta_i) \Lambda_i (X_{i1} Y_{i1} - \eta_i X_{i1}^2) \cdot \Lambda_i (X_{i2} Y_{i2} - \eta_i X_{i2}^2) - \theta. \end{aligned} \quad (16)$$

One can verify that $\mathbb{E}[\psi(W_i; \theta_{i0}, \eta_i, \Lambda_i)]$ and all its first and second derivatives in (η_i, Λ_i) at $(\eta_{i0}, \Lambda_{i0})$ vanish. The moment function is therefore second-order Neyman-orthogonal.

Both constructions in this section are special cases of the general affine formula in Theorem 1 below. The next section presents the general construction for arbitrary d_η and arbitrary order q , first for affine g and then for the fully nonlinear case.

4 General construction of orthogonal moment functions

We now present the general construction of higher-order orthogonal moment functions. The target moment function $m(W_i, \theta, \eta)$ is allowed to be nonlinear in η throughout. We first

treat the case where the nuisance-identifying moment function $g(W_i, \theta, \eta)$ is affine in η , and then extend to the case where g is also nonlinear.

4.1 Moment conditions affine in the nuisance parameter

Assume throughout this subsection that the moment function $g(W_i, \theta, \eta)$ is affine in η , so that $\partial_\eta g(W_i, \theta, \eta)$ does not depend on η , whereas the target moment function $m(W_i, \theta, \eta)$ can be nonlinear in η . This covers many important cases. Linear regression and linear instrumental variables (Example 1) have g and m both affine in η , while the average effect of a generated regressor (Example 2) has g affine and m nonlinear in η .

We begin by introducing notation for multilinear derivatives that is used throughout this section. For $r \geq 1$, write $\partial_\eta^r m(W_i, \theta, \eta)$ for the r -th derivative tensor of m with respect to η . For vectors $v^{(1)}, \dots, v^{(r)} \in \mathbb{R}^{d_\eta}$, denote the multilinear action by

$$\langle \partial_\eta^r m(W_i, \theta, \eta), v^{(1)} \otimes \dots \otimes v^{(r)} \rangle = \sum_{j_1, \dots, j_r} \frac{\partial^r m(W_i, \theta, \eta)}{\partial \eta_{j_1} \dots \partial \eta_{j_r}} v_{j_1}^{(1)} \dots v_{j_r}^{(r)}.$$

For $r = 1$ this gives the directional derivative $\partial_\eta m(W_i, \theta, \eta)' v^{(1)}$. For $r = 2$ it gives the quadratic form $v^{(1)'} \partial_{\eta\eta'}^2 m(W_i, \theta, \eta) v^{(2)}$.

For an integer $q \geq 0$, the orthogonal moment function is built from factors Z_k , each using k independent copies of W_i . The outer sum aggregates over all ways to distribute at most q copies among r such factors, with the r -th derivative of m providing the contraction at the root. Concretely, define

$$\psi^{(q)}(W; \theta, \eta, \lambda) := \sum_{r=0}^q \frac{1}{r!} \sum_{\substack{k_1, \dots, k_r \geq 1 \\ k_1 + \dots + k_r \leq q}} (-1)^{\sum_{s=1}^r k_s} \binom{q}{\sum_{s=1}^r k_s} \left\langle \partial_\eta^r m(W_1, \theta, \eta), \bigotimes_{s=1}^r Z_{k_s}(W^{(s)}; \theta, \eta, \lambda) \right\rangle, \quad (17)$$

where the tensor product $\bigotimes_{s=1}^r Z_{k_s}$ denotes $Z_{k_1} \otimes \dots \otimes Z_{k_r}$ (i.e., $\langle \partial_\eta^r m, \bigotimes_s Z_{k_s} \rangle$ is the r -linear contraction of $\partial_\eta^r m$ with the vectors Z_{k_1}, \dots, Z_{k_r}), the argument $W = (W_1, W^{(1)}, \dots, W^{(q)})$ consists of $q + 1$ independent copies of W_i , and

$$Z_{k_s}(W^{(s)}; \theta, \eta, \lambda) = \left(\prod_{\ell=1}^{k_s-1} \Lambda(\lambda) \partial_\eta g(W_{(s,\ell)}, \theta, \eta) \right) \Lambda(\lambda) g(W_{(s,k_s)}, \theta, \eta). \quad (18)$$

Each block $W^{(s)} = (W_{(s,1)}, \dots, W_{(s,k_s)})$ uses k_s independent copies, the copy W_1 enters only through $\partial_\eta^r m(W_1, \theta, \eta)$, and the $r = 0$ term uses the empty-product convention.

Each Z_{k_s} pre-multiplies $k_s - 1$ Jacobian factors $\Lambda \partial_\eta g$ onto a terminal factor Λg . The

binomial weight $(-1)^{\sum_{s=1}^r k_s} \binom{q}{\sum_{s=1}^r k_s}$ ensures that all monomials of degree 1 through q in the nuisance estimation error cancel in expectation.

Explicit formulas for small q . To make formula (17) concrete, we display $\psi^{(q)}$ for $q = 0, 1, 2$. Write $m := m(W_1, \theta, \eta)$, $m_\eta := \partial_\eta m(W_1, \theta, \eta)$, and $m_{\eta\eta} := \partial_{\eta\eta}^2 m(W_1, \theta, \eta)$.

Case $q = 0$.

$$\psi^{(0)}(W_1; \theta, \eta) = m(W_1, \theta, \eta).$$

Case $q = 1$.

$$\psi^{(1)}(W_1, W_2; \theta, \eta, \lambda) = m - m'_\eta \Lambda(\lambda) g(W_2, \theta, \eta).$$

Case $q = 2$.

$$\begin{aligned} \psi^{(2)}(W_1, W_2, W_3; \theta, \eta, \lambda) &= m - 2 m'_\eta \Lambda(\lambda) g(W_2, \theta, \eta) \\ &\quad + m'_\eta \left(\Lambda(\lambda) \partial_\eta g(W_3, \theta, \eta) \right) \Lambda(\lambda) g(W_2, \theta, \eta) \\ &\quad + \frac{1}{2} \left(\Lambda(\lambda) g(W_2, \theta, \eta) \right)' m_{\eta\eta} \left(\Lambda(\lambda) g(W_3, \theta, \eta) \right). \end{aligned} \quad (19)$$

The $q = 2$ moment function uses $L = 3$ independent copies. Explicit formulas for $q = 3$ and $q = 4$, using $L = 4$ and $L = 5$ copies respectively, are given in Appendix A.2.

Assumption 3.

- (i) $g(W_i, \theta, \eta)$ is affine in η .
- (ii) $m(W_i, \theta, \eta)$ is $(q + 1)$ -times continuously differentiable in η in a neighborhood of η_0 .
- (iii) The map $\lambda \mapsto \Lambda(\lambda)$ in Assumption 2 is $(q + 1)$ -times continuously differentiable in a neighborhood of λ_0 .
- (iv) For all $r \leq q + 1$, there is a neighborhood \mathcal{N} of η_0 and an integrable function $M_r(W_i)$ with $\sup_{\eta \in \mathcal{N}} \|\partial_\eta^r m(W_i, \theta_0, \eta)\| \leq M_r(W_i)$ almost surely.
- (v) $\mathbb{E}[\|g(W_i, \theta_0, \eta_0)\|] < \infty$ and $\mathbb{E}[\|\partial_\eta g(W_i, \theta_0, \eta_0)\|] < \infty$.

Theorem 1 (q -th order orthogonality, affine case). Define the population moment associated with $\psi^{(q)}$ in (17) by $\Psi^{(q)}(\theta, \eta, \lambda) := \mathbb{E}[\psi^{(q)}(W_1, \dots, W_L; \theta, \eta, \lambda)]$. Under Assumptions 1, 2, and 3,

$$\Psi^{(q)}(\theta_0, \eta_0, \lambda_0) = 0,$$

and, for all multi-indices (α, β) with $1 \leq |\alpha| + |\beta| \leq q$,

$$\partial_\eta^\alpha \partial_\lambda^\beta \Psi^{(q)}(\theta_0, \eta_0, \lambda_0) = 0.$$

That is, $\psi^{(q)}$ is q -th order Neyman-orthogonal in the sense of Definition 1.

The proof is in Appendix B.1.

Remark 2 (Comparison with standard first-order orthogonality). For $q = 1$, our moment function reduces to $\psi^{(1)} = m(W_1, \theta, \eta) - \partial_\eta m(W_1, \theta, \eta)' \Lambda(\lambda) g(W_2, \theta, \eta)$, which uses two independent copies and has the Jacobian inverse Λ as nuisance parameter. The standard first-order Neyman-orthogonal moment function (Newey, 1994; Chernozhukov et al., 2018) is $\psi_{\text{std}}^{(1)}(W_i; \theta, \eta, \lambda) = m(W_i, \theta, \eta) - \lambda' g(W_i, \theta, \eta)$, which uses a single observation and has a d_g -dimensional nuisance vector $\lambda_0 = \Lambda_0' \partial_\eta \mathbb{E}[m(W_i, \theta_0, \eta_0)]$. The standard moment function is simpler: it combines Λ_0 and $\partial_\eta \mathbb{E}[m]$ into a single nuisance λ and avoids the second copy. Our construction separates these ingredients: Λ handles the Jacobian inverse while $\partial_\eta m(W_1, \theta, \eta)$ enters stochastically through the independent copy W_1 . This separation enables the natural extension to $q \geq 2$.

4.2 Nonlinear g : general moment conditions

When the moment function $g(W_i, \theta, \eta)$ is nonlinear in η , additional correction terms are needed beyond the affine orthogonal moment function. The target m was already allowed to be nonlinear in the previous subsection, but we now also allow for nonlinearity of g . We denote the affine orthogonal moment function from (17) by $\psi_{\text{aff}}^{(q)}$ and write the corrected orthogonal moment function as

$$\psi^{(q)}(W, \theta, \eta, \lambda) = \psi_{\text{aff}}^{(q)}(W, \theta, \eta, \lambda) + \text{correction terms.} \quad (20)$$

The reason $\psi_{\text{aff}}^{(q)}$ fails to be orthogonal for nonlinear g is that the proof of Theorem 1 relies on $\Lambda_0 \mathbb{E}[g(W_i, \theta_0, \eta_0 + \delta)]$ being linear in δ . When g is nonlinear, this expression contains quadratic and higher-order terms in δ that the affine construction does not cancel. The correction terms involve the higher-order derivatives $\partial_\eta^p g(W_i, \theta, \eta)$ for $p \geq 2$. For $q = 1$ no correction is needed: $\psi^{(1)} = \psi_{\text{aff}}^{(1)}$ is first-order Neyman-orthogonal even when g is nonlinear.

Before turning to the construction, we set up notation used throughout this subsection. For $p \geq 1$, the p -th derivative tensor $\partial_\eta^p g(W_i, \theta, \eta) \in \mathbb{R}^{d_g} \otimes (\mathbb{R}^{d_\eta})^{\otimes p}$ has one d_g -output slot and p input slots in \mathbb{R}^{d_η} , with entries $[\partial_\eta^p g]_{k, j_1, \dots, j_p} = \partial^p g_k(W_i, \theta, \eta) / \partial \eta_{j_1} \cdots \partial \eta_{j_p}$. For $p = 1$ this is the $d_g \times d_\eta$ Jacobian. To express contractions of such tensors with vectors in \mathbb{R}^{d_η} , we

extend the angle-bracket notation of Section 4.1 as follows: for a tensor A with p input slots in \mathbb{R}^{d_η} and output in a vector space V , and for $v_1, \dots, v_p \in \mathbb{R}^{d_\eta}$,

$$A[v_1, \dots, v_p] \in V$$

denotes the contraction of A along its inputs, leaving the V -output unchanged. For the scalar function m , the bracket and the angle-bracket of Section 4.1 agree, $\partial_\eta^r m[v_1, \dots, v_r] = \langle \partial_\eta^r m, v_1 \otimes \dots \otimes v_r \rangle \in \mathbb{R}$. For the vector-valued g , $\partial_\eta^p g[v_1, \dots, v_p] \in \mathbb{R}^{d_g}$, and pre-multiplying by $\Lambda(\lambda)$ contracts the d_g -output slot to give a vector in \mathbb{R}^{d_η} .

4.2.1 Second-order orthogonality via reparameterization

The orthogonal moment function for the general nonlinear case can be derived from the affine formula via a change of nuisance parameter $\phi = f(\eta)$, and the goal of this subsection is to show that derivation for $q = 2$. Let $f : \mathbb{R}^{d_\eta} \rightarrow \mathbb{R}^{d_\eta}$ be a parameter transformation, define $\bar{g}(W_i, \theta, \phi) := g(W_i, \theta, f^{-1}(\phi))$.

We want to apply the affine moment function construction to \bar{g} , but it is important to note that \bar{g} being affine in ϕ is not a necessary condition for that construction to be applicable. Inspecting the proof of Theorem 1 in the appendix, one sees that all that is actually needed at $q = 2$ is the condition

$$\Lambda_0 \mathbb{E} \left[\partial_\phi^2 \bar{g}(W_i, \theta_0, \phi_0) \right] = 0, \quad (21)$$

where $\partial_\phi^2 \bar{g}(W_i, \theta, \phi) \in \mathbb{R}^{d_g} \otimes (\mathbb{R}^{d_\eta})^{\otimes 2}$ follows the same convention as $\partial_\eta^p g$ above, with η replaced by ϕ and g by \bar{g} . Pre-multiplication by Λ_0 contracts the d_g -output slot, and (21) requires the resulting $d_\eta \times d_\eta \times d_\eta$ tensor to vanish. Define $T_2 := \Lambda_0 \mathbb{E} \left[\partial_\eta^2 g(W_i, \theta_0, \eta_0) \right] \in \mathbb{R}^{d_\eta \times d_\eta \times d_\eta}$. Then a simple explicit choice for f that guarantees (21) is given by

$$\phi = f(\eta) := \eta + \frac{1}{2} T_2 [\eta - \eta_0, \eta - \eta_0]. \quad (22)$$

Because (21) holds for this choice of f ,⁴ the affine $q = 2$ moment function from (17), with g replaced by \bar{g} and η replaced by ϕ , is second-order Neyman-orthogonal at $(\theta_0, \phi_0, \lambda_0)$. Defining $\bar{m}(W_i, \theta, \phi) := m(W_i, \theta, f^{-1}(\phi))$ analogously to \bar{g} , this orthogonal moment function

⁴From (22) we have $\phi_0 = \eta_0$ and $\partial_\eta f(\eta_0) = I_{d_\eta}$, so by the inverse function theorem f^{-1} exists locally with $\partial_\phi f^{-1}(\phi_0) = I_{d_\eta}$ and $\partial_\phi^2 f^{-1}(\phi_0) = -T_2$. The chain rule applied to $\bar{g}(W_i, \theta_0, \phi) = g(W_i, \theta_0, f^{-1}(\phi))$ at ϕ_0 then gives $\partial_\phi^2 \bar{g}(W_i, \theta_0, \phi_0) = \partial_\eta^2 g(W_i, \theta_0, \eta_0) + \partial_\eta g(W_i, \theta_0, \eta_0) \cdot (-T_2)$, and pre-multiplying by Λ_0 and taking expectations yields $T_2 - T_2 = 0$.

reads, evaluated at $\phi = f(\eta)$,

$$\begin{aligned}\bar{\psi}^{(2)}(W, \theta, \eta, \lambda) &= \bar{m}(W_1, \theta, \phi) - 2 \partial_\phi \bar{m}(W_1, \theta, \phi)' \Lambda(\lambda) \bar{g}(W_2, \theta, \phi) \\ &\quad + \partial_\phi \bar{m}(W_1, \theta, \phi)' \left(\Lambda(\lambda) \partial_\phi \bar{g}(W_3, \theta, \phi) \right) \Lambda(\lambda) \bar{g}(W_2, \theta, \phi) \\ &\quad + \frac{1}{2} \left(\Lambda(\lambda) \bar{g}(W_2, \theta, \phi) \right)' \partial_{\phi\phi}^2 \bar{m}(W_1, \theta, \phi) \left(\Lambda(\lambda) \bar{g}(W_3, \theta, \phi) \right).\end{aligned}$$

By construction $\bar{\psi}^{(2)}$ is second-order Neyman-orthogonal at $(\theta_0, \eta_0, \lambda_0)$, but its expression in η -coordinates still involves \bar{m}, \bar{g} and their ϕ -derivatives, which is not yet a closed-form expression in η . We obtain such an expression by chain-rule expansion of each \bar{m}, \bar{g} derivative at $\phi = f(\eta_0) = \eta_0$, using $\partial_\phi f^{-1}(\phi_0) = I_{d_\eta}$ and $\partial_\phi^2 f^{-1}(\phi_0) = -T_2$. The first three terms of $\bar{\psi}^{(2)}$ reduce to the corresponding terms of $\psi_{\text{aff}}^{(2)}$, and the second-order chain rule for \bar{m} produces the identity

$$\partial_{\phi\phi}^2 \bar{m}(W_1, \theta_0, \phi_0) = \partial_{\eta\eta}^2 m(W_1, \theta_0, \eta_0) - \partial_\eta m(W_1, \theta_0, \eta_0) \cdot T_2,$$

which splits the fourth term of $\bar{\psi}^{(2)}$ into the fourth term of $\psi_{\text{aff}}^{(2)}$ plus a T_2 -correction. The result is that the value of $\bar{\psi}^{(2)}$ at the truth $(\theta_0, \eta_0, \lambda_0)$, and its η - and λ -derivatives at the truth up to total order two, coincide with those of the moment function⁵

$$\begin{aligned}\psi^{(2)}(W, \theta, \eta, \lambda) &= \psi_{\text{aff}}^{(2)}(W_1, W_2, W_3, \theta, \eta, \lambda) \\ &\quad - \frac{1}{2} \partial_\eta m(W_1, \theta, \eta)' T_2 \left[\Lambda(\lambda) g(W_2, \theta, \eta), \Lambda(\lambda) g(W_3, \theta, \eta) \right].\end{aligned}\quad (23)$$

However, the expression for $\psi^{(2)}(W, \theta, \eta, \lambda)$ in the last display is not yet a usable moment function, because T_2 is unknown. To make it feasible we replace $T_2 = \Lambda_0 \mathbb{E}[\partial_\eta^2 g(W_4, \theta_0, \eta_0)]$ by $\Lambda(\lambda) \partial_\eta^2 g(W_4, \theta, \eta)$, where W_4 is an independent copy of the data. This delivers a moment function that uses $L = 4$ independent copies of W_i and that by construction is second-order Neyman-orthogonal:

$$\begin{aligned}\psi^{(2)}(W_1, W_2, W_3, W_4, \theta, \eta, \lambda) &= \psi_{\text{aff}}^{(2)}(W_1, W_2, W_3, \theta, \eta, \lambda) \\ &\quad - \frac{1}{2} \partial_\eta m(W_1, \theta, \eta)' \Lambda(\lambda) \partial_\eta^2 g(W_2, \theta, \eta) \left[\Lambda(\lambda) g(W_3, \theta, \eta), \Lambda(\lambda) g(W_4, \theta, \eta) \right] \\ &= m(W_1, \theta, \eta) - 2 \partial_\eta m(W_1, \theta, \eta)' \Lambda(\lambda) g(W_2, \theta, \eta) \\ &\quad + \partial_\eta m(W_1, \theta, \eta)' \left(\Lambda(\lambda) \partial_\eta g(W_3, \theta, \eta) \right) \Lambda(\lambda) g(W_2, \theta, \eta)\end{aligned}$$

⁵The two functions $\bar{\psi}^{(2)}$ and $\psi^{(2)}$ are not equal as functions of (η, λ) — the chain-rule identities used to derive (23) hold only at $\eta = \eta_0$, since $\partial_\phi f^{-1}(f(\eta))$ depends on η for $\eta \neq \eta_0$ — but they have matching Taylor expansions at the truth up to order two, which is what orthogonality requires.

$$\begin{aligned}
& + \frac{1}{2} \left(\Lambda(\lambda) g(W_2, \theta, \eta) \right)' \partial_{\eta\eta}^2 m(W_1, \theta, \eta) \left(\Lambda(\lambda) g(W_3, \theta, \eta) \right) \\
& - \frac{1}{2} \partial_{\eta} m(W_1, \theta, \eta)' \Lambda(\lambda) \partial_{\eta}^2 g(W_2, \theta, \eta) \left[\Lambda(\lambda) g(W_3, \theta, \eta), \Lambda(\lambda) g(W_4, \theta, \eta) \right], \quad (24)
\end{aligned}$$

where we swapped the labeling of W_1, \dots, W_4 in the final term. The replacement of T_2 by $\Lambda(\lambda) \partial_{\eta}^2 g(W_i, \theta, \eta)$ introduces additional dependence on (η, λ) that was not present in the ϕ -coordinate derivation, but this does not affect second-order Neyman-orthogonality.

4.2.2 Useful notation: indexing terms in $\psi^{(q)}$ by rooted trees

We now introduce notation that turns out to be useful for expressing $\psi^{(q)}$ at general q . We motivate it by rewriting $\psi^{(2)}$ from (24). Figure 1 shows how the five terms of $\psi^{(2)}$ can each be represented by a rooted tree built from three kinds of nodes. The *root* (shaded blue) carries either m or one of its η -derivatives $\partial_{\eta}^r m$, where r is the number of children of the root. The *leaves* (shaded gray) each carry Λg . The *non-root, non-leaf nodes* (unshaded) each carry $\Lambda \partial_{\eta}^p g$, where p is the number of children of the node. Each node is evaluated on its own independent copy of W_i . The tree, which we denote by τ , then uniquely determines the corresponding term in $\psi^{(2)}$, which we denote by $\kappa_{\tau}(W, \theta, \eta, \lambda)$.

It turns out that all terms in the orthogonal moment function $\psi^{(q)}$ can be indexed by rooted trees in the same way for any order $q \in \{1, 2, 3, \dots\}$,

$$\psi^{(q)}(W, \theta, \eta, \lambda) = \sum_{\tau \in \mathcal{T}_q} c_{q,\tau} \kappa_{\tau}(W, \theta, \eta, \lambda), \quad (25)$$

where \mathcal{T}_q is a finite set of rooted trees, $c_{q,\tau} \in \mathbb{R}$ is a combinatorial coefficient, and the kernel $\kappa_{\tau}(W, \theta, \eta, \lambda)$ is determined by τ via the rules illustrated in Figure 1.

A *rooted tree* τ is a finite tree (a connected acyclic graph) with one node singled out as the *root*. Let $d(\tau)$ denote the number of non-root nodes of τ with at most one child. The middle column of Figure 1 reports $d(\tau)$ for each of the five trees: the trivial tree has $d = 0$, the root with one leaf child has $d = 1$, and the remaining three trees all have $d = 2$. The set \mathcal{T}_q in (25) is then

$$\mathcal{T}_q := \left\{ \tau \text{ rooted tree} : d(\tau) \leq q \right\}. \quad (26)$$

For $q = 0$, \mathcal{T}_0 only contains the single-node tree. For $q = 1$, \mathcal{T}_1 contains the single-node tree and the tree with one root and one leaf child. For $q = 2$, the five elements of \mathcal{T}_2 are shown in Figure 1. For $q = 3$ we have $|\mathcal{T}_3| = 13$, shown in Figure 2. The cardinalities grow rapidly, with $|\mathcal{T}_4| = 40$ and $|\mathcal{T}_5| = 130$.

Each rooted tree $\tau \in \mathcal{T}_q$ translates into a kernel $\kappa_{\tau}(W, \theta, \eta, \lambda)$ by populating the nodes with the algebraic objects already used in $\psi^{(2)}$. The root carries m or one of its η -derivatives,





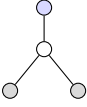
Rooted tree τ	$d(\tau)$	τ with factors	$\kappa_\tau(W, \theta, \eta, \lambda)$
	0	$m(W_1)$	$m(W_1, \theta, \eta)$
	1	$\partial_\eta m(W_1)$ $\Lambda g(W_2)$	$\partial_\eta m(W_1, \theta, \eta)' \Lambda(\lambda) g(W_2, \theta, \eta)$
	2	$\partial_\eta m(W_1)$ $\Lambda \partial_\eta g(W_3)$ $\Lambda g(W_2)$	$\partial_\eta m(W_1, \theta, \eta)' (\Lambda(\lambda) \partial_\eta g(W_3, \theta, \eta)) \Lambda(\lambda) g(W_2, \theta, \eta)$
	2	$\partial_{\eta\eta}^2 m(W_1)$ $\Lambda g(W_2)$ $\Lambda g(W_3)$	$(\Lambda(\lambda) g(W_2, \theta, \eta))' \partial_{\eta\eta}^2 m(W_1, \theta, \eta) (\Lambda(\lambda) g(W_3, \theta, \eta))$
	2	$\partial_\eta m(W_1)$ $\Lambda \partial_\eta^2 g(W_2)$ $\Lambda g(W_3)$ $\Lambda g(W_4)$	$\partial_\eta m(W_1, \theta, \eta)' \times \Lambda(\lambda) \partial_\eta^2 g(W_2, \theta, \eta) [\Lambda(\lambda) g(W_3, \theta, \eta), \Lambda(\lambda) g(W_4, \theta, \eta)]$

Figure 1: The five terms of $\psi^{(2)}$ in (24), indexed by rooted trees. The root (blue) carries m or its η -derivatives. Non-root non-leaf nodes carry $\Lambda \partial_\eta^p g$ where p is the node's number of children. Leaves (gray) carry Λg .

with the order of differentiation equal to the number of children of the root. Each leaf carries Λg . Each non-root, non-leaf node carries $\Lambda \partial_\eta^p g$, where p is the number of children of the node. Each node is evaluated on its own independent copy of W_i , and the resulting expression is the contraction along the parent–child edges of the tree, exactly as in Figure 1. Applying this construction to the seven trees in the top row of Figure 2 gives the affine moment function $\psi_{\text{aff}}^{(3)}$, where for compactness we suppress the arguments (θ, η) in m and g and the argument λ in Λ :

$$\begin{aligned}
\psi_{\text{aff}}^{(3)} &= m(W_1) - 3 \partial_\eta m(W_1)' \Lambda g(W_2) \\
&\quad + 3 \partial_\eta m(W_1)' (\Lambda \partial_\eta g(W_3)) \Lambda g(W_2) \\
&\quad - \partial_\eta m(W_1)' (\Lambda \partial_\eta g(W_3)) (\Lambda \partial_\eta g(W_4)) \Lambda g(W_2)
\end{aligned}$$

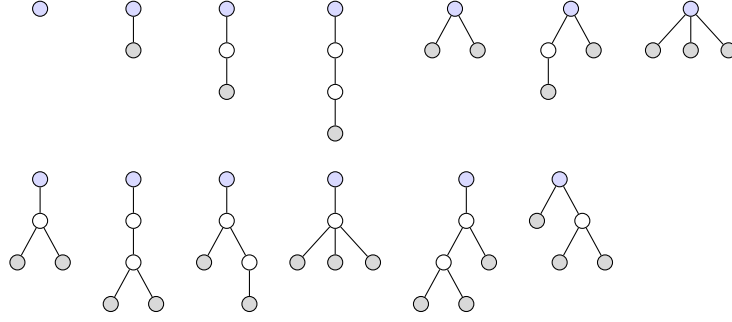


Figure 2: The thirteen elements of \mathcal{T}_3 . The first row shows the seven trees in which every non-root node has at most one child — these correspond to the terms of the affine moment function $\psi_{\text{aff}}^{(3)}$. The second row shows the six correction trees, each containing at least one non-root node with two or more children.

$$\begin{aligned}
& + \frac{3}{2} \partial_\eta^2 m(W_1) \left[\Lambda g(W_2), \Lambda g(W_3) \right] \\
& - \partial_\eta^2 m(W_1) \left[\Lambda g(W_2), \left(\Lambda \partial_\eta g(W_4) \right) \Lambda g(W_3) \right] \\
& - \frac{1}{6} \partial_\eta^3 m(W_1) \left[\Lambda g(W_2), \Lambda g(W_3), \Lambda g(W_4) \right], \tag{27}
\end{aligned}$$

while the six trees in the bottom row of Figure 2 correspond to the additional correction terms that are needed when g is nonlinear in η ,

$$\begin{aligned}
\psi^{(3)} & = \psi_{\text{aff}}^{(3)} \\
& - 2 \partial_\eta m(W_1)' \Lambda \partial_\eta^2 g(W_2) \left[\Lambda g(W_3), \Lambda g(W_4) \right] \\
& + \frac{1}{2} \partial_\eta m(W_1)' \Lambda \partial_\eta g(W_2) \Lambda \partial_\eta^2 g(W_3) \left[\Lambda g(W_4), \Lambda g(W_5) \right] \\
& + \partial_\eta m(W_1)' \Lambda \partial_\eta^2 g(W_2) \left[\Lambda \partial_\eta g(W_3) \Lambda g(W_4), \Lambda g(W_5) \right] \\
& + \frac{1}{6} \partial_\eta m(W_1)' \Lambda \partial_\eta^3 g(W_2) \left[\Lambda g(W_3), \Lambda g(W_4), \Lambda g(W_5) \right] \\
& - \frac{1}{2} \partial_\eta m(W_1)' \Lambda \partial_\eta^2 g(W_2) \left[\Lambda \partial_\eta^2 g(W_3) \left[\Lambda g(W_4), \Lambda g(W_5) \right], \Lambda g(W_6) \right] \\
& + \frac{1}{2} \partial_\eta^2 m(W_1) \left[\Lambda g(W_2), \Lambda \partial_\eta^2 g(W_3) \left[\Lambda g(W_4), \Lambda g(W_5) \right] \right]. \tag{28}
\end{aligned}$$

The combinatorial pre-factors in (27) and (28) are exactly the values needed for third-order Neyman orthogonality. The affine pre-factors can be obtained from (17) by collecting the ordered chain-length tuples (k_1, \dots, k_r) that correspond to the same unordered tree. The correction term pre-factors are the values of the closed-form $c_{q,\tau}$ that we introduce next. Note that the substantive content of the tree formalism is our closed form expression for $c_{q,\tau}$, while the index set \mathcal{T}_q and the construction rule for κ_τ are purely notational bookkeeping tools.

4.2.3 Closed form for $c_{q,\tau}$ and the main result

To state the closed form for $c_{q,\tau}$, we first need to define two further integers associated with each rooted tree τ . We have already introduced $d(\tau)$ as the number of non-root nodes of τ with at most one child. Let $|\tau|$ denote the total number of non-root nodes of τ (equivalently, the number of edges of τ). Let $\text{Aut}(\tau)$ denote the automorphism group of τ , that is, the set of bijections of the nodes of τ that fix the root and keep the tree structure unchanged. Finally, let $|\text{Aut}(\tau)|$ be its order (i.e. the cardinality of $\text{Aut}(\tau)$). Then,

$$c_{q,\tau} := \frac{(-1)^{|\tau|}}{|\text{Aut}(\tau)|} \binom{q + |\tau| - d(\tau)}{|\tau|}. \quad (29)$$

Figure 3 shows the values of $|\tau|$, $d(\tau)$, and $|\text{Aut}(\tau)|$ for all rooted trees τ with $d(\tau) \leq 3$. The values of $|\tau|$ and $d(\tau)$ are read off by counting the relevant nodes. The values of $|\text{Aut}(\tau)|$ in this figure are also straightforward: for many of the trees, only the trivial bijection preserves the tree structure, so $|\text{Aut}(\tau)| = 1$. For each of the remaining trees, exactly one node has c isomorphic leaf children, which can be permuted in $c!$ ways, giving $|\text{Aut}(\tau)| = c!$. For larger values of $d(\tau)$ the automorphism group can of course be more complicated.

As a check, applying (29) at $q = 3$ to the values in Figure 3 reproduces the 13 coefficients in (27) and (28). For example, the root with three leaf children gives $c_{3,\tau} = -\binom{3}{3}/3! = -1/6$ (matching the coefficient of $\partial_\eta^3 m[\Lambda g, \Lambda g, \Lambda g]$ in (27)), etc.⁶

Tree-indexed expansions of this form, including the appearance of the symmetry factor $1/|\text{Aut}(\tau)|$, are a well-known structure in the literature on multivariate Taylor expansions, see Appendix A.1 for a discussion of the Butcher B -series. That appendix also gives a simple recursion for computing $|\text{Aut}(\tau)|$.⁷

⁶It is also instructive to verify the consistency between (29) and the affine moment function in (17). The trees that contribute to $\psi_{\text{aff}}^{(q)}$ are precisely those in which only the root has more than one child, equivalently those with $|\tau| = d(\tau)$. For such a tree, let $r \geq 0$ be the number of children of the root, and let $k_s \geq 1$ be the chain length of each child for $s \in \{1, \dots, r\}$. We then have $|\tau| = d(\tau) = k_1 + \dots + k_r$. The automorphism group permutes branches of equal length, so if the multiset $\{k_1, \dots, k_r\}$ has n_ℓ elements equal to ℓ , then $|\text{Aut}(\tau)| = \prod_\ell n_\ell!$, and (29) simplifies to

$$c_{q,\tau} = \frac{(-1)^{|\tau|}}{\prod_\ell n_\ell!} \binom{q}{|\tau|}. \quad (30)$$

The same unordered tree τ is the image of $r!/\prod_\ell n_\ell!$ ordered branch tuples (k_1, \dots, k_r) , each contributing the term $\frac{(-1)^{|\tau|}}{r!} \binom{q}{|\tau|} \kappa_\tau$ to the affine moment function $\psi_{\text{aff}}^{(q)}$ in (17). Summing across these tuples cancels the $r!$ and reproduces exactly $c_{q,\tau} \kappa_\tau$ with $c_{q,\tau}$ as in (30), confirming that $\psi_{\text{aff}}^{(q)}$ in (17) matches the affine part of (25).

⁷Using this recursion we have numerically verified that (29) delivers orthogonal moment functions for q up to 10 in the scalar case $d_\eta = d_g = 1$, and for q up to 8 in the multivariate case. Enumerating all $|\mathcal{T}_{10}| = 110, 135$ trees and calculating all corresponding $c_{q,\tau}$ is possible within seconds.














Rooted tree τ	$ \tau $	$d(\tau)$	$ \text{Aut}(\tau) $
	0	0	1
	1	1	1
	2	2	1
	2	2	2
	3	2	2
	3	3	1
	3	3	1
	3	3	6
	4	3	2
	4	3	1
	4	3	6
	4	3	2
	5	3	2

Figure 3: The integers $|\tau|$, $d(\tau)$, and $|\text{Aut}(\tau)|$ for the 13 trees with $d(\tau) \leq 3$.

With $c_{q,\tau}$ in hand, we can state our main result for the nonlinear case.

Assumption 4.

- (i) Assumption 3(ii)–(iv) hold.
- (ii) $g(W_i, \theta, \eta)$ is $(q + 1)$ -times continuously differentiable in η in a neighborhood of η_0 .
- (iii) For all $p \leq q + 1$, there is a neighborhood \mathcal{N} of η_0 and an integrable function $M_p(W_i)$ with $\sup_{\eta \in \mathcal{N}} \|\partial_\eta^p g(W_i, \theta_0, \eta)\| \leq M_p(W_i)$ almost surely.

Theorem 2 (q -th order Neyman orthogonality, nonlinear case). *Let $q \geq 1$ be an integer, and suppose Assumptions 1, 2, and 4 hold. Then the moment function $\psi^{(q)}(W, \theta, \eta, \lambda)$ in (25) with $c_{q,\tau}$ given by (29) is q -th order Neyman-orthogonal in the sense of Definition 1. That is, $\Psi^{(q)}(\theta_0, \eta_0, \lambda_0) = 0$, and, for all multi-indices (α, β) with $1 \leq |\alpha| + |\beta| \leq q$,*

$$\partial_\eta^\alpha \partial_\lambda^\beta \Psi^{(q)}(\theta_0, \eta_0, \lambda_0) = 0.$$

The proof is in Appendix B.2.

The expression (29) applies to both the affine and correction parts of $\psi^{(q)}$. Trees with $|\tau| = d(\tau)$ are exactly those in which every non-root node is either a leaf or a one-child node; for these, (29) reduces to (30), the binomial weights of the affine score (17). Trees with $|\tau| > d(\tau)$ — those with at least one non-root node with at least two children — are the correction trees, which contribute terms involving the higher-order derivatives $\partial_\eta^p g$ for $p \geq 2$. When g is affine in η , all such kernels vanish, (25) reduces to $\psi^{(q)} = \psi_{\text{aff}}^{(q)}$, and Theorem 2 recovers Theorem 1.

The total number of independent copies of W_i used by $\psi^{(q)}$ equals one plus the maximum number of nodes in any tree $\tau \in \mathcal{T}_q$, which is bounded by $L \leq 2q$ for $q \geq 2$ (compared with $L = 1 + q$ in the affine case). The bound is attained by the binary tree with $d(\tau) = q$ in which every non-root, non-leaf node has exactly two children.

4.3 Reducing the nuisance dimension

To reduce the dimension of the nuisance parameter λ , the key observation is that the moment function $g(W_i, \theta, \eta)$ in (1) can be replaced by a different function $\tilde{g}(W, \theta, \eta)$ that may depend on multiple independent copies of W_i .⁸ The replacement is valid as long as $\mathbb{E}[\tilde{g}(W, \theta_0, \eta_0)] = 0$ and the corresponding Jacobian $\tilde{J}_0 = \mathbb{E}[\partial_\eta \tilde{g}(W, \theta_0, \eta_0)]$ has a left inverse Λ_0 with $\Lambda_0 \tilde{J}_0 = I_{d_\eta}$. The orthogonality construction in Sections 4.1–4.2 applies with g replaced by \tilde{g} throughout.

⁸As before, here we write W without subscript when the argument requires several copies.

The advantage is that \tilde{g} can be designed so that the nuisance parameter λ has much lower dimension.

There is, however, a trade-off: reducing d_λ comes at the cost of increasing the number of independent copies L needed by the orthogonal moment function. The original moment function $g(W_i, \theta, \eta)$ depends on a single observation, so the q -th order affine orthogonal moment function uses $L = 1+q$ copies. The transformed function $\tilde{g}(W, \theta, \eta)$ may itself require multiple copies (e.g., d_η copies for the determinant construction below), and each appearance of \tilde{g} in the orthogonal moment function then uses that many copies, increasing L accordingly. In practice, the choice of \tilde{g} balances the difficulty of estimating a high-dimensional nuisance λ (large d_λ , small L) against the variance cost of a higher-order U-statistic (small d_λ , large L).

In Sections 4.1 and 4.2, the orthogonal moment functions were stated in terms of a moment function g that depends on a single observation W_j . When \tilde{g} requires a block of several independent copies, the construction is applied in exactly the same way, with one bookkeeping change: each occurrence of $g(W_j, \theta, \eta)$ or $\partial_\eta g(W_j, \theta, \eta)$ in the formulas of those sections is replaced by \tilde{g} or $\partial_\eta \tilde{g}$ evaluated on a fresh block of independent copies of the size required by \tilde{g} . We give examples of constructions for \tilde{g} below.

Determinant construction (exactly identified case). Suppose $d_g = d_\eta$ and J_0 is invertible. Define $\tilde{g}(W_1, \dots, W_{d_\eta}; \theta, \eta)$ componentwise by

$$\tilde{g}_r(W_1, \dots, W_{d_\eta}; \theta, \eta) = \det \left[\begin{array}{c} \partial_{\eta_1} g(W_1, \theta, \eta), \dots, \partial_{\eta_{r-1}} g(W_{r-1}, \theta, \eta), g(W_r, \theta, \eta), \\ \partial_{\eta_{r+1}} g(W_{r+1}, \theta, \eta), \dots, \partial_{\eta_{d_\eta}} g(W_{d_\eta}, \theta, \eta) \end{array} \right],$$

where $r = 1, \dots, d_\eta$, and the $d_\eta \times d_\eta$ matrix inside the determinant has $g(W_r, \theta, \eta)$ in its r -th column and $\partial_{\eta_j} g(W_j, \theta, \eta)$ in its j -th column for $j \neq r$. Then $\mathbb{E}[\tilde{g}_r] = 0$ follows from $\mathbb{E}[g(W_i, \theta_0, \eta_0)] = 0$ and the multilinearity of the determinant. The Jacobian satisfies $\mathbb{E}[\partial_\eta \tilde{g}(W, \theta_0, \eta_0)] = \det(J_0) \cdot I_{d_\eta}$, so $\Lambda(\lambda) = \lambda \cdot I_{d_\eta}$ with

$$\lambda_0 = \frac{1}{\det(J_0)}.$$

This gives $d_\lambda = 1$. The trade-off is that each evaluation of \tilde{g} uses d_η independent copies of W_i , so the total number of copies for the q -th order moment function becomes $L = 1 + q \cdot d_\eta$.

Determinant construction (overidentified case). When $d_g > d_\eta$, the Jacobian J_0 is no longer square, so the construction above does not directly apply. The idea is to first project

the d_g -dimensional vectors g and $\partial_{\eta_j}g$ down to d_η dimensions using independent copies of $\partial_{\eta_j}g$, and then take the determinant. Concretely, define $\tilde{g}(W_1, \dots, W_{2d_\eta}; \theta, \eta)$ componentwise by

$$\tilde{g}_r(W_1, \dots, W_{2d_\eta}; \theta, \eta) = \det \left[(\partial_{\eta_j}g(W_1, \theta, \eta))' \partial_{\eta_1}g(W_2, \theta, \eta), \dots, \right. \\ \left. (\partial_{\eta_j}g(W_{2r-1}, \theta, \eta))' g(W_{2r}, \theta, \eta), \dots, (\partial_{\eta_j}g(W_{2d_\eta-1}, \theta, \eta))' \partial_{\eta_{d_\eta}}g(W_{2d_\eta}, \theta, \eta) \right],$$

where $r = 1, \dots, d_\eta$. Each column of the $d_\eta \times d_\eta$ matrix is formed by pre-multiplying a d_g -vector ($\partial_{\eta_j}g$ or g) by the $d_\eta \times d_g$ matrix $(\partial_{\eta_j}g)'$ from an independent copy, yielding a d_η -vector. The r -th column uses $g(W_{2r}, \theta, \eta)$ (without a ∂_{η_r} derivative), while all other columns use $\partial_{\eta_j}g(W_{2j}, \theta, \eta)$. Then $\mathbb{E}[\tilde{g}_r] = 0$ by the same multilinearity argument as before, the Jacobian satisfies $\mathbb{E}[\partial_{\eta} \tilde{g}(W, \theta_0, \eta_0)] = \det(J'_0 J_0) \cdot I_{d_\eta}$, and

$$\Lambda(\lambda) = \lambda \cdot I_{d_\eta}, \quad \lambda_0 = \frac{1}{\det(J'_0 J_0)},$$

again giving $d_\lambda = 1$. Each evaluation of \tilde{g} now uses $2d_\eta$ independent copies of W_i .

5 Implementation and simulation

In this section we present an implementation of our orthogonal moment functions in the setup of Section 3, together with Monte Carlo simulation results. The simulation design is a stylized version of the setup in [Kline, Rose and Walters \(2022\)](#) (though we do not attempt to calibrate the design to their empirical results). There are N firms, and T job applications per firm. Applications feature the race $X_{it,1} \in \{0, 1\}$ of applicants and their gender $X_{it,2} \in \{0, 1\}$. Binary callback indicators are generated as

$$Y_{it} = \mathbf{1}\{\varepsilon_{it} \geq \beta_{i0} + \beta_{i1}X_{it,1} + \beta_{i2}X_{it,2}\}, \quad (31)$$

where ε_{it} are i.i.d. standard logistic independent of $X_{it,1}, X_{it,2}$, and observations are independent across i and t .

Job applications are randomly generated within firms but – for example due to implementation constraints – assignment is heterogeneous across firms. Specifically, we assume that firms belong to one of two types $Z_i \in \{1, 2\}$, where Z_i are i.i.d. Bernoulli with probability 1/2. When $Z_i = 1$, $(X_{it,1} = 1, X_{it,2} = 1)$ and $(X_{it,1} = 0, X_{it,2} = 0)$ are each drawn with probability 3/8, and $(X_{it,1} = 0, X_{it,2} = 1)$ and $(X_{it,1} = 1, X_{it,2} = 0)$ are each drawn with probability 1/8. When $Z_i = 2$, $(X_{it,1} = 1, X_{it,2} = 1)$ and $(X_{it,1} = 0, X_{it,2} = 0)$ are each

drawn with probability $1/8$, and $(X_{it,1} = 0, X_{it,2} = 1)$ and $(X_{it,1} = 1, X_{it,2} = 0)$ are each drawn with probability $3/8$. For $j \in \{0, 1, 2\}$, $\beta_{ij} = Z_i + V_{ij}$, where the V_{ij} are independent standard normal, independent of Z_i .

Let $X_{it} = (1, X_{it,1}, X_{it,2})'$, and let

$$\eta_{i0} = \left(\mathbb{E} \left[\sum_{t=1}^T X_{it} X_{it}' \right] \right)^{-1} \mathbb{E} \left[\sum_{t=1}^T X_{it} Y_{it} \right].$$

We estimate the parameters

$$\theta_{10} = \frac{1}{N} \sum_{i=1}^N \eta_{i10}, \quad \theta_{20} = \frac{1}{N} \sum_{i=1}^N \eta_{i10}^2,$$

which together are informative about the level and dispersion of race-based employment discrimination (here, the coefficient η_{i10} corresponds to race). The true values in the DGP are $\theta_{10} = -0.0844$, which corresponds to a 8.44 percentage point lower callback probability for blacks compared to whites, and $\theta_{20} = 0.0177$, which corresponds to a standard deviation of 0.1030.

We report results based on 1000 simulations. We compute the OLS estimators as

$$\hat{\eta}_i = \left(\sum_{t=1}^T X_{it} X_{it}' \right)^{-1} \sum_{t=1}^T X_{it} Y_{it},$$

and report the resulting estimates of θ_1 and θ_2 (denoted as OLS). Note that, given the design, $\sum_{t=1}^T X_{it} X_{it}'$ is singular with positive probability. We report averages of $\hat{\eta}_{i1}$ and $\hat{\eta}_{i1}^2$ on the subset of units for which singularity does not occur, and compare those to the averages of η_{i10} and η_{i10}^2 for the same subset of units. We proceed similarly for the other estimators described below.

Next, we compute orthogonal estimators to order 2 (denoted as ORTH). Let, for all $s_1, s_2 \in \{1, \dots, T\}$,

$$\hat{\eta}_{i, -(s_1, s_2)} = \left(\sum_{t=1}^T \mathbf{1}\{t \neq s_1, s_2\} X_{it} X_{it}' \right)^{-1} \sum_{t=1}^T \mathbf{1}\{t \neq s_1, s_2\} X_{it} Y_{it},$$

and

$$\hat{\Lambda}_{i, -(s_1, s_2)} = - \left(\frac{1}{T-2} \sum_{t=1}^T \mathbf{1}\{t \neq s_1, s_2\} X_{it} X_{it}' \right)^{-1}. \quad (32)$$

For θ_1 , the orthogonal estimator is the second element (i.e., the one corresponding to η_{i1})

of

$$\frac{2}{T(T-1)} \sum_{s_1=1}^T \sum_{s_2=s_1+1}^T \left[\widehat{\eta}_{i,-(s_1,s_2)} - \left(2I_3 + \widehat{\Lambda}_{i,-(s_1,s_2)} X_{i,s_2} X'_{i,s_2} \right) \widehat{\Lambda}_{i,-(s_1,s_2)} X_{i,s_1} (Y_{i,s_1} - X'_{i,s_1} \widehat{\eta}_{i,-(s_1,s_2)}) \right],$$

where I_3 is the 3×3 identity matrix, averaged across units. For θ_2 , the orthogonal estimator is, for D the 3×3 matrix with a one in the position (2, 2) and zeroes everywhere else,

$$\begin{aligned} & \frac{2}{T(T-1)} \sum_{s_1=1}^T \sum_{s_2=s_1+1}^T \left[\widehat{\eta}'_{i,-(s_1,s_2)} D \widehat{\eta}_{i,-(s_1,s_2)} \right. \\ & \quad - 2 \widehat{\eta}'_{i,-(s_1,s_2)} D \left(2I_3 + \widehat{\Lambda}_{i,-(s_1,s_2)} X_{i,s_2} X'_{i,s_2} \right) \widehat{\Lambda}_{i,-(s_1,s_2)} X_{i,s_1} (Y_{i,s_1} - X'_{i,s_1} \widehat{\eta}_{i,-(s_1,s_2)}) \\ & \quad \left. + \left(\widehat{\Lambda}_{i,-(s_1,s_2)} X_{i,s_2} (Y_{i,s_2} - X'_{i,s_2} \widehat{\eta}_{i,-(s_1,s_2)}) \right)' D \left(\widehat{\Lambda}_{i,-(s_1,s_2)} X_{i,s_1} (Y_{i,s_1} - X'_{i,s_1} \widehat{\eta}_{i,-(s_1,s_2)}) \right) \right], \end{aligned}$$

also averaged across units.

In addition to estimators based on the plug-in estimates (32) of Λ_i , we also report estimators based on empirical Bayes regularization. Let $p_{ikl} = \Pr(X_{it1} = k, X_{it2} = \ell)$ for all $(k, \ell) \in \{0, 1\}^2$. To regularize Λ_i , we endow $(p_{i00}, p_{i01}, p_{i10}, p_{i11})$ with a Dirichlet prior with parameters $(\alpha_{00}, \alpha_{01}, \alpha_{10}, \alpha_{11})$. Let $\alpha = \alpha_{00} + \alpha_{01} + \alpha_{10} + \alpha_{11}$, $\pi_{kl} = \alpha_{kl}/\alpha$, and

$$\Pi = \begin{pmatrix} 1 & \pi_{10} + \pi_{11} & \pi_{01} + \pi_{11} \\ \pi_{10} + \pi_{11} & \pi_{10} + \pi_{11} & \pi_{11} \\ \pi_{01} + \pi_{11} & \pi_{11} & \pi_{01} + \pi_{11} \end{pmatrix}.$$

We estimate $-\Lambda_i^{-1} = \mathbb{E}[X_{it} X'_{it}]$ as the posterior mean

$$-\widehat{\Lambda}_i^{-1} = \frac{T}{T + \alpha} \left(\frac{1}{T} \sum_{t=1}^T X_{it} X'_{it} \right) + \frac{\alpha}{T + \alpha} \Pi.$$

Finally, to select $(\alpha_{00}, \alpha_{01}, \alpha_{10}, \alpha_{11})$, we maximize the marginal likelihood across firms with respect to α , while fixing $\widehat{\pi}_{kl} = \frac{1}{NT} \sum_i \sum_t \mathbf{1}\{X_{it1} = k, X_{it2} = \ell\}$ to the pooled mean estimator. We estimate α as

$$\widehat{\alpha} = \operatorname{argmax}_{\alpha > 0} N [\ln \Gamma(\alpha) - \ln \Gamma(\alpha + T)] + \sum_{i=1}^N \sum_{(k,\ell) \in \{0,1\}^2} [\ln \Gamma(\widehat{\pi}_{kl} \alpha + n_{ikl}) - \ln \Gamma(\widehat{\pi}_{kl} \alpha)],$$

where $n_{ikl} = \sum_{t=1}^T \mathbf{1}\{X_{it1} = k, X_{it2} = \ell\}$. We apply the regularization scheme to both the OLS and ORTH estimates. In the latter case, we apply the above formulas to subsets of

observations, leaving out two t indices and averaging across pairs of indices *ex post*.

Figures 4 and 5 report the Monte Carlo performance of the OLS (in blue) and ORTH (in red) estimators for θ_1 and θ_2 , respectively, as the number of applications per firm T varies between $T = 20$ and $T = 100$. Each figure plots the mean estimate across simulations (solid line), the true value (dashed horizontal line), and 90% simulation bands defined by the 5th and 95th quantiles across simulations (shaded regions). The plots on the left-hand side show the estimators based on regularized Λ_i 's, while the plots on the right-hand side correspond to un-regularized Λ_i 's.

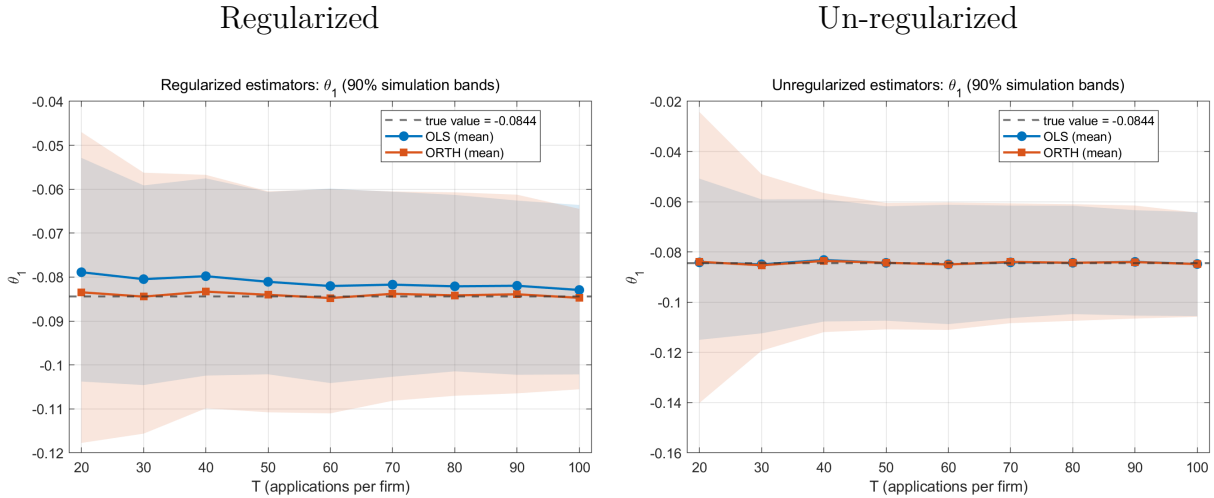


Figure 4: Monte Carlo estimates of θ_1 as a function of T . Solid lines show the mean across simulations for OLS and ORTH; the dashed line is the true value; shaded regions are 90% simulation bands.

Starting with θ_{10} – the average of η_{i10} – the right graph in Figure 4 shows that both the OLS and the second-order orthogonalized estimator are close to unbiased irrespective of T . The left graph, in turn, shows that while OLS is slightly biased as a result of the regularization of Λ_{i0} , the orthogonal estimator remains virtually unbiased. In addition, the confidence bands show that the orthogonalization is associated to some increase in sampling variability.

Shifting attention to θ_{20} – the average of η_{i10}^2 – the right graph in Figure 5 shows that both estimators are biased when one does not rely on a regularized estimator of Λ_{i0} . Moreover, the bias is higher for the orthogonal estimator, and it only becomes lower than the bias of OLS when $T \geq 60$. Lastly, the variance increases associated with the orthogonalization is larger than in the case of θ_{10} , especially for lower values of T . Hence, while orthogonalization reduces bias for large enough T , an estimator based on un-regularized estimates of Λ_{i0} may perform poorly for lower values of T .

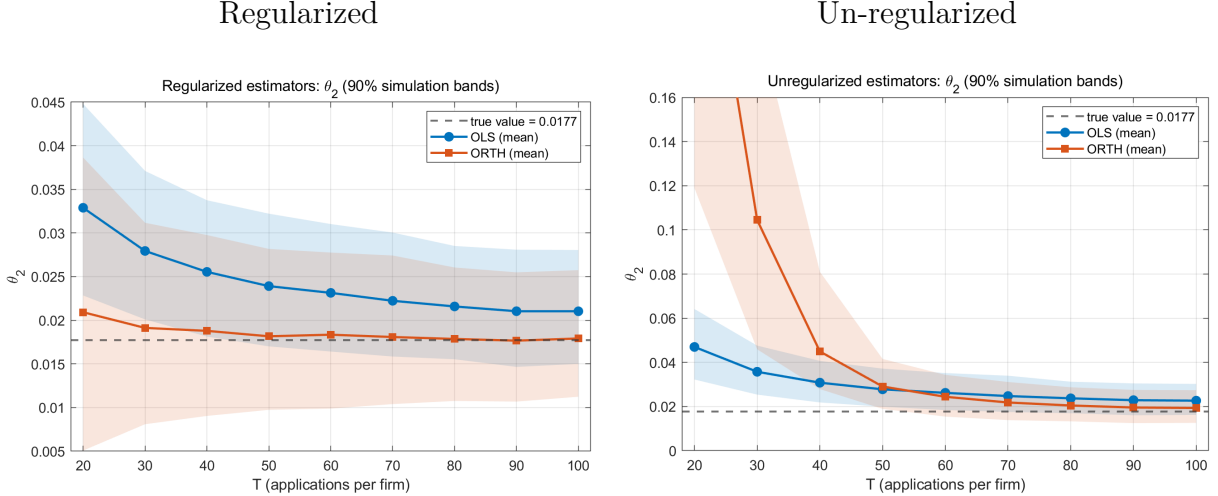


Figure 5: Monte Carlo estimates of θ_2 as a function of T . Solid lines show the mean across simulations for OLS and ORTH; the dashed line is the true value; shaded regions are 90% simulation bands.

The left graph in Figure 5 shows that the situation improves greatly when relying on regularized estimates of Λ_{i0} . The bias of OLS is reduced relative to the un-regularized case, without a noticeable increase in variability. Moreover, the bias of the orthogonal estimator decreases substantially relative to the un-regularized case, and in fact the orthogonal estimator has small bias, lower than the one of OLS for all values of T . This suggests that the combination of orthogonal moments – to reap the benefits of higher-order Neyman orthogonality – and regularization of the nuisance parameters – to ensure lower estimation error – can be particularly appealing in applications.

6 Asymptotic theory

In this section we establish the asymptotic distribution of the estimator based on a q -th order Neyman-orthogonal moment function. The assumptions and result cover grouped data applications in which each unit contributes a within-unit U-statistic over within-group observations, including the heterogeneous-coefficients application of Sections 3 and 5 as a special case. We state the result for a d_θ -dimensional target parameter θ_0 , with an orthogonal moment function ψ taking values in \mathbb{R}^{d_ψ} for some $d_\psi \geq d_\theta$. Under the assumptions below, the estimator is \sqrt{N} -consistent and asymptotically normal, where N denotes the number of units (groups), even when the preliminary nuisance estimates converge at a rate slower than \sqrt{N} .

The analysis uses the unit-by-unit view introduced in Section 3. For each unit $i \in$

$\{1, \dots, N\}$, the T observations W_{i1}, \dots, W_{iT} form the sample on which the orthogonal moment function $\psi(W_i; \theta, \nu_i)$ is constructed, and preliminary estimation of the unit-specific nuisance $\nu_{i0} = (\eta_{i0}, \lambda_{i0})$ is carried out on a held-out subset of observations for the same unit. The cross-sectional dimension $i = 1, \dots, N$ enters only at the final step, through the sample average that defines the estimator of θ_0 . As a consequence, the \sqrt{N} -rate and the asymptotic normality in Theorem 3 below are driven by a cross-sectional central limit theorem. The within-unit structure of ψ affects only the asymptotic variance.

The analysis closely parallels Section 6 of [Bonhomme, Jochmans and Weidner \(2025\)](#). The main adaptation is that the nuisance parameter η_i in that paper is replaced by the pair $\nu_i = (\eta_i, \lambda_i)$, and that the moment function is our orthogonal U-statistic rather than a conditional-likelihood score. Because the limiting distribution is driven by the cross-sectional average, with the within-unit structure only entering through the variance of the moment function, the proof of [Bonhomme, Jochmans and Weidner \(2025\)](#) applies after a straightforward notational translation. We record this translation in Appendix B.3 rather than reproducing the full proof.

6.1 Setup

We observe an independent cross-section of N units. Each unit contributes a bundle $W_i = (W_{i1}, \dots, W_{iT})$ of T observations. Both N and T grow as $N \rightarrow \infty$; we suppress the dependence of T on N in the notation. For each unit i , let $\nu_{i0} = (\eta_{i0}, \lambda_{i0})$ denote the true value of the combined nuisance parameter, where λ_{i0} is the finite-dimensional parameter encoding the unit-level Jacobian inverse Λ_{i0} as in Section 2. The dimension $d_{\nu,i}$ is bounded uniformly in i and N .

The target parameter $\theta_0 \in \mathbb{R}^{d_\theta}$ is identified by

$$\frac{1}{N} \sum_{i=1}^N \mathbb{E}[\psi(W_i; \theta_0, \nu_{i0})] = 0, \quad (33)$$

where $\psi(W_i; \theta, \nu_i) \in \mathbb{R}^{d_\psi}$ is a q -th order Neyman-orthogonal moment function with respect to ν_i in the sense of Definition 1. We allow $d_\psi \geq d_\theta$ to permit overidentification. In applications such as the heterogeneous coefficients example of Section 3, $\theta_0 = (1/N) \sum_i m(\eta_{i0})$ is a sample-dependent parameter; the analysis below covers this case, with the understanding that θ_0 may depend on N .

Let $\hat{\nu}_i$ be a preliminary estimator of ν_{i0} . We assume throughout that $\hat{\nu}_i$ is constructed from a held-out subset of observations, independent of the observations used to evaluate ψ . Concretely, for each unit i we partition the within-unit index $\{1, \dots, T\}$ into disjoint

subsets \mathcal{S}_1 and \mathcal{S}_2 , estimate $\hat{\nu}_i$ from $\{W_{it} : t \in \mathcal{S}_1\}$, and evaluate ψ using observations in \mathcal{S}_2 . The sample is split within each unit, not across units. Efficiency can be improved by cross-fitting, by swapping the roles of \mathcal{S}_1 and \mathcal{S}_2 and averaging the resulting moment functions. The analysis below applies to either a single split or the cross-fitted estimator.

The estimator $\hat{\theta}$ is defined by the GMM problem

$$\hat{\theta} = \operatorname{argmin}_{\theta \in \Theta} \left\| \frac{1}{N} \sum_{i=1}^N \psi(W_i; \theta, \hat{\nu}_i) \right\|_{\Omega}^2, \quad (34)$$

where $\Theta \subseteq \mathbb{R}^{d_\theta}$ is the parameter space, Ω is a symmetric positive-definite $d_\psi \times d_\psi$ weight matrix, and $\|x\|_{\Omega}^2 = x' \Omega x$. When $d_\psi = d_\theta$ (just-identification), $\hat{\theta}$ solves the moment equation $(1/N) \sum_i \psi(W_i; \hat{\theta}, \hat{\nu}_i) = 0$.

6.2 Assumptions and main result

Assumption 5.

(i) As $N \rightarrow \infty$, $(\hat{\theta}, \hat{\nu}_1, \dots, \hat{\nu}_N)$ is contained in a convex neighborhood \mathcal{B}_N of $(\theta_0, \nu_{10}, \dots, \nu_{N0})$. Let $\mathcal{B}_{N,i}$ denote the intersection of \mathcal{B}_N with the parameter subspace for unit i .

(ii) $\max_i d_{\nu,i} = O(1)$.

(iii) Each component of the moment function $\psi(W_i; \theta, \nu_i)$ is $(q + 1)$ times continuously differentiable in (θ, ν_i) , and all its partial derivatives up to order $(q + 1)$ are bounded in absolute value (componentwise) by $C_{N,i}(W_i) \geq 0$ uniformly over $\mathcal{B}_{N,i}$, with

$$\frac{1}{N} \sum_{i=1}^N \mathbb{E}[C_{N,i}(W_i)^2] = O(1).$$

(iv) $\hat{\theta} - \theta_0 = o_P(1)$, and

$$\frac{1}{N} \sum_{i=1}^N \mathbb{E}[\|\hat{\nu}_i - \nu_{i0}\|^{2(q+1)}] = o(N^{-1}).$$

(v) The probability limit

$$G = \operatorname{plim}_{N \rightarrow \infty} \frac{1}{N} \sum_{i=1}^N \frac{\partial \psi(W_i; \theta_0, \nu_{i0})}{\partial \theta'}$$

exists as a $d_\psi \times d_\theta$ matrix, and $G' \Omega G$ is nonsingular.

Part (ii) mirrors the bounded-nuisance-dimension condition of [Bonhomme, Jochmans and Weidner \(2025\)](#). In the grouped data example of Section 3, $d_{\nu,i} = d_\eta + d_\eta(d_\eta + 1)/2$,

which is independent of N and T . Part (iv) is the key rate condition on the preliminary estimator. In the grouped data example, $\hat{\nu}_i$ is an OLS-based estimator computed on $|\mathcal{S}_1|$ observations, so $\mathbb{E}[\|\hat{\nu}_i - \nu_{i0}\|^{2(q+1)}] = O(T^{-(q+1)})$ if $|\mathcal{S}_1|/T$ tends to a non-zero constant, and the rate condition is satisfied whenever $N = o(T^{q+1})$. Higher-order orthogonality thus permits inference in grouped settings where the group sizes are small relative to the number of groups. Part (v) is a standard rank condition.

Assumption 6.

- (i) Each component of the moment function $\psi(W_i; \theta, \nu_i)$ is Neyman-orthogonal to order q in ν_i in the sense of Definition 1, and $(1/N) \sum_i \mathbb{E}[\psi(W_i; \theta_0, \nu_{i0})] = 0$.
- (ii) $\hat{\nu}_i$ is independent of the observations in W_i used to evaluate $\psi(W_i; \theta, \hat{\nu}_i)$, for every i .
- (iii) W_1, \dots, W_N are independent across i .
- (iv) With $\xi_{N,i} = G' \Omega \psi(W_i; \theta_0, \nu_{i0}) \in \mathbb{R}^{d_\theta}$, the Lindeberg condition holds, and

$$V = \text{plim}_{N \rightarrow \infty} \frac{1}{N} \sum_{i=1}^N \text{Var}(\xi_{N,i})$$

exists as a $d_\theta \times d_\theta$ positive definite matrix.

Part (ii) is achieved by sample splitting within each unit, as described above. Part (iii) requires independence across units. It is straightforward to modify the variance formula in Theorem 3 below to accommodate particular forms of dependence across units, by substituting an appropriate expression for V .

Theorem 3. Under Assumptions 5 and 6, with the same $q \in \{1, 2, 3, \dots\}$,

$$\sqrt{N} (\hat{\theta} - \theta_0) \xrightarrow{d} \mathcal{N}(0, (G' \Omega G)^{-1} V (G' \Omega G)^{-1}).$$

The proof is in Appendix B.3.

The asymptotic variance depends on the orthogonality order q through the form of the asymptotic variance, although we have left this dependence implicit in the notation. As in Bonhomme, Jochmans and Weidner (2025), the use of cross-fitting mitigates the variability introduced by a single split.

References

- Angrist, J. D. and B. Frandsen (2022). Machine labor. *Journal of Labor Economics* 40, S97–S140.
- Belloni, A., V. Chernozhukov, and C. Hansen (2014). Inference on treatment effects after selection among high-dimensional controls. *Review of Economic Studies* 81, 608–650.
- Bickel, P. J., C. A. J. Klaassen, Y. Ritov, and J. A. Wellner (1993). *Efficient and Adaptive Estimation for Semiparametric Models*. Baltimore: Johns Hopkins University Press.
- Bonhomme, S., K. Jochmans, and M. Weidner (2025). A Neyman-orthogonalization approach to the incidental parameter problem. *Mimeo*.
- Butcher, J. C. (1963). Coefficients for the study of Runge–Kutta integration processes. *Journal of the Australian Mathematical Society* 3, 185–201.
- Cattaneo, M. D., M. Jansson, and X. Ma (2019). Two-step estimation and inference with possibly many included covariates. *The Review of Economic Studies* 86, 1095–1122.
- Cayley, A. (1857). On the theory of the analytical forms called trees. *Philosophical Magazine* 13(85), 172–176.
- Chamberlain, G. (1992). Efficiency bounds for semiparametric regression. *Econometrica* 60, 567–596.
- Chernozhukov, V., D. Chetverikov, M. Demirer, E. Duflo, C. Hansen, W. Newey, and J. Robins (2018). Double/debiased machine learning for treatment and structural parameters. *Econometrics Journal* 21, C1–C68.
- Chetverikov, D., J. R.-V. Sorensen, and A. Tsyvinski (2026). Triple/double-debiased Lasso. *arXiv preprint arXiv:2603.20134*.
- Hahn, J. and W. K. Newey (2004). Jackknife and analytical bias reduction for nonlinear panel models. *Econometrica* 72, 1295–1319.
- Hairer, E., C. Lubich, and G. Wanner (2006). *Geometric Numerical Integration: Structure-Preserving Algorithms for Ordinary Differential Equations* (2nd ed.), Volume 31 of *Springer Series in Computational Mathematics*. Berlin: Springer-Verlag.
- Hansen, L. P. (1982). Large sample properties of generalized method of moments estimators. *Econometrica* 50, 1029–1054.

- Javanmard, A. and A. Montanari (2014). Confidence intervals and hypothesis testing for high-dimensional regression. *Journal of Machine Learning Research* 15, 2869–2909.
- Jochmans, K. and M. Weidner (2019). Fixed-effect regressions on network data. *Econometrica* 87, 1543–1560.
- Kline, P., E. K. Rose, and C. R. Walters (2022). Systemic discrimination among large U.S. employers. *Quarterly Journal of Economics* 137, 1963–2036.
- Kline, P., R. Saggio, and M. S¸olvsten (2020). Leave-out estimation of variance components. *Econometrica* 88, 1859–1898.
- Li, H., B. Lindsay, and R. Waterman (2003). Efficiency of projected score methods in rectangular array asymptotics. *Journal of the Royal Statistical Society, Series B* 65, 191–208.
- Mackey, L., V. Syrgkanis, and I. Zadik (2018). Orthogonal machine learning: Power and limitations. In *International Conference on Machine Learning*, pp. 3375–3383. PMLR.
- McLachlan, R. I., K. Modin, H. Munthe-Kaas, and O. Verdier (2017). Butcher series: A story of rooted trees and numerical methods for evolution equations. *Asia Pacific Mathematics Newsletter*. arXiv:1512.00906.
- Mikusheva, A. and M. S¸olvsten (2025). Linear regression with weak exogeneity. *Working paper*.
- Newey, W. K. (1994). The asymptotic variance of semiparametric estimators. *Econometrica* 62, 1349–1382.
- Neyman, J. (1959). Optimal asymptotic tests of composite hypotheses. In *U. Grenander (Ed.), Probability and Statistics*, 416–444.
- Neyman, J. and E. L. Scott (1948). Consistent estimates based on partially consistent observations. *Econometrica* 16, 1–32.
- Robins, J. M., L. Li, R. Mukherjee, E. T. Tchetgen, and A. van der Vaart (2017). Minimax estimation of a functional on a structured high-dimensional model. *Annals of Statistics* 45, 1951–1987.
- Robins, J. M., L. Li, E. T. Tchetgen, and A. van der Vaart (2008). Higher order influence functions and minimax estimation of nonlinear functionals. In *Probability and Statistics: Essays in Honor of David A. Freedman*, Volume 2, pp. 335–421. IMS Collections.

- Sur, P. and E. J. Candès (2019). A modern maximum-likelihood theory for high-dimensional logistic regression. *Proceedings of the National Academy of Sciences* 116, 14516–14525.
- Valiente, G. (2002). *Algorithms on Trees and Graphs*. Berlin: Springer-Verlag.
- van de Geer, S., P. Bühlmann, Y. Ritov, and R. Dezeure (2014). On asymptotically optimal confidence regions and tests for high-dimensional models. *Annals of Statistics* 42, 1166–1202.
- van der Vaart, A. (2014). Higher order tangent spaces and influence functions. *Statistical Science* 29, 679–686.
- Wüthrich, K. and Y. Zhu (2023). Omitted variable bias of Lasso-based inference methods: A finite sample analysis. *Review of Economics and Statistics* 105, 982–997.
- Zhang, C.-H. and S. S. Zhang (2014). Confidence intervals for low dimensional parameters in high dimensional linear models. *Journal of the Royal Statistical Society: Series B* 76, 217–242.

APPENDIX

A Additional discussion and results

A.1 Tree-indexed Taylor expansions

The closed-form coefficient

$$c_{q,\tau} = \frac{(-1)^{|\tau|}}{|\text{Aut}(\tau)|} \binom{q + |\tau| - d(\tau)}{|\tau|}$$

of Theorem 2 resembles coefficients that appear in the literature on Taylor expansions indexed by rooted trees. A close match is the B -series of Butcher (1963), which expresses derivatives of the solution to an ordinary differential equation (ODE) as rooted-tree sums that share the same $1/|\text{Aut}(\tau)|$ factor. The other parts of $c_{q,\tau}$ are specific to the orthogonalization problem and do not appear in B -series. We recall the B -series example below to demystify the $1/|\text{Aut}(\tau)|$ factor, and close with a recursion for computing $|\text{Aut}(\tau)|$.

A.1.1 The B -series for a differential equation

Consider the ordinary differential equation $y'(t) = f(y(t))$ with $y(0) = y_0$, where $y : \mathbb{R} \rightarrow \mathbb{R}^n$ and $f : \mathbb{R}^n \rightarrow \mathbb{R}^n$ are smooth functions. Here, f is given and the goal is to determine y from the ODE. Differentiating $y' = f(y)$ repeatedly and evaluating at $t = 0$, the chain rule gives

$$\begin{aligned} y'(0) &= f, \\ y''(0) &= f'(f), \\ y'''(0) &= f''(f, f) + f'(f'(f)), \\ y^{(4)}(0) &= f'''(f, f, f) + 3f''(f'(f), f) + f'(f''(f, f)) + f'(f'(f'(f))), \end{aligned}$$

all evaluated at y_0 , with $f^{(k)}$ the k -th derivative tensor of f acting on k inputs. Each term in $y^{(k)}(0)$ corresponds to a rooted tree, and for $k \geq 1$ we have

$$y^{(k)}(0) = \sum_{|\tau|=k} \frac{k!}{|\text{Aut}(\tau)| \tau!} F_\tau(y_0), \quad (35)$$

where the sum runs over all rooted trees τ on k nodes. The integer $|\tau|$ is the total number of nodes of τ .⁹ The map $F_\tau(y_0)$ is the elementary differential built from f and its derivatives

⁹In the rest of the paper, $|\tau|$ denotes the number of *non-root* nodes. The conventions differ by a -1 but $|\text{Aut}(\tau)|$ is unchanged from the main text.

by a dictionary very similar to how we construct κ_τ in the main text.¹⁰ The integer $\tau!$ is called the tree factorial, and is simply the product over all nodes v of τ of the number of nodes in the subtree of τ rooted at v . $|\text{Aut}(\tau)|$ is the order of the group of bijections of the nodes of τ that fix the root and preserve the parent-child relation, exactly as in the main text.

An early version of this rooted-tree picture appears in [Cayley \(1857\)](#), and the B -series form¹¹ is developed in [Butcher \(1963\)](#). Textbook treatments are [Hairer, Lubich and Wanner \(2006\)](#), Ch. III.1, and the survey of [McLachlan, Modin, Munthe-Kaas and Verdier \(2017\)](#).

This relatively simple example already illustrates how sums over rooted trees with combinatorial factors of the form $1/|\text{Aut}(\tau)|$ arise naturally in the mathematics literature on multivariate Taylor expansions.

A.1.2 Computing $|\text{Aut}(\tau)|$

The order $|\text{Aut}(\tau)|$ admits a simple recursion. Let the root of τ have r children. Deleting the root splits τ into r rooted subtrees, one at each child. Group these subtrees into isomorphism classes (as rooted trees). For $i = 1, \dots, s$, let τ_i denote a representative subtree of the i -th class and n_i the number of subtrees in that class, so $n_1 + \dots + n_s = r$. An automorphism of τ fixes the root, permutes the n_i subtrees within each class freely, and acts on each subtree by one of its own automorphisms. This gives

$$|\text{Aut}(\tau)| = \prod_{i=1}^s n_i! |\text{Aut}(\tau_i)|^{n_i},$$

with the recursion bottoming out at the trivial single-node tree, where $|\text{Aut}(\bullet)| = 1$. The tree-isomorphism test that determines the classes admits linear-time implementations ([Valiente, 2002](#), Ch. 4).

A.2 Explicit formulas at $q = 4$

This appendix gives the explicit orthogonal moment function $\psi^{(4)}$ at $q = 4$, in both the affine and nonlinear cases. The construction follows Section 4.2: $\psi^{(4)} = \sum_{\tau \in \mathcal{T}_4} c_{4,\tau} \kappa_\tau$, where \mathcal{T}_4

¹⁰Each node of τ with k children carries $f^{(k)}(y_0)$, so each leaf carries $f(y_0)$ and the root with r children carries $f^{(r)}(y_0)$. The map $F_\tau(y_0)$ is the contraction of these tensors along the parent-child edges of τ . This mirrors the construction of κ_τ in the main text, with f playing the role of both g and m and no Λ since no implicit-equation inversion is involved.

¹¹Substituting (35) into the Taylor series $y(t) = \sum_{k \geq 0} t^k y^{(k)}(0)/k!$ gives the standard B -series form $y(t) = y_0 + \sum_{\tau} t^{|\tau|} F_\tau(y_0)/(|\text{Aut}(\tau)| \tau!)$, where the sum now runs over all rooted trees with at least one node.

contains $|\mathcal{T}_4| = 40$ trees and the closed-form coefficient (29) is

$$c_{q,\tau} = \frac{(-1)^{|\tau|}}{|\text{Aut}(\tau)|} \binom{q + |\tau| - d(\tau)}{|\tau|}.$$

We display \mathcal{T}_4 , list $c_{4,\tau}$ for every tree, and compute $\psi^{(4)}$. Throughout we write $g(W_j)$ for $g(W_j, \theta, \eta)$, $\partial_\eta^p g(W_j)$ for $\partial_\eta^p g(W_j, \theta, \eta)$, $m_\eta = \partial_\eta m(W_1, \theta, \eta)$, $m_{\eta\eta} = \partial_{\eta\eta}^2 m(W_1, \theta, \eta)$, $m_{\eta\eta\eta} = \partial_\eta^3 m(W_1, \theta, \eta)$, $m_{\eta\eta\eta\eta} = \partial_\eta^4 m(W_1, \theta, \eta)$, and $\Lambda = \Lambda(\lambda)$.

The 40 trees in \mathcal{T}_4

Among the 40 trees, 12 have $|\tau| = d(\tau)$ (no branches apart from the root, this is the affine trees, Figure 6) and 28 have $|\tau| > d(\tau)$ (at least one branch outside of the root, these are the correction terms, Figure 7).

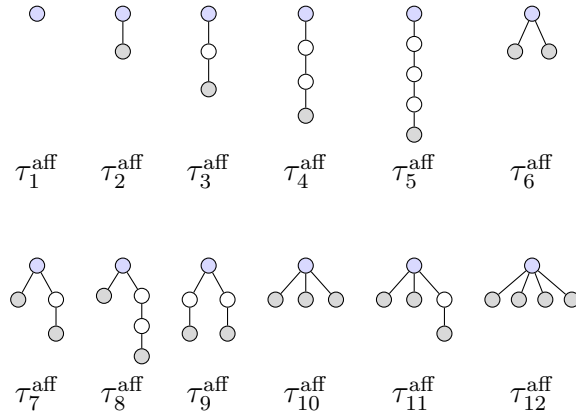


Figure 6: The 12 trees in \mathcal{T}_4 that correspond to affine terms in $\psi^{(4)}$.

Example: κ_τ and $|\text{Aut}(\tau)|$ for τ_{15}^{corr}

We illustrate the construction of the kernel κ_τ from Section 4.2.2 and the recursive computation of $|\text{Aut}(\tau)|$ from Appendix A.1 on the balanced tree τ_{15}^{corr} , in which the root has one child, that child has two children, and each of those grandchildren has two leaf children. Labeling the root by W_1 , the top internal by W_2 , the two middle internals by W_3, W_4 , and the four leaves by W_5, W_6, W_7, W_8 (any choice of distinct copies works), bottom-up evaluation gives, at each leaf, $\Lambda g(W_j)$; at each middle internal, $\Lambda \partial_\eta^2 g(W_j)$ contracted with the values of its two leaf-children; at the top internal, $\Lambda \partial_\eta^2 g(W_2)$ contracted with the values of the two middle internals; and at the root, m'_η contracted with the value of the top internal. The

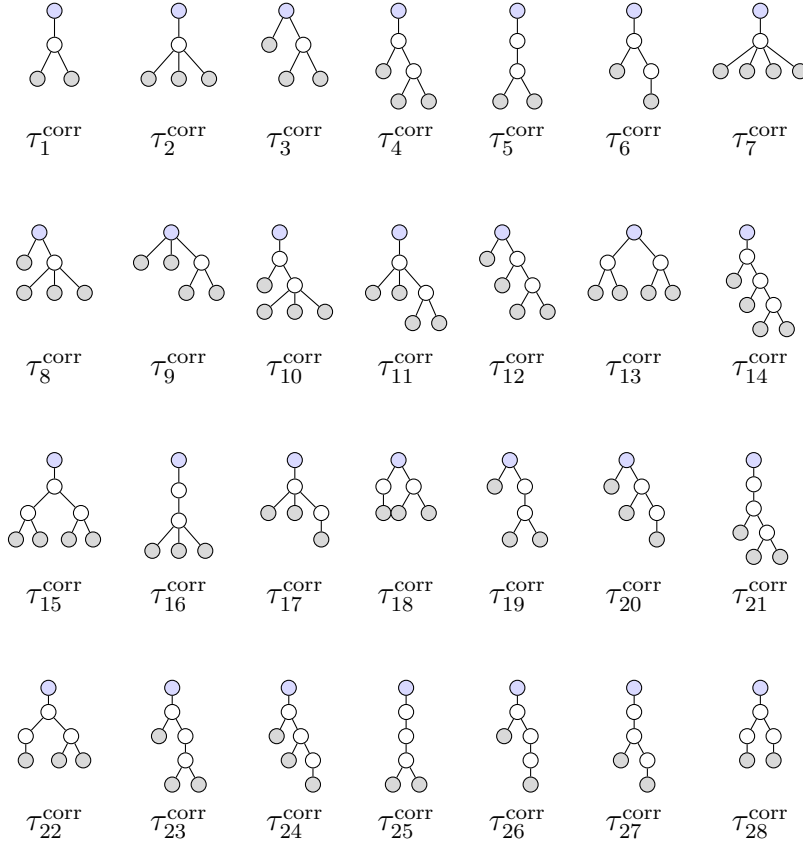


Figure 7: The 28 trees in \mathcal{T}_4 that correspond to non-linear correction terms in $\psi^{(4)}$.

result is

$$\kappa_{\tau_{15}^{\text{corr}}} = m'_\eta \Lambda \partial_\eta^2 g(W_2) \left[\Lambda \partial_\eta^2 g(W_3) \left[\Lambda g(W_5), \Lambda g(W_6) \right], \Lambda \partial_\eta^2 g(W_4) \left[\Lambda g(W_7), \Lambda g(W_8) \right] \right].$$

For $|\text{Aut}(\tau_{15}^{\text{corr}})|$, apply the recursion at each node: at each middle internal, the two leaf children form a single isomorphism class with $n = 2$, contributing $2! \cdot 1^2 = 2$, so each middle subtree has $|\text{Aut}| = 2$. At the top internal, the two middle children are isomorphic ($n = 2$, factor $2!$) and each has $|\text{Aut}| = 2$, so the top subtree has $|\text{Aut}| = 2! \cdot 2^2 = 8$. At the root, there is a single child ($n = 1$, factor $1!$), so $|\text{Aut}(\tau_{15}^{\text{corr}})| = 1! \cdot 8 = 8$. With $|\tau_{15}^{\text{corr}}| = 7$ and $d(\tau_{15}^{\text{corr}}) = 4$, the closed form gives $c_{4, \tau_{15}^{\text{corr}}} = (-1)^7 \binom{7}{4} / 8 = -1/8$.

The same recursion gives the values $|\text{Aut}(\tau_5^{\text{aff}})| = 1$ (chain of length 4, no symmetry), $|\text{Aut}(\tau_{12}^{\text{aff}})| = 4! = 24$ (root with four leaf children, all isomorphic), $|\text{Aut}(\tau_7^{\text{corr}})| = 4! = 24$ ((1, 4)-star), $|\text{Aut}(\tau_9^{\text{corr}})| = 2! \cdot 2 = 4$ ((3, 2)-star: two leaf children of the root form one class, the third child has two leaves of its own and so contributes $|\text{Aut}| = 2$), and $|\text{Aut}(\tau_{13}^{\text{corr}})| = 2! \cdot 2^2 = 8$ (root with two isomorphic children, each carrying two leaves).

Coefficient table

Tables 1 and 2 list $(|\tau|, d(\tau), |\text{Aut}(\tau)|, c_{4,\tau})$ for every tree in Figures 6 and 7, computed from (29) at $q = 4$.

τ_i^{aff}	$ \tau $	d	$ \text{Aut} $	$c_{4,\tau_i^{\text{aff}}}$
τ_1^{aff}	0	0	1	1
τ_2^{aff}	1	1	1	-4
τ_3^{aff}	2	2	1	6
τ_4^{aff}	3	3	1	-4
τ_5^{aff}	4	4	1	1
τ_6^{aff}	2	2	2	3
τ_7^{aff}	3	3	1	-4
τ_8^{aff}	4	4	1	1
τ_9^{aff}	4	4	2	1/2
τ_{10}^{aff}	3	3	6	-2/3
τ_{11}^{aff}	4	4	2	1/2
τ_{12}^{aff}	4	4	24	1/24

Table 1: Closed-form coefficients $c_{4,\tau}$ for the 12 trees in \mathcal{T}_4 with no branches.

Remark 3. The 12 affine coefficients of Table 1 are exactly the binomial weights of the affine score (17) at $q = 4$. This is the special case where every non-root node is a leaf or chain (so $|\tau| = d(\tau)$). The closed form (29) reduces to $c_{q,\tau} = (-1)^{d(\tau)} \binom{q}{d(\tau)} / |\text{Aut}(\tau)|$, which after grouping the ordered chain-length tuples (k_1, \dots, k_r) in (17) into tree classes recovers the affine binomial weights $(-1)^K \binom{q}{K} / |\text{Aut}|$.

Affine moment function

Reading off the coefficients in Table 1 and the kernels from Figure 6 via the construction rule of Section 4.2.2,

$$\begin{aligned}
\psi_{\text{aff}}^{(4)} &= m - 4 m'_\eta \Lambda g(W_2) \\
&\quad + 6 m'_\eta \left(\Lambda \partial_\eta g(W_3) \right) \Lambda g(W_2) \\
&\quad - 4 m'_\eta \left(\Lambda \partial_\eta g(W_3) \right) \left(\Lambda \partial_\eta g(W_4) \right) \Lambda g(W_2) \\
&\quad + m'_\eta \left(\Lambda \partial_\eta g(W_3) \right) \left(\Lambda \partial_\eta g(W_4) \right) \left(\Lambda \partial_\eta g(W_5) \right) \Lambda g(W_2) \\
&\quad + 3 \left(\Lambda g(W_2) \right)' m_{\eta\eta} \left(\Lambda g(W_3) \right) \\
&\quad - 4 \left(\Lambda g(W_2) \right)' m_{\eta\eta} \left(\Lambda \partial_\eta g(W_4) \right) \Lambda g(W_3) \\
&\quad + \left(\Lambda g(W_2) \right)' m_{\eta\eta} \left(\Lambda \partial_\eta g(W_4) \right) \left(\Lambda \partial_\eta g(W_5) \right) \Lambda g(W_3)
\end{aligned}$$

τ_i^{corr}	$ \tau $	d	$ \text{Aut} $	$c_{4,\tau_i^{\text{corr}}}$	τ_i^{corr}	$ \tau $	d	$ \text{Aut} $	$c_{4,\tau_i^{\text{corr}}}$
τ_1^{corr}	3	2	2	-5	τ_{15}^{corr}	7	4	8	-1/8
τ_2^{corr}	4	3	6	5/6	τ_{16}^{corr}	5	4	6	-1/6
τ_3^{corr}	4	3	2	5/2	τ_{17}^{corr}	5	4	2	-1/2
τ_4^{corr}	5	3	2	-3	τ_{18}^{corr}	5	4	2	-1/2
τ_5^{corr}	4	3	2	5/2	τ_{19}^{corr}	5	4	2	-1/2
τ_6^{corr}	4	3	1	5	τ_{20}^{corr}	5	4	1	-1
τ_7^{corr}	5	4	24	-1/24	τ_{21}^{corr}	6	4	2	1/2
τ_8^{corr}	5	4	6	-1/6	τ_{22}^{corr}	6	4	2	1/2
τ_9^{corr}	5	4	4	-1/4	τ_{23}^{corr}	6	4	2	1/2
τ_{10}^{corr}	6	4	6	1/6	τ_{24}^{corr}	6	4	1	1
τ_{11}^{corr}	6	4	4	1/4	τ_{25}^{corr}	5	4	2	-1/2
τ_{12}^{corr}	6	4	2	1/2	τ_{26}^{corr}	5	4	1	-1
τ_{13}^{corr}	6	4	8	1/8	τ_{27}^{corr}	5	4	1	-1
τ_{14}^{corr}	7	4	2	-1/2	τ_{28}^{corr}	5	4	2	-1/2

Table 2: Closed-form coefficients $c_{4,\tau}$ for the 28 trees in \mathcal{T}_4 with at least one branch. Entries are computed from (29) at $q = 4$.

$$\begin{aligned}
& + \frac{1}{2} \left((\Lambda \partial_\eta g(W_3)) \Lambda g(W_2) \right)' m_{\eta\eta} \left((\Lambda \partial_\eta g(W_5)) \Lambda g(W_4) \right) \\
& - \frac{2}{3} m_{\eta\eta\eta} [\Lambda g(W_2), \Lambda g(W_3), \Lambda g(W_4)] \\
& + \frac{1}{2} m_{\eta\eta\eta} [\Lambda g(W_2), \Lambda g(W_3), (\Lambda \partial_\eta g(W_5)) \Lambda g(W_4)] \\
& + \frac{1}{24} m_{\eta\eta\eta\eta} [\Lambda g(W_2), \Lambda g(W_3), \Lambda g(W_4), \Lambda g(W_5)].
\end{aligned}$$

The 12 terms correspond to $\tau_1^{\text{aff}}, \tau_2^{\text{aff}}, \tau_3^{\text{aff}}, \tau_4^{\text{aff}}, \tau_5^{\text{aff}}, \tau_6^{\text{aff}}, \tau_7^{\text{aff}}, \tau_8^{\text{aff}}, \tau_9^{\text{aff}}, \tau_{10}^{\text{aff}}, \tau_{11}^{\text{aff}}, \tau_{12}^{\text{aff}}$ in that order. The construction uses $L = 5$ independent copies W_1, \dots, W_5 .

Full moment function

Reading off the 28 correction-tree coefficients in Table 2 and the kernels from Figure 7 via the construction rule of Section 4.2.2,

$$\begin{aligned}
\psi^{(4)} &= \psi_{\text{aff}}^{(4)} \\
& - 5 m'_\eta \Lambda \partial_\eta^2 g(W_2) [\Lambda g(W_3), \Lambda g(W_4)] \\
& + \frac{5}{6} m'_\eta \Lambda \partial_\eta^3 g(W_2) [\Lambda g(W_3), \Lambda g(W_4), \Lambda g(W_5)] \\
& + \frac{5}{2} \partial_\eta^2 m(W_1) [\Lambda g(W_2), \Lambda \partial_\eta^2 g(W_3) [\Lambda g(W_4), \Lambda g(W_5)]] \\
& - 3 m'_\eta \Lambda \partial_\eta^2 g(W_2) [\Lambda g(W_3), \Lambda \partial_\eta^2 g(W_4) [\Lambda g(W_5), \Lambda g(W_6)]] \\
& + \frac{5}{2} m'_\eta \Lambda \partial_\eta g(W_2) \Lambda \partial_\eta^2 g(W_3) [\Lambda g(W_4), \Lambda g(W_5)]
\end{aligned}$$

$$\begin{aligned}
& + 5 m'_\eta \Lambda \partial_\eta^2 g(W_2) \left[\Lambda \partial_\eta g(W_3) \Lambda g(W_4), \Lambda g(W_5) \right] \\
& - \frac{1}{24} m'_\eta \Lambda \partial_\eta^4 g(W_2) \left[\Lambda g(W_3), \Lambda g(W_4), \Lambda g(W_5), \Lambda g(W_6) \right] \\
& - \frac{1}{6} \partial_\eta^2 m(W_1) \left[\Lambda g(W_2), \Lambda \partial_\eta^3 g(W_3) [\Lambda g(W_4), \Lambda g(W_5), \Lambda g(W_6)] \right] \\
& - \frac{1}{4} \partial_\eta^3 m(W_1) \left[\Lambda g(W_2), \Lambda g(W_3), \Lambda \partial_\eta^2 g(W_4) [\Lambda g(W_5), \Lambda g(W_6)] \right] \\
& + \frac{1}{6} m'_\eta \Lambda \partial_\eta^2 g(W_2) \left[\Lambda g(W_3), \Lambda \partial_\eta^3 g(W_4) [\Lambda g(W_5), \Lambda g(W_6), \Lambda g(W_7)] \right] \\
& + \frac{1}{4} m'_\eta \Lambda \partial_\eta^3 g(W_2) \left[\Lambda g(W_3), \Lambda g(W_4), \Lambda \partial_\eta^2 g(W_5) [\Lambda g(W_6), \Lambda g(W_7)] \right] \\
& + \frac{1}{2} \partial_\eta^2 m(W_1) \left[\Lambda g(W_2), \Lambda \partial_\eta^2 g(W_3) [\Lambda g(W_4), \Lambda \partial_\eta^2 g(W_5) [\Lambda g(W_6), \Lambda g(W_7)]] \right] \\
& + \frac{1}{8} \partial_\eta^2 m(W_1) \left[\Lambda \partial_\eta^2 g(W_2) [\Lambda g(W_3), \Lambda g(W_4)], \Lambda \partial_\eta^2 g(W_5) [\Lambda g(W_6), \Lambda g(W_7)] \right] \\
& - \frac{1}{2} m'_\eta \Lambda \partial_\eta^2 g(W_2) \left[\Lambda g(W_3), \Lambda \partial_\eta^2 g(W_4) [\Lambda g(W_5), \Lambda \partial_\eta^2 g(W_6) [\Lambda g(W_7), \Lambda g(W_8)]] \right] \\
& - \frac{1}{8} m'_\eta \Lambda \partial_\eta^2 g(W_2) \left[\Lambda \partial_\eta^2 g(W_3) [\Lambda g(W_5), \Lambda g(W_6)], \Lambda \partial_\eta^2 g(W_4) [\Lambda g(W_7), \Lambda g(W_8)] \right] \\
& - \frac{1}{6} m'_\eta \Lambda \partial_\eta g(W_2) \Lambda \partial_\eta^3 g(W_3) \left[\Lambda g(W_4), \Lambda g(W_5), \Lambda g(W_6) \right] \\
& - \frac{1}{2} m'_\eta \Lambda \partial_\eta^3 g(W_2) \left[\Lambda g(W_3), \Lambda g(W_4), \Lambda \partial_\eta g(W_5) \Lambda g(W_6) \right] \\
& - \frac{1}{2} \partial_\eta^2 m(W_1) \left[\Lambda \partial_\eta g(W_2) \Lambda g(W_3), \Lambda \partial_\eta^2 g(W_4) [\Lambda g(W_5), \Lambda g(W_6)] \right] \\
& - \frac{1}{2} \partial_\eta^2 m(W_1) \left[\Lambda g(W_2), \Lambda \partial_\eta g(W_3) \Lambda \partial_\eta^2 g(W_4) [\Lambda g(W_5), \Lambda g(W_6)] \right] \\
& - \partial_\eta^2 m(W_1) \left[\Lambda g(W_2), \Lambda \partial_\eta^2 g(W_3) [\Lambda g(W_4), \Lambda \partial_\eta g(W_5) \Lambda g(W_6)] \right] \\
& + \frac{1}{2} m'_\eta \Lambda \partial_\eta g(W_2) \Lambda \partial_\eta^2 g(W_3) \left[\Lambda g(W_4), \Lambda \partial_\eta^2 g(W_5) [\Lambda g(W_6), \Lambda g(W_7)] \right] \\
& + \frac{1}{2} m'_\eta \Lambda \partial_\eta^2 g(W_2) \left[\Lambda \partial_\eta g(W_3) \Lambda g(W_4), \Lambda \partial_\eta^2 g(W_5) [\Lambda g(W_6), \Lambda g(W_7)] \right] \\
& + \frac{1}{2} m'_\eta \Lambda \partial_\eta^2 g(W_2) \left[\Lambda g(W_3), \Lambda \partial_\eta g(W_4) \Lambda \partial_\eta^2 g(W_5) [\Lambda g(W_6), \Lambda g(W_7)] \right] \\
& + m'_\eta \Lambda \partial_\eta^2 g(W_2) \left[\Lambda g(W_3), \Lambda \partial_\eta^2 g(W_4) [\Lambda g(W_5), \Lambda \partial_\eta g(W_6) \Lambda g(W_7)] \right] \\
& - \frac{1}{2} m'_\eta \Lambda \partial_\eta g(W_2) \Lambda \partial_\eta g(W_3) \Lambda \partial_\eta^2 g(W_4) \left[\Lambda g(W_5), \Lambda g(W_6) \right] \\
& - m'_\eta \Lambda \partial_\eta^2 g(W_2) \left[\Lambda g(W_3), \Lambda \partial_\eta g(W_4) \Lambda \partial_\eta g(W_5) \Lambda g(W_6) \right] \\
& - m'_\eta \Lambda \partial_\eta g(W_2) \Lambda \partial_\eta^2 g(W_3) \left[\Lambda g(W_4), \Lambda \partial_\eta g(W_5) \Lambda g(W_6) \right] \\
& - \frac{1}{2} m'_\eta \Lambda \partial_\eta^2 g(W_2) \left[\Lambda \partial_\eta g(W_3) \Lambda g(W_4), \Lambda \partial_\eta g(W_5) \Lambda g(W_6) \right].
\end{aligned}$$

The 28 correction terms correspond to $\tau_1^{\text{corr}}, \dots, \tau_{28}^{\text{corr}}$ in that order. The construction uses $L = 8$ independent copies in the worst case (τ_{14}^{corr} and τ_{15}^{corr} , each with $|\tau| = 7$ non-root nodes plus the root copy W_1).

B Proofs

Notation. Throughout the proofs in this appendix, we write $g(W_i, \eta)$ for $g(W_i, \theta_0, \eta)$ and $m(W_i, \eta)$ for $m(W_i, \theta_0, \eta)$, suppressing the fixed argument θ_0 for brevity.

For integers a, b with $b \geq 0$, the binomial coefficient $\binom{a}{b}$ is defined as

$$\binom{a}{b} := \frac{a(a-1) \cdots (a-b+1)}{b!}.$$

In particular, $\binom{a}{0} = 1$ for all $a \geq 0$ (empty product in the numerator), and $\binom{a}{b} = 0$ whenever $0 \leq a < b$ (the numerator contains a zero factor).

In the proofs below we use formal power series in the variables x and t to read off numerical coefficients from algebraic expansions. For example, the geometric series identity $1/(1-x) = \sum_{k \geq 0} x^k$ is used as a symbolic rewriting rule. Since these series serve as combinatorial bookkeeping only, the domain of x and the convergence of the series do not matter. Every expansion is eventually truncated to a finite-degree polynomial. For a formal power series $\varphi(x) = \sum_{k \geq 0} \varphi_k x^k$, we write $[x^K] \varphi(x) := \varphi_K$ for the coefficient of x^K in $\varphi(x)$.

B.1 Proof of Theorem 1

The proof relies on the following lemma.

Lemma 1 (Weighted composition sum). *Fix integers $c_1, \dots, c_r \geq 0$ with $r \geq 1$. Let r_0 be the number of indices $s \in \{1, \dots, r\}$ with $c_s = 0$, and $r_1 = r - r_0$ the number of indices with $c_s \geq 1$. Define*

$$S(c_1, \dots, c_r) := \sum_{K=r}^q (-1)^K \binom{q}{K} \sum_{\substack{k_1, \dots, k_r \geq 1 \\ k_1 + \dots + k_r = K}} \prod_{s=1}^r (-1)^{c_s} \binom{k_s}{c_s}.$$

We extend the definition to $r = 0$ by the empty-product convention $S(\emptyset) := 1 = (-1)^0$. Then:

- (i) If $r \geq 1$, $\sum_{s=1}^r c_s + r \leq q$, and $r_1 \geq 1$, then $S(c_1, \dots, c_r) = 0$.
- (ii) $S(\underbrace{0, \dots, 0}_r) = (-1)^r$ for every $r \geq 0$.

Proof. Throughout the proof, write $\sigma := \sum_{s=1}^r c_s$ for brevity. Let x be a formal variable. For each integer $c \geq 0$, define the formal power series $f_c(x) := \sum_{k \geq 1} (-1)^c \binom{k}{c} x^k$. We have $f_0(x) = x/(1-x)$. For $c \geq 1$, differentiating the geometric series $1/(1-x) = \sum_{k \geq 0} x^k$ a

total of c times in x and dividing by $c!$ gives $\sum_{k \geq 0} \binom{k}{c} x^k = x^c / (1-x)^{c+1}$, hence $f_c(x) = (-1)^c x^c / (1-x)^{c+1}$.

Set $\varphi(x) := \prod_{s=1}^r f_{c_s}(x)$. Expanding the product, the coefficient of x^K in $\varphi(x)$ is obtained by summing over all ways to choose one term from each factor whose powers add to K :

$$[x^K] \varphi(x) = \sum_{\substack{k_1, \dots, k_r \geq 1 \\ k_1 + \dots + k_r = K}} \prod_{s=1}^r (-1)^{c_s} \binom{k_s}{c_s},$$

where the constraint $k_s \geq 1$ reflects that each f_{c_s} has no x^0 term. This is exactly the inner sum in the definition of S . Substituting,

$$S = \sum_{K=0}^q (-1)^K \binom{q}{K} [x^K] \varphi(x),$$

where the lower limit is extended from $K = r$ to $K = 0$ since $[x^K] \varphi(x) = 0$ for $K < r$.

The binomial coefficients $(-1)^K \binom{q}{K}$ are the coefficients of $(1-x)^q = \sum_{j=0}^q (-1)^j \binom{q}{j} x^j$. Expanding the product $(1-x)^q \varphi(x)$ in the same way, the coefficient of x^q is

$$[x^q] \left((1-x)^q \varphi(x) \right) = \sum_{j=0}^q (-1)^j \binom{q}{j} [x^{q-j}] \varphi(x) = (-1)^q S,$$

where the last equality follows from the substitution $K := q - j$ (using $\binom{q}{q-K} = \binom{q}{K}$ and $(-1)^{q-K} = (-1)^q (-1)^K$). Hence $S = (-1)^q [x^q] \left((1-x)^q \varphi(x) \right)$.

Multiplying out the closed forms for f_0 and f_{c_s} ($c_s \geq 1$), we have $\varphi(x) = (-1)^\sigma x^{\sigma+r_0} / (1-x)^{\sigma+r}$, so

$$(1-x)^q \varphi(x) = (-1)^\sigma x^{\sigma+r_0} (1-x)^{q-\sigma-r}. \quad (36)$$

Under the condition $\sigma + r \leq q$ of part (i), the exponent $q - \sigma - r$ is nonnegative, so the right-hand side of (36) is a polynomial in x of degree $(\sigma + r_0) + (q - \sigma - r) = q - r_1$.

For part (i): if $r_1 \geq 1$, the polynomial in (36) has degree at most $q - 1$, so $[x^q]$ of it is zero, and $S = 0$.

For part (ii): if all $c_s = 0$, then $\sigma = 0$, $r_0 = r$, $r_1 = 0$, and (36) reduces to $x^r (1-x)^{q-r}$. Its coefficient of x^q is $(-1)^{q-r}$, so $S = (-1)^q \cdot (-1)^{q-r} = (-1)^r$. \square

Proof of Theorem 1. Write $\delta := \eta - \eta_0 \in \mathbb{R}^{d_\eta}$ and $B(\lambda) := I_{d_\eta} - \Lambda(\lambda) J_0$, so $B(\lambda_0) = 0$. Define $A(\lambda) := \Lambda(\lambda) J_0 = I_{d_\eta} - B(\lambda)$. Throughout the proof, write $K := \sum_{s=1}^r k_s$.

Since $g(W_i, \eta)$ is affine in η , we have $\mathbb{E}[g(W_i, \eta_0 + \delta)] = J_0 \delta$ and $\mathbb{E}[\partial_\eta g(W_i, \eta)] = J_0$

(constant in η). Taking expectations in (17) and using independence across the copies,

$$\Psi^{(q)}(\eta, \lambda) = \sum_{r=0}^q \frac{1}{r!} \sum_{\substack{k_1, \dots, k_r \geq 1 \\ k_1 + \dots + k_r \leq q}} (-1)^K \binom{q}{K} \left\langle \mathbb{E}[\partial_\eta^r m(W_i, \eta)], \bigotimes_{s=1}^r A(\lambda)^{k_s} \delta \right\rangle. \quad (37)$$

We expand each input as $A(\lambda)^{k_s} \delta = (I - B)^{k_s} \delta = \sum_{c_s=0}^{k_s} (-1)^{c_s} \binom{k_s}{c_s} B^{c_s} \delta$, and Taylor-expand $\mathbb{E}[\partial_\eta^r m(W_i, \eta)]$ around η_0 in δ to order $q - r$, with a remainder of order $\|\delta\|^{q-r+1}$ (justified by Assumption 3(ii) and (iv), which give $(q + 1)$ -times continuous differentiability of m in η on a neighborhood of η_0 together with the integrable envelopes that allow expectation and differentiation to be interchanged). The polynomial part of $\Psi^{(q)}$ is a finite sum of multilinear contractions

$$\left\langle M_j, B^{b_1} \delta \otimes \dots \otimes B^{b_j} \delta \right\rangle, \quad M_j := \mathbb{E}[\partial_\eta^j m(W_i, \eta_0)],$$

with $j \geq r$. Of the j exponents b_s , exactly r equal the binomial parameters c_s , and the other $j - r$ are zero (from the Taylor expansion). The Taylor remainder contributes only terms of total (δ, B) -degree at least $q + 1$, which therefore have all partial derivatives of total order $\leq q$ vanishing at (η_0, λ_0) (since $\delta = 0$ and $B = 0$ at the truth). It thus suffices to check that every monomial of the polynomial part with total degree $1, \dots, q$ has zero coefficient.

Fix one such monomial. Its total degree, counting δ and B factors together, equals $\sum_s b_s + j$, and we focus on monomials whose total degree lies between 1 and q .

If some $b_s \geq 1$, those nonzero entries must all come from the binomial expansion (Taylor contributes only $b_s = 0$). Collecting in $\Psi^{(q)}$ all configurations that produce this monomial, the coefficient is a finite linear combination of values $S(c_1, \dots, c_r)$ with combinatorial weights that depend on j, r , and the placement of the Taylor-generated zero-exponent slots, but not on (k_1, \dots, k_r) . Each binomial-expansion factor adds one δ to the count j , so $r \leq j$. Combined with the total-degree bound, this gives $\sum_s c_s + r \leq q$. With $r_1 \geq 1$ in the notation of Lemma 1, part (i) of the lemma gives $S(c_1, \dots, c_r) = 0$, so the coefficient vanishes.

If all $b_s = 0$, the monomial is $\langle M_j, \delta^{\otimes j} \rangle$, and r ranges over $0, 1, \dots, j$. At each fixed r , the contribution collects a factor $1/r!$ from (37) and $1/(j - r)!$ from the order- $(j - r)$ Taylor expansion, and Lemma 1(ii) sums the remaining factors to $(-1)^r$. Combining,

$$\sum_{r=0}^j \frac{(-1)^r}{r! (j - r)!} = \frac{1}{j!} \sum_{r=0}^j \binom{j}{r} (-1)^r = \frac{(1 - 1)^j}{j!} = 0,$$

for all $j \geq 1$.

Hence every monomial of total (δ, B) -degree between 1 and q has zero coefficient in $\Psi^{(q)}(\eta, \lambda)$. Since $B(\lambda_0) = 0$ and $\delta = 0$ at $\eta = \eta_0$, this gives $\partial_\eta^\alpha \partial_\lambda^\beta \Psi^{(q)}(\eta_0, \lambda_0) = 0$ for all

$$1 \leq |\alpha| + |\beta| \leq q.$$

Finally, evaluating (37) at $\eta = \eta_0$, every term with $r \geq 1$ contains a factor $A(\lambda)^{k_s} \delta = 0$, so only the $r = 0$ term survives, giving $\Psi^{(q)}(\eta_0, \lambda_0) = \mathbb{E}[m(W_i, \eta_0)] = 0$. \square

B.2 Proof of Theorem 2

B.2.1 The case without W_i

We initially consider the case without the data W_i , where neither g nor m nor κ_τ nor $\psi^{(q)}$ has any W_i dependence. The general case is reduced to this one below. As elsewhere in Appendix B, the dependence on θ is suppressed throughout. Thus, we are given two deterministic functions $g : \mathbb{R}^{d_\eta} \rightarrow \mathbb{R}^{d_g}$ and $m : \mathbb{R}^{d_\eta} \rightarrow \mathbb{R}$, both $(q+1)$ -times continuously differentiable on a neighborhood of η_0 , and satisfying $g(\eta_0) = 0$, $m(\eta_0) = 0$, and $\partial_\eta g(\eta_0)$ of full column rank. The matrix $\Lambda(\lambda) \in \mathbb{R}^{d_\eta \times d_g}$ and the value λ_0 are as in Assumption 2, with $\Lambda_0 := \Lambda(\lambda_0)$ satisfying $\Lambda_0 \partial_\eta g(\eta_0) = I_{d_\eta}$. Define the primitives

$$A(\eta, \lambda) := \Lambda(\lambda) \partial_\eta g(\eta), \quad G_p(\eta, \lambda) := \Lambda(\lambda) \partial_\eta^p g(\eta) \quad (p \geq 2), \quad E(\eta, \lambda) := I_{d_\eta} - A(\eta, \lambda).$$

Each $G_p(\eta, \lambda)$ is a symmetric p -multilinear map $(\mathbb{R}^{d_\eta})^p \rightarrow \mathbb{R}^{d_\eta}$. We write $G_p[v_1, \dots, v_p]$ for its evaluation on p inputs $v_1, \dots, v_p \in \mathbb{R}^{d_\eta}$, and $G_p[v^{\otimes p}] := G_p[v, v, \dots, v]$ when all p inputs are equal. The same bracket notation is used for $\partial_\eta^r m(\eta)$, which is a symmetric r -multilinear map $(\mathbb{R}^{d_\eta})^r \rightarrow \mathbb{R}$.

Remember that τ is a rooted tree. The tree kernel of τ is built bottom-up by the rules of Section 4.2.2, with all W -arguments removed. Each leaf carries $\Lambda(\lambda) g(\eta)$, each chain node carries $A(\eta, \lambda)$, each branch with $p \geq 2$ children carries $G_p(\eta, \lambda)$, the root with r children carries $\partial_\eta^r m(\eta)$, and the contraction is along parent–child edges. We denote the resulting scalar by $\kappa_\tau(\eta, \lambda)$. With the closed-form coefficient

$$c_{q,\tau} = \frac{(-1)^{|\tau|}}{|\text{Aut}(\tau)|} \binom{q + |\tau| - d(\tau)}{|\tau|}$$

and support set $\mathcal{T}_q := \{\tau : d(\tau) \leq q\}$, the moment function is

$$\psi^{(q)}(\eta, \lambda) := \sum_{\tau \in \mathcal{T}_q} c_{q,\tau} \kappa_\tau(\eta, \lambda).$$

The aim is to show

$$\partial_\eta^\alpha \partial_\lambda^\beta \psi^{(q)}(\eta, \lambda) \Big|_{(\eta_0, \lambda_0)} = 0, \tag{38}$$

for all $0 \leq |\alpha| + |\beta| \leq q$. Write $\rho := \|\eta - \eta_0\| + \|\lambda - \lambda_0\|$, where $\|\cdot\|$ denotes the Euclidean

vector norm on \mathbb{R}^{d_η} and \mathbb{R}^{d_λ} , respectively. Consider ρ on a neighborhood of (η_0, λ_0) . The factors $\Lambda(\lambda)g(\eta)$ and $E(\eta, \lambda)$ are both $O(\rho)$.¹² The factors G_p and the derivatives $\partial_\eta^r m$ are bounded on a neighborhood of (η_0, λ_0) .

Main idea of the proof.

We consider the formal Lagrange–Bürmann inverse $\delta_{\text{LB}}(\varepsilon; \eta, \lambda)$, which is defined as the unique formal-power-series solution in ε at $\varepsilon = 0$, with $\delta_{\text{LB}} = 0$ at $\varepsilon = 0$, of the implicit equation

$$\Lambda(\lambda)g(\eta + \delta_{\text{LB}}) = \Lambda(\lambda)g(\eta) - \varepsilon. \quad (39)$$

Setting $\varepsilon = \Lambda(\lambda)g(\eta)$ in (39) makes the right-hand side equal to 0, so $\Lambda(\lambda)g(\eta + \delta_{\text{LB}}) = 0$. By the inverse function theorem applied at η_0 (where $\Lambda_0 \partial_\eta g(\eta_0) = I$ is invertible by Assumption 2), η_0 is the unique solution of $\Lambda(\lambda)g(\cdot) = 0$ near η_0 , so $\eta + \delta_{\text{LB}} = \eta_0$. Substituting into m gives the formal-series identity

$$m\left(\eta + \delta_{\text{LB}}\left(\Lambda(\lambda)g(\eta); \eta, \lambda\right)\right) = \underbrace{m(\eta_0)}_{=0}, \quad (40)$$

for all (η, λ) . Equivalently, defining

$$\Phi(\varepsilon; \eta, \lambda) := m\left(\eta + \delta_{\text{LB}}(\varepsilon; \eta, \lambda)\right), \quad (41)$$

the identity (40) reads $\Phi\left(\Lambda(\lambda)g(\eta); \eta, \lambda\right) = 0$, for all (η, λ) .

The main idea of the proof is the following. The function $\psi^{(q)}(\eta, \lambda)$ is the order- q Taylor expansion of $\Phi(\varepsilon; \eta, \lambda)$ in the variable $\varepsilon \in \mathbb{R}^{d_\eta}$ at $\varepsilon = 0$, evaluated at $\varepsilon = \Lambda(\lambda)g(\eta)$. Since $\Phi(\Lambda g) = 0$ exactly and $\Lambda(\lambda)g(\eta) = O(\rho)$ near (η_0, λ_0) , the Taylor remainder is $O(\rho^{q+1})$. Hence $\psi^{(q)}(\eta, \lambda) = O(\rho^{q+1})$ near (η_0, λ_0) , which implies the orthogonality (38) (a smooth function with this bound has all partial derivatives of total order $\leq q$ vanishing at the origin).

The Taylor expansion of Φ in ε has a tree structure that yields the closed form for $\psi^{(q)}$. Iterating the L–B equation (39) expresses δ_{LB} as a sum indexed by rooted trees (a B -series, see Appendix A.1 and Hairer, Lubich and Wanner, 2006, Ch. III.1), with leaves carrying ε and edges carrying the matrix inverse $A^{-1} = (I - E)^{-1}$ (where $E := I - \Lambda \partial_\eta g$, also $O(\rho)$ near (η_0, λ_0)). Substituting into $m(\eta + \delta_{\text{LB}})$, expanding A^{-1} in powers of E , and collecting all terms of order $\leq q$ in ρ gives the rooted-tree support set \mathcal{T}_q and the closed-form coefficients

¹²For Λg we have $\Lambda(\lambda)g(\eta_0) = 0$ for every λ since $g(\eta_0) = 0$, so $\Lambda(\lambda)g(\eta) = O(\|\eta - \eta_0\|)$ uniformly in λ on a neighborhood of λ_0 , and in particular $O(\rho)$. For E we have $E(\eta_0, \lambda_0) = I - \Lambda_0 \partial_\eta g(\eta_0) = 0$ by Assumption 2, so $E(\eta, \lambda) = O(\rho)$.

$c_{q,\tau}$, with the sign $(-1)^{|\tau|}$ from inverting the linearization once per non-root edge, the factor $1/|\text{Aut}(\tau)|$ from grouping orderings of children by the automorphism group of τ , and the binomial $\binom{q+|\tau|-d(\tau)}{|\tau|}$ from counting how the truncation budget is distributed across the tree's edges (with $d(\tau)$ leaf-or-chain edges and $|\tau| - d(\tau)$ branch edges).

Defining the function F_q .

Throughout this part of the proof, the formal-power-series manipulations serve as bookkeeping for the order- q Taylor expansion of $\psi^{(q)}$ in $(\eta - \eta_0, \lambda - \lambda_0)$ at the truth. Under the smoothness assumptions, this Taylor polynomial is a finite, well-defined object, and the conclusion $\psi^{(q)} = O(\rho^{q+1})$ follows from Taylor's theorem with remainder applied to the C^{q+1} function $\psi^{(q)}$ once we show that the Taylor polynomial vanishes identically.

Although several steps below are written as infinite series (the Neumann expansion $A_t^{-1} = \sum_{j \geq 0} t^j E^j$, the B -series of Lemma 2, and so on), none of the arguments require these series to converge as analytic functions. At every step we keep only the terms of total (ε, t) -degree at most $q + 1$, which form a finite polynomial in the smooth-function primitives $A, E, G_p, \partial_\eta^r m$. The omitted higher-degree terms are absorbed into a Taylor remainder that is $O(\rho^{q+1})$ by Taylor's theorem applied to the underlying C^{q+1} functions. The proof is therefore valid under C^{q+1} smoothness; the formal-series language only streamlines the bookkeeping of the finite-order computation.

Introduce a formal variable t , and let $\delta(\varepsilon, t; \eta, \lambda) \in \mathbb{R}^{d_\eta}$ be the unique formal power series¹³ in (ε, t) with $\delta = 0$ at $\varepsilon = 0$ and at $t = 0$ that satisfies

$$(1 - t) E(\eta, \lambda) \delta + \Lambda(\lambda) [g(\eta + \delta) - g(\eta)] = -t \varepsilon. \quad (42)$$

The factors $\Lambda(\lambda) g(\eta)$ and $E(\eta, \lambda)$ enter (42) symmetrically. The factor $\Lambda(\lambda) g(\eta)$ enters via the right-hand side $-t \varepsilon$ (we set $\varepsilon = \Lambda(\lambda) g(\eta)$ below), and E enters via the linear-in- δ term $(1 - t) E \delta$ on the left-hand side. At $t = 1$, equation (42) reduces to $\Lambda(\lambda) g(\eta + \delta) = \Lambda(\lambda) g(\eta) - \varepsilon$. Setting $\varepsilon = \Lambda(\lambda) g(\eta)$ gives $\Lambda(\lambda) g(\eta + \delta) = 0$, with unique formal solution $\delta = \eta_0 - \eta$ (the only solution of $\Lambda(\lambda) g(\cdot) = 0$ near η_0 , by the inverse function theorem applied at η_0 where $\Lambda_0 \partial_\eta g(\eta_0) = I$ is invertible). Substituting into $m(\eta + \delta)$ at $t = 1$,

$$m(\eta + \delta(\Lambda(\lambda) g(\eta), 1; \eta, \lambda)) = m(\eta_0). \quad (43)$$

¹³The implicit function theorem applied to (42) at the base point $(\varepsilon, t, \eta, \lambda) = (0, 0, \eta_0, \lambda_0)$ guarantees existence and uniqueness of δ , both as a formal power series and as a C^{q+1} function on a neighborhood of the base point. The linearization of the left-hand side in δ is $(1 - t) E + A = I - tE$, which evaluates to I_{d_η} at the base point since $E(\eta_0, \lambda_0) = 0$, and remains invertible for (η, λ) in a neighborhood of (η_0, λ_0) and t in a neighborhood of $[0, 1]$.

At $t = 0$, equation (42) reduces to $E\delta + \Lambda[g(\eta + \delta) - g(\eta)] = 0$. Linearizing the left-hand side in δ gives $(E + A)\delta = \delta$, which forces $\delta = 0$.

For computation, expand $\Lambda[g(\eta + \delta) - g(\eta)] = A\delta + \sum_{p \geq 2} (1/p!) G_p[\delta^{\otimes p}]$ in (42) by Taylor's formula. Using $A + (1 - t)E = I - tE$, the equation rearranges to

$$(I - tE)\delta = -t\varepsilon - \sum_{p \geq 2} \frac{1}{p!} G_p[\delta^{\otimes p}]. \quad (44)$$

Define the order- q partial sum

$$F_q(\varepsilon; \eta, \lambda) := [t^q] \frac{m(\eta + \delta(\varepsilon, t; \eta, \lambda))}{1 - t}, \quad (45)$$

where $[t^q] \varphi(t)$ denotes the coefficient of t^q in the formal power series $\varphi(t)$. Multiplying any series $\sum_k x_k t^k$ by $1/(1 - t) = \sum_{n \geq 0} t^n$ converts the coefficient of t^q in the product into the partial sum $\sum_{k=0}^q x_k$ of the original series. Hence

$$F_q(\varepsilon; \eta, \lambda) = \sum_{k=0}^q [t^k] m(\eta + \delta(\varepsilon, t; \eta, \lambda)). \quad (46)$$

By Taylor's theorem in t at $t = 0$ to order q with remainder, this partial sum is the order- q Taylor polynomial of the C^{q+1} function $t \mapsto m(\eta + \delta(\varepsilon, t; \eta, \lambda))$ evaluated at $t = 1$.

To prove the Neyman orthogonality property (38) we will establish the two facts

$$F_q(\Lambda(\lambda)g(\eta); \eta, \lambda) = \psi^{(q)}(\eta, \lambda) \quad (47)$$

and

$$F_q(\Lambda(\lambda)g(\eta); \eta, \lambda) = O(\rho^{q+1}), \quad (48)$$

on a neighborhood of (η_0, λ_0) . Together, (47) and (48) give $\psi^{(q)}(\eta, \lambda) = O(\rho^{q+1})$ on the same neighborhood, and a smooth function with this bound near the origin has all partial derivatives of total order $\leq q$ vanishing at the origin. This is (38).

Closed-form expression for F_q .

So far F_q has been defined only implicitly via the coefficient extraction (45). We now derive an explicit closed-form expression for F_q as a sum over rooted trees, from which (47) is read off directly. The remainder bound (48) is established in the part that follows.

Iterating (44) expands δ , and hence $m(\eta + \delta)$, as a sum indexed by rooted trees (a *B-series*, see Appendix A.1 and Hairer, Lubich and Wanner, 2006, Ch. III.1).

Lemma 2 (Tree expansion of $m(\eta + \delta)$). *As a formal power series in ε , t , and the primitives $A_t^{-1} := (I - tE)^{-1}$, G_p ($p \geq 2$), $\partial_\eta^r m$,*

$$m(\eta + \delta(\varepsilon, t; \eta, \lambda)) = \sum_{\tau} \frac{(-1)^{|\tau|} t^{\ell(\tau)}}{|\text{Aut}(\tau)|} V_{\tau}(\varepsilon, t; \eta, \lambda), \quad (49)$$

where the sum is over rooted trees τ in which the root has $r \geq 0$ children and every non-root node has either 0 children (a leaf) or ≥ 2 children (a branch). The elementary differential $V_{\tau}(\varepsilon, t; \eta, \lambda)$ is built by the same bottom-up tree-contraction rule as the kernel κ_{τ} at the start of Section B.2, with two changes. Leaves carry ε in place of Λg . And the value at each non-root node is left-multiplied by the matrix $A_t^{-1} = (I - tE)^{-1}$ before being passed up to its parent. The branch and root rules are unchanged: a branch with p children carrying vectors v_1, \dots, v_p evaluates to $G_p[v_1, \dots, v_p]$, and the root with r children carrying vectors v_1, \dots, v_r evaluates to $\partial_\eta^r m[v_1, \dots, v_r]$. Here $|\tau|$ is the number of non-root nodes of τ and $\ell(\tau)$ is the number of leaves.¹⁴

The proof is given in Section B.2.3. Each $A_t^{-1} = (I - tE)^{-1}$ in the elementary differential V_{τ} sits on a non-root edge of τ , namely the parent edge of some non-root node v . Expanding¹⁵ $A_t^{-1} = \sum_{j_v \geq 0} t^{j_v} E^{j_v}$ and binomial-expanding each power $E^{j_v} = (I - A)^{j_v} = \sum_{k_v=0}^{j_v} \binom{j_v}{k_v} (-1)^{k_v} A^{k_v}$, we replace each A_t^{-1} by a sum over pairs (j_v, k_v) with $0 \leq k_v \leq j_v$. Each factor of $A = \Lambda \partial_\eta g$ inserted on an edge corresponds to one chain node on that edge. To distinguish the original tree of Lemma 2 (with no chain nodes) from the chain-extended tree obtained by inserting k_v chain nodes on the parent edge of each non-root v , we write τ' for the original (the *skeleton*) and reuse τ for the chain-extended tree from this point onward. The final sum will be indexed by τ in the chain-extended sense. The chain-extended trees τ are exactly the rooted trees of the main text (Section 4.2.2), where chain nodes are allowed alongside leaves and branches; the skeleton τ' is the special subset where chain nodes are excluded, as in Lemma 2.

For the chain-extended tree τ we have $\ell(\tau) = \ell(\tau')$ (leaves are preserved), $\text{ch}(\tau) = \sum_v k_v$ (total chain nodes), and the number of branches is the same in τ and τ' (chain insertion does not create branches). Hence $|\tau| = |\tau'| + \text{ch}(\tau)$ and $d(\tau) = \ell(\tau) + \text{ch}(\tau) = d(\tau') + \text{ch}(\tau)$

¹⁴For a concrete example, take τ to be the tree where the root has one child which is itself a branch with two leaves. Then $|\tau| = 3$ (one branch and two leaves), $\ell(\tau) = 2$, $|\text{Aut}(\tau)| = 2$ (the two leaves are isomorphic), and

$$V_{\tau}(\varepsilon, t; \eta, \lambda) = \partial_\eta m[A_t^{-1} G_2[A_t^{-1} \varepsilon, A_t^{-1} \varepsilon]].$$

Reading bottom-up: each leaf produces ε , multiplied by A_t^{-1} on its parent edge; the branch contracts G_2 with the two child values $A_t^{-1} \varepsilon$, then is itself multiplied by A_t^{-1} on its own parent edge; the root contracts $\partial_\eta m$ with the resulting vector.

¹⁵The formal identity $(I - X)^{-1} = \sum_{j \geq 0} X^j$ applied with $X = tE$.

(using $d(\tau') = \ell(\tau')$ since τ' has no chain nodes), so $|\tau| - d(\tau) = |\tau'| - d(\tau')$. Substituting these expansions into (49), and writing $\text{nonroot}(\tau')$ for the set of non-root nodes of τ' ,

$$m(\eta + \delta) = \sum_{\tau'} \frac{(-1)^{|\tau'|}}{|\text{Aut}(\tau')|} t^{\ell(\tau')} \sum_{\substack{\{j_v, k_v\} \\ 0 \leq k_v \leq j_v}} \prod_{v \in \text{nonroot}(\tau')} \binom{j_v}{k_v} (-1)^{k_v} t^{j_v} V_{\tau', \{k_v\}}, \quad (50)$$

where $V_{\tau', \{k_v\}}(\varepsilon; \eta, \lambda)$ is the elementary differential of the chain-extended tree τ , built by the same bottom-up tree-contraction rule as $V_{\tau'}$ in Lemma 2 but with each A_t^{-1} on a non-root edge of τ' replaced by A^{k_v} (a string of k_v chain factors A) on that edge.¹⁶ By construction $V_{\tau', \{k_v\}}$ contains no matrix inverses, since every A_t^{-1} has been replaced by a (possibly empty) string of chain primitives A .

Apply $F_q = [t^q](\cdot)/(1-t)$ to (50) evaluated at $\varepsilon = \Lambda(\lambda)g(\eta)$. The total t -power of a contribution is $t^{\ell(\tau') + \sum_v j_v}$. Since

$$[t^q] \frac{t^N}{1-t} = \begin{cases} 1 & \text{if } N \leq q, \\ 0 & \text{otherwise,} \end{cases}$$

the partial-sum extraction imposes the constraint $\sum_v j_v \leq q - \ell(\tau')$. Hence

$$F_q(\Lambda g; \eta, \lambda) = \sum_{\tau'} \frac{(-1)^{|\tau'|}}{|\text{Aut}(\tau')|} \sum_{\{k_v\}} (-1)^{\sum_v k_v} N_{\tau'}(\{k_v\}; q) V_{\tau', \{k_v\}}, \quad (51)$$

where

$$N_{\tau'}(\{k_v\}; q) := \sum_{\substack{j_v \geq k_v \\ \sum_v j_v \leq q - \ell(\tau')}} \prod_v \binom{j_v}{k_v}. \quad (52)$$

The sum $N_{\tau'}$ admits the following closed form.

Lemma 3 (Multidimensional hockey-stick). *For nonnegative integers a_1, \dots, a_n and $M \geq 0$,*

$$\sum_{\substack{x_1, \dots, x_n \geq 0 \\ \sum_i x_i \leq M}} \prod_{i=1}^n \binom{x_i + a_i}{a_i} = \binom{M + n + \sum_i a_i}{n + \sum_i a_i}. \quad (53)$$

(Proof in Section B.2.3.) To apply Lemma 3 to $N_{\tau'}(\{k_v\}; q)$, substitute $j'_v := j_v - k_v \geq 0$.

¹⁶Continuing the example from Lemma 2, take τ' to be the tree where the root has one branch child carrying two leaves, and set $k_v = 1$ on the branch's parent edge and $k_v = 0$ on the two leaf-edges. Then

$$V_{\tau', \{k_v\}}(\varepsilon; \eta, \lambda) = \partial_\eta m[AG_2[\varepsilon, \varepsilon]],$$

namely each A_t^{-1} in the Lemma 2 footnote example has been replaced by the matrix power A^{k_v} for that edge: $A^0 = I$ on the two leaf-edges and $A^1 = A$ on the branch's parent edge.

Then $\binom{j_v}{k_v} = \binom{j'_v + k_v}{k_v}$, and the constraint $\sum_v j_v \leq q - \ell$ becomes $\sum_v j'_v \leq q - \ell - K$ with $K := \sum_v k_v$. The lemma with $n = |\tau'|$, $a_v = k_v$, and $M = q - \ell - K$ gives

$$N_{\tau'}(\{k_v\}; q) = \binom{(q - \ell - K) + |\tau'| + K}{|\tau'| + K} = \binom{q + |\tau'| - \ell}{|\tau|},$$

using $|\tau| = |\tau'| + K$. Since $|\tau'| - \ell(\tau') = |\tau| - d(\tau)$ (both equal the number of branches, preserved under chain insertion), this simplifies to

$$N_{\tau'}(\{k_v\}; q) = \binom{q + |\tau| - d(\tau)}{|\tau|}. \quad (54)$$

The right-hand side depends on the chain pattern $\{k_v\}$ only through $|\tau|$ and $d(\tau)$.

The total sign factor on τ in (51) is

$$(-1)^{|\tau'|} \cdot (-1)^{\sum_v k_v} = (-1)^{|\tau'| + \text{ch}(\tau)} = (-1)^{|\tau|},$$

using $|\tau| = |\tau'| + \text{ch}(\tau)$.

To convert the sum (51) indexed by $(\tau', \{k_v\})$ into a sum indexed by chain-extended trees τ , count the number of labeled $\{k_v\}$ -assignments that produce a fixed τ . By the ‘‘orbit–stabilizer theorem’’¹⁷ applied to the action of $\text{Aut}(\tau')$ on chain assignments,

$$\#\{\text{labeled } \{k_v\} \text{ producing } \tau\} = \frac{|\text{Aut}(\tau')|}{|\text{Aut}(\tau)|}.$$

Chain nodes within a single chain are linearly ordered by their depth from the parent edge, so no automorphism of τ can permute nodes within a chain. Hence the stabilizer of the chain-length assignment $v \mapsto k_v$ inside $\text{Aut}(\tau')$ is canonically isomorphic to $\text{Aut}(\tau)$, and the orbit–stabilizer count is exact. Combining with the prefactor $1/|\text{Aut}(\tau')|$ in (51),

$$\frac{1}{|\text{Aut}(\tau')|} \cdot \frac{|\text{Aut}(\tau')|}{|\text{Aut}(\tau)|} = \frac{1}{|\text{Aut}(\tau)|}.$$

Reindexing the sum (51) from $(\tau', \{k_v\})$ to τ and using (54),

$$F_q(\Lambda g; \eta, \lambda) = \sum_{\tau: d(\tau) \leq q} \frac{(-1)^{|\tau|}}{|\text{Aut}(\tau)|} \binom{q + |\tau| - d(\tau)}{|\tau|} V_\tau(\eta, \lambda). \quad (55)$$

¹⁷Applied here, this is the following counting statement. The symmetries of the skeleton τ' act on the chain-length labels $\{k_v\}$ by permuting non-root nodes. Two assignments produce the same chain-extended tree τ precisely when they are related by such a symmetry. The number of assignments producing a given τ therefore equals $|\text{Aut}(\tau')|$ divided by the number of symmetries of τ' that leave the assignment unchanged.

The constraint $d(\tau) \leq q$ in the summation comes from $k_v \leq j_v$ for each v , combined with the partial-sum extraction's constraint $\sum_v j_v \leq q - \ell(\tau')$. Together these give $d(\tau) = \ell(\tau') + \sum_v k_v \leq \ell(\tau') + \sum_v j_v \leq q$. Conversely, any chain-extended τ with $d(\tau) \leq q$ arises from an admissible choice of $\{j_v\}$ (e.g. $j_v = k_v$).

The elementary differential $V_\tau(\eta, \lambda)$ matches the kernel $\kappa_\tau(\eta, \lambda)$ defined at the start of this subsection verbatim. Leaves carry Λg , chain nodes carry $A = \Lambda \partial_\eta g$, branches carry $G_p = \Lambda \partial_\eta^p g$, the root carries $\partial_\eta^r m$, and the contraction is along parent–child edges. The summation index τ ranges over $\mathcal{T}_q = \{\tau : d(\tau) \leq q\}$, so equation (55) reads

$$F_q(\Lambda g; \eta, \lambda) = \sum_{\tau \in \mathcal{T}_q} c_{q,\tau} \kappa_\tau(\eta, \lambda) = \psi^{(q)}(\eta, \lambda),$$

which is the identity (47).

Bound on the remainder.

By Taylor's theorem in t at $t = 0$ (to order q , with remainder), applied to the C^{q+1} function $h(t) := m(\eta + \delta(\Lambda g, t; \eta, \lambda))$ on $[0, 1]$,

$$h(1) = \sum_{k=0}^q \frac{h^{(k)}(0)}{k!} + R_{q+1}(\eta, \lambda),$$

where R_{q+1} is the order- $(q+1)$ Taylor remainder. The partial sum on the right equals $F_q(\Lambda g; \eta, \lambda)$ by (46), and the left side equals $m(\eta_0) = 0$ by (43). Hence $F_q(\Lambda g; \eta, \lambda) = -R_{q+1}(\eta, \lambda)$, so it suffices to bound $R_{q+1} = O(\rho^{q+1})$.

We bound R_{q+1} via the order- $(q+1)$ Taylor coefficient $[t^{q+1}]h(t) = h^{(q+1)}(0)/(q+1)!$, computed using the tree expansion of Lemma 2. Substitute the Neumann identity $A_t^{-1} = \sum_{j \geq 0} t^j E^j$ on each non-root edge (without further binomial-expanding E^{j_v}). At $\varepsilon = \Lambda(\lambda)g(\eta)$, the coefficient $[t^{q+1}]m(\eta + \delta)$ is a finite sum of monomials, one for each chain-free tree τ and each choice of $\{j_v\}_{v \in \text{nonroot}(\tau)}$ with $j_v \geq 0$ and $\ell(\tau) + \sum_v j_v = q+1$. Each monomial has $\ell(\tau)$ leaves carrying Λg , one factor E^{j_v} on the parent edge of each non-root v , branch factors G_p on each branch with p children, and a root factor $\partial_\eta^r m$.

We bound each monomial. The leaf factors Λg are $O(\rho)$. Each E^{j_v} is $O(\rho^{j_v})$, being a product of j_v copies of E which are each $O(\rho)$. The branch factors G_p and the root derivative $\partial_\eta^r m$ are bounded on a neighborhood of (η_0, λ_0) . Each monomial is therefore $O(\rho^{\ell(\tau) + \sum_v j_v}) = O(\rho^{q+1})$. Hence $h^{(q+1)}(0)/(q+1)! = O(\rho^{q+1})$. By the same tree analysis applied at the expansion point $t = s$ for any $s \in [0, 1]$ (which is admissible because the linearization $I - tE$ remains invertible there), $h^{(q+1)}(s)/(q+1)! = O(\rho^{q+1})$ uniformly in $s \in [0, 1]$. The integral form of the remainder then gives $R_{q+1} = O(\rho^{q+1})$, which is (48).

Combining (47) and (48) gives $\psi^{(q)}(\eta, \lambda) = O(\rho^{q+1})$ on a neighborhood of (η_0, λ_0) , completing the proof of (38).

B.2.2 Reduction from the case with W_i

The data-dependent orthogonal moment function of Theorem 2 is

$$\psi^{(q)}(W; \eta, \lambda) = \sum_{\tau \in \mathcal{T}_q} c_{q,\tau} \kappa_\tau(W; \eta, \lambda),$$

with the same coefficients $c_{q,\tau}$ and tree set \mathcal{T}_q as in the previous subsection. The kernel $\kappa_\tau(W; \eta, \lambda)$ is built by the same bottom-up rules as the deterministic kernel above, but with each node carrying a function of an independent copy of W_i in addition to (η, λ) . Each leaf carries $\Lambda(\lambda) g(W_j, \eta)$, each chain node carries $\Lambda(\lambda) \partial_\eta g(W_j, \eta)$, each branch with p children carries $\Lambda(\lambda) \partial_\eta^p g(W_j, \eta)$, and the root with r children carries $\partial_\eta^r m(W_1, \eta)$, with the W_j i.i.d. copies of W_i assigned one to each node of τ . The orthogonality requirement of Definition 1 concerns the population moment $\Psi^{(q)}(\theta_0, \eta, \lambda) := \mathbb{E}[\psi^{(q)}(W; \eta, \lambda)]$.

At each node v of τ , the kernel κ_τ depends on W through a single tensor evaluated on a copy of W specific to that node. Denote this tensor by X_v . The kernel κ_τ is multilinear in the tuple $(X_v)_v$, since it is a contraction along parent–child edges. The X_v are evaluated on independent copies of W_i , so they are mutually independent random variables. By independence and multilinearity, $\mathbb{E}[\kappa_\tau(W; \eta, \lambda)]$ equals the same contraction with each X_v replaced by $\mathbb{E}[X_v]$. For each non-root v , $X_v = \Lambda(\lambda) \partial_\eta^p g(W_j, \eta)$ with $p \in \{0, 1, 2, \dots\}$ equal to the node’s number of children, and

$$\mathbb{E}[\Lambda(\lambda) \partial_\eta^p g(W_j, \eta)] = \Lambda(\lambda) \partial_\eta^p \bar{g}(\eta), \quad \bar{g}(\eta) := \mathbb{E}[g(W_i, \theta_0, \eta)],$$

using interchange of expectation and differentiation (Assumption 4(iii)). For the root, $X_1 = \partial_\eta^r m(W_1, \eta)$ and $\mathbb{E}[X_1] = \partial_\eta^r \bar{m}(\eta)$ with $\bar{m}(\eta) := \mathbb{E}[m(W_i, \theta_0, \eta)]$, by the analogous interchange under Assumption 3(iv) (inherited via Assumption 4(i)). Substituting, $\mathbb{E}[\kappa_\tau(W; \eta, \lambda)] = \bar{\kappa}_\tau(\eta, \lambda)$, where $\bar{\kappa}_\tau$ is the deterministic kernel of the previous subsection built from \bar{g} and \bar{m} in place of g and m . Summing over τ ,

$$\Psi^{(q)}(\theta_0, \eta, \lambda) = \sum_{\tau \in \mathcal{T}_q} c_{q,\tau} \bar{\kappa}_\tau(\eta, \lambda) = \bar{\psi}^{(q)}(\eta, \lambda), \quad (56)$$

where $\bar{\psi}^{(q)}$ is the deterministic moment function of the previous subsection with g, m replaced by \bar{g}, \bar{m} .

The functions \bar{g}, \bar{m} satisfy the hypotheses of the previous subsection. We have $\bar{g}(\eta_0) =$

$\mathbb{E}[g(W_i, \theta_0, \eta_0)] = 0$ and $\bar{m}(\eta_0) = \mathbb{E}[m(W_i, \theta_0, \eta_0)] = 0$ from the moment conditions (1). The Jacobian $\partial_\eta \bar{g}(\eta_0) = J_0$ has full column rank by Assumption 2. For smoothness, Assumption 4 allows interchange of expectation and differentiation up to order $q + 1$, so \bar{g} and \bar{m} are $(q+1)$ -times continuously differentiable on a neighborhood of η_0 . The previous subsection therefore gives

$$\partial_\eta^\alpha \partial_\lambda^\beta \bar{\psi}^{(q)}(\eta, \lambda) \Big|_{(\eta_0, \lambda_0)} = 0,$$

for all $0 \leq |\alpha| + |\beta| \leq q$. By (56), the same holds for $\Psi^{(q)}(\theta_0, \eta, \lambda)$, which is Theorem 2. \square

B.2.3 Proofs of Lemmas 2 and 3

Proof of Lemma 2. Rewrite (44) as

$$\delta = -A_t^{-1} \left[t\varepsilon + \sum_{p \geq 2} \frac{1}{p!} G_p[\delta^{\otimes p}] \right]$$

and iterate. At each iteration, every occurrence of δ on the right-hand side is replaced by one of two patterns. Either it becomes the leaf pattern $-A_t^{-1} \cdot t\varepsilon$, contributing a single ε -leaf with A_t^{-1} on its parent edge and no further δ to expand; or it becomes the branch pattern $-A_t^{-1} \cdot (1/p!) G_p[\delta^{\otimes p}]$ for some $p \geq 2$, contributing a branch with p children carrying G_p , with A_t^{-1} on its parent edge and p further subtrees of δ to expand. The iteration terminates after finitely many steps at any fixed (ε, t) -degree, and gives a formal sum

$$\delta = \sum_{\tau^{\text{ord}}} (-1)^{|\tau^{\text{ord}}|} t^{\ell(\tau^{\text{ord}})} \prod_{v \text{ branch}} \frac{1}{p_v!} V_{\tau^{\text{ord}}}^\delta(\varepsilon, t; \eta, \lambda),$$

where τ^{ord} is a rooted tree of the type described in Lemma 2 with the children of each node further distinguished by position, and $|\tau^{\text{ord}}|$ and $\ell(\tau^{\text{ord}})$ are the number of non-root nodes and number of leaves of τ^{ord} , respectively. Substituting this δ -expansion into the Taylor expansion $m(\eta + \delta) = \sum_{r \geq 0} (1/r!) \partial_\eta^r m(\eta) [\delta^{\otimes r}]$, every δ in $\delta^{\otimes r}$ becomes a δ -tree, giving

$$m(\eta + \delta) = \sum_{\tau^{\text{ord}}} (-1)^{|\tau^{\text{ord}}|} t^{\ell(\tau^{\text{ord}})} \frac{1}{r!} \prod_{v \text{ branch}} \frac{1}{p_v!} V_{\tau^{\text{ord}}}^\delta(\varepsilon, t; \eta, \lambda), \quad (57)$$

where r is the number of children of the root of τ^{ord} .

Group these trees by their underlying tree τ , forgetting the ordering of children at each node. By the ‘‘orbit–stabilizer theorem’’ applied to the action of $S_r \times \prod_v S_{p_v}$ on labelings of the children at each node, the number of distinct τ^{ord} representing a fixed τ is $r! \prod_v p_v! / |\text{Aut}(\tau)|$. Here the numerator $r! \prod_v p_v!$ counts all orderings of children (with p_v the number of children

of node v), and $|\text{Aut}(\tau)|$ is the size of the stabilizer (the order of the automorphism group of τ). Multiplying by the per-ordering weight $\frac{1}{r!} \prod_v \frac{1}{p_v!}$ in (57) collapses to $\frac{1}{|\text{Aut}(\tau)|}$, and substituting back gives (49). \square

Proof of Lemma 3. The formal-power-series identity $\sum_{x \geq 0} \binom{x+a}{a} t^x = (1-t)^{-(a+1)}$ follows from differentiating $1/(1-t) = \sum_{x \geq 0} t^x$ a total of a times in t and dividing by $a!$. Multiplying n such identities,

$$\prod_{i=1}^n \sum_{x_i \geq 0} \binom{x_i + a_i}{a_i} t^{x_i} = (1-t)^{-s} = \sum_{X \geq 0} \binom{X + s - 1}{s - 1} t^X,$$

with $s := n + \sum_i a_i$. The left-hand side, expanded as a single power series in t , has coefficient of t^X equal to $\sum_{x_1 + \dots + x_n = X} \prod_i \binom{x_i + a_i}{a_i}$. Equating coefficients gives this sum equal to $\binom{X+s-1}{s-1}$. Summing over $X = 0, \dots, M$ and using the standard one-dimensional hockey-stick identity¹⁸ gives (53). \square

B.3 Proof of Theorem 3

Theorem 3 is a direct consequence of the asymptotic result in Section 6 of Bonhomme, Jochmans and Weidner (2025) after a change of notation. We record the translation here rather than repeat the proof.

Identify Z_i , $\eta_{n,i}$, μ , and u in Bonhomme, Jochmans and Weidner (2025) with W_i , $\nu_i = (\eta_i, \lambda_i)$, θ , and ψ in our notation, respectively. The number of cross-sectional units, denoted N in both papers, plays the same role. The parameter θ of Bonhomme, Jochmans and Weidner (2025) has no counterpart in our setting, so their Assumption 1(vi) on its asymptotic linearity, and the corresponding $G_\theta \psi_{n,i}$ term appearing in their expansion lemma, are dropped. In the grouped data setting we consider, the moment function $\psi(W_i; \theta, \nu_i)$ is bounded per unit rather than scaling with the number of observations per unit, so n_i in their Assumption 1(iv) can be set to one and the total sample size $n = \sum_i n_i$ collapses to N . The \sqrt{n} -asymptotics of their Theorem 2 then become the \sqrt{N} -asymptotics of our Theorem 3.

Under this mapping, our Assumption 5 corresponds to their Assumption 1, and our Assumption 6 corresponds to their Assumption 2. The expansion lemma and the final central limit theorem of their Section 6 apply verbatim and yield Theorem 3.

¹⁸ $\sum_{X=0}^M \binom{X+s-1}{s-1} = \binom{M+s}{s}$, which follows by induction on M from Pascal's rule $\binom{n+1}{k} = \binom{n}{k-1} + \binom{n}{k}$.