FAST ONLINE INFERENCE ON SEMIPARAMETRIC MODELS

By

Xiaohong Chen, Elie Tamer and Qingsong Yao

July 2026

COWLES FOUNDATION DISCUSSION PAPER NO. 2548

COWLES FOUNDATION FOR RESEARCH IN ECONOMICS

YALE UNIVERSITY
Box 208281
New Haven, Connecticut 06520-8281

http://cowles.yale.edu/

Fast Online Inference on Semiparametric Models*

Xiaohong Chen[†] Elie Tamer[‡] Qingsong Yao[§]

July 27, 2026

Abstract

This paper develops a framework for fast online inference on semiparametric models with large sample sizes and possibly many covariates. The computational algorithm itself is the object of statistical study: after a globally consistent warm start in the first phase, the path of averaged online iterates generated in the second phase automatically delivers estimators with optimal convergence rates and valid confidence sets. Both phases require only a single pass over the data stream and are well suited to streaming data or to settings with storage/privacy constraints. For semiparametric monotone index models, the averaged trajectory of the second phase lead to estimators that are automatically orthogonalized and satisfy the laws of the iterated logarithms, and policy functionals are updated along the same trajectory at negligible additional cost. The averaged trajectories satisfy functional central limit theorems, which yield fast online inference via random scaling and bypass the explicit variance estimation that complicates inference for semiparametric models. Applied to a fixed large sample, our online algorithm achieves substantial computational gains over corresponding offline procedures without sacrificing statistical performance. Monte Carlo experiments show adequate behavior. Our methods are applied to 19 million traffic-stop online records from the North Carolina State Patrol (Pierson et al. 2020) and to the international trade data of Helpman et al. (2008) with over 300 regressors. We also compare our estimator to its parametric benchmark in both empirical illustrations.

*First version: arXiv:2603.08614v1. We acknowledge helpful comments from participants of the Econometrics Seminars at Brown, Renmin and USP Universities, the 2026 Cowles Summer Conference in New Haven and the 2026 CIREQ-HEC Women in Econometrics Conference in Montreal.

†Dept. of Economics, Yale University.

‡Dept. of Economics, Harvard University.

§Dept. of Economics, Louisiana State University.

1. Introduction

An extensive literature in econometrics examines inference on flexible semiparametric models (Powell 1994). Traditionally, this literature has treated the estimator as given: one posits that an estimator exists and then derives its statistical properties—global consistency, convergence rates, limiting distributions, and consistent standard errors—while the computational machinery for actually obtaining that estimator is treated as a separate concern, with no bearing on the large-sample theory. Computation, however, has become a first-order issue in modern economic research. Data relevant to applied economists increasingly arrive as streams—the Stanford Open Policing Project has recorded hundreds of millions of traffic stops and is updated frequently (Pierson et al. 2020); scanner systems (Broda and Weinstein 2010; Atkin et al. 2018; DellaVigna and Gentzkow 2019), credit registries (Jiménez et al. 2014; Agarwal et al. 2018)-, and digital platforms (Cavallo 2017) generate observations at a pace that makes repeated full-sample re-estimation impractical and—when records cannot be retained for privacy or memory reasons— not feasible. Even in conventional fixed-sample settings, semiparametric models with many covariates and large sample sizes are costly to compute and pose significant challenges for inference. This has led some practitioners to fall back on convenient parametric specifications—such as Logit or Probit links in binary response models—even when those assumptions are unwarranted and can lead to bias in policy evaluations.

In light of these challenges, we revisit inference for an important class of semiparametric models and take a fundamentally different approach: the computational algorithm itself becomes the object of statistical study. We propose a novel online algorithm designed for these models and use the sequence of iterates it generates as the statistical scaffold for deriving inference results, rather than treating computation and statistics as separate enterprises. After a quick initial warm start phase, the algorithm treats the current estimates of the parametric and nonparametric components as summary statistics, uses a small batch of newly arrived data to update both components jointly, and outputs the trajectory of the Polyak–Ruppert (PR) averaged iterates (Ruppert 1988; Polyak and Juditsky 1992) together with the averaged policy-functional iterates. Coupling computation with statistics pays off on both fronts. Theoretically, the PR averaged estimates converge to the truth almost surely at optimal rates and are automatically orthogonalized, and the trajectory of the PR averages satisfies a functional central limit theorem (FCLT). The FCLT enables us to construct fast online confidence sets via self-normalization (or "random scaling" of Lee et al. (2022)), bypassing the complicated variance estimation and/or bootstrapping for inference on semiparametric models using full sample data. Practically, the automatic orthogonalization guarantees negligible first-order bias, which, together with the FCLT, yields fast and valid confidence sets even when the number of covariates is large. Each data point is processed only once, so the method is naturally suited to streaming data and to storage- or privacy-constrained environments; and the same trajectory also delivers almost surely consistent estimates of, and valid inference on, policy-relevant functionals—such as average marginal effects—at negligible additional cost.

The above described online principle applies to general semiparametric models. For a large class of semiparametric nonlinear models with many covariates, however, it is difficult to design a warm start phase that can quickly enter the stable local neighborhood of the global true parameter values regardless of the initialization of the algorithm, as well as difficult to establish almost sure convergence with optimal rates and trajectory FCLTs under lower level sufficient conditions. In this paper, we focus on a leading class of semiparametric nonlinear models, the following semiparameric monotone index model:

$$Y = F_0(Z_0) + \varepsilon, \qquad Z_0 := x_0 + X^\top \theta_0, \qquad \mathbb{E}(\varepsilon \mid x_0, X) = 0,$$

where F_0 is the unknown and monotone link function, $\theta \in \mathbb{R}^p$ is unknown parameter of interest with a fixed but possibly large p (e.g., $p = 50, 100, 500$ in our Monte Carlo simulations, and $p > 300$ in our trade application), the normalization on x_0 gives an interpretable index direction, and some of the covariates X could be discrete (as in our applications and Monte Carlo experiments). An important parameter of interest is the average marginal effects (AME) parameter $\tau_0 := \mathbb{E}\left[\partial \mathbb{E}(Y|x_0, X)/\partial X\right] \equiv \mathbb{E}\left[\nabla_z F_0(Z_0)\,\theta_0\right]$.

This monotone index class contains many workhorse models used in applied microeconomics, including binary choice, censored regression, and duration models (Powell 1994). Despite many theoretical works on identification, estimation and inference for θ_0, F_0, τ_0 in the literature, due to the computational difficulties of the existing procedures, there are very few empirical applications in economics allowing for unknown link function F_0 with many covariates and large sample sizes. Our paper will propose a novel, computationally fast online estimation and inference method for these semiparametric nonlinear models, in which the unknown monotone link function F_0 is approximated by a sequence of B-spline sieves[1]. The semiparametric online problem is theoretically challenging for two reasons: the joint sieve nonlinear least-squares (NLS) criterion for θ_0, F_0 is not globally convex; and the sieve dimen-

[1] B-spline sieve is known to preserve monotonic shape. (Chen 2007). Other shape-preserving sieves could also be used to approximate unknown monotone link function F_0; see Section Section 7.1.1.

sion needs to grow slowly as data accumulate. These features place our problem outside the existing literature on online M-estimation of θ_0, F_0 and inference on θ_0, τ_0 (see the related literature below for more discussions).

Contributions This paper makes three main contributions. The first is algorithmic. We propose a two-phase online learner for semiparametric index models. The warm-start phase uses a kernelized rank-based update, inspired by the maximum rank correlation of Han (1987) and the smoothed maximum score of Horowitz (1992), to obtain an almost surely consistent estimate of the index parameter θ_0 from an arbitrary initialization. Our warm-start algorithm for θ is globally convex and stable, and hence converges to the true parameter θ_0 regardless of the choice of initialization. This warm start in turn generates a consistent index $Z_0 = x_0 + X^\top\theta_0$ estimate, on which a B-spline sieve least-squares update provides an almost sure consistent warm start for the nonparametric link component F_0. Initialized at these warm starts for (θ, F), the second phase performs joint local stochastic-gradient updates for (θ_0, F_0) via a sieve NLS criterion. To make sieve learning genuinely online, we introduce a re-embedding operator that maps previously learned sieve coefficients into the newly expanded sieve space; this preserves the fitted function exactly when the sieve dimension grows and allows the online path to continue without restarting the optimization.

The second contribution is theoretical. For the warm-start phase, we establish an almost-sure global consistency with rate for the online warm-start estimator of θ_0, together with an almost-sure sup-norm convergence rate for online sieve estimator of F_0 with generated regressors. For the joint local refinement phase, we prove an asymptotic linear representation for the PR averages. The representation shows that the leading score for θ_0 is automatically orthogonalized: the joint local update removes the first-order effect of estimating the unknown link F_0, without requiring a researcher to estimate a separate high-dimensional Riesz representer in order to construct an orthogonal score manually. The PR-averaged estimator obeys an almost-sure rate-optimal law of the iterated logarithm, and satisfies an FCLT. We further establish a corresponding limit theory for the online plug-in estimator of the AME τ_0, so the same path supports inference on objects directly tied to policy evaluations.

The third contribution is practical. After a quick warm start, the procedure uses each new small fixed batch once, updates the parameter, sieve coefficients, and policy-functional running sums, and avoids repeated full-sample optimization. This is valuable for streaming data, in environments in which storage is costly or restricted, and for conventional large offline samples in which repeated full-sample nonconvex optimization is slow. In our Monte

Carlo experiments, the method remains accurate in dimensions as large as $p = 500$. In online vs offline timing comparisons with $p = 100$, online local refinement is roughly one to two orders of magnitude faster than full-sample sieve NLS while delivering comparable or smaller estimation error. Because the construction of confidence sets for θ_0, τ_0 uses only the stored trajectory of averaged updates via random scaling, it completely avoids the computational bottlenecks of conventional semiparametric inferences that require either additional nonparametric estimation of Riesz representers (to compute semiparametric standard errors) or additional numerical methods such as bootstraps.

Empirical Applications We illustrate the method in two empirical settings. The first uses the bilateral trade panel of Helpman et al. (2008), a binary-choice application with more than 300 regressors once exporter, importer, and year fixed effects are included. A calibrated Probit experiment first verifies that the online semiparametric estimator recovers the parametric truth when the Probit link is correctly specified. We then apply our method to the observed trade outcomes without imposing a parametric link. The online semiparametric estimates agree closely with the parametric benchmarks for some geographic determinants of trade such as shared land border and landlock, but they differ substantially for the institutional and informational gravity variables—colonial ties, common language, shared religion, currency union, and free trade agreements. The estimated coefficient on colonial ties, for example, is less than half of its Probit counterpart. On the probability scale, the average marginal effect of a free trade agreement is smaller than the Probit benchmark by about three percentage points—a modest difference in absolute size, but an economically meaningful one relative to the baseline probability of positive trade. This illustrates how the shape of the link function—typically fixed by assumption in parametric models—can impact estimated policy magnitudes.

The second application uses North Carolina traffic-stop records from the Stanford Open Policing Project (Pierson et al. 2020). After cleaning, the data contain more than 19 million stops ordered in time—a genuinely online administrative stream. The outcome is whether a stop results in a search. We process the post-warm-start records in their recorded order, each exactly once, and update the average marginal effects of race and gender along the online path. The warm start takes about two minutes using the first half of the 19 million data points; the online learning then processes the remaining 9.6 million post-June-2009 records in roughly 17 seconds—about half a million records per second—for a total of 2.3 minutes. The final estimates are statistically distinguishable from zero and economically sizable relative to the low unconditional search rate. Most importantly, as in the trade application, the semiparametric estimates deviate substantially from the Probit and Logit benchmarks: the parametric models overstate the Black–Hispanic search-probability gap by more than a quarter and understate the White–Hispanic gap by roughly a half. Because all specifications are estimated on the same sample with the same covariates, these discrepancies are attributable to the shape of the link function alone. Consistent with the infra-marginality concerns emphasized in the policing literature (Simoiu et al. 2017; Pierson et al. 2020), we interpret the estimated effects as descriptive rather than causal. The point of the exercise is to show that semiparametric inference can be implemented at the scale, and in the order, in which policy-relevant administrative data are generated.

Extensions We note that the Phase II algorithm and its properties remain valid with any other consistent warm starts. Indeed, our Monte Carlo studies in Section 6.3 show that the online Phase II algorithms perform virtually the same regardless whether our online Phase I algorithm or the global optimization algorithm SMCO of (Chen et al. 2026) is used. The only difference is computational speed in the warm start phase: our online Phase I exploits the semiparametric monotone index model structure so is faster to find a good warm start than the generic global optimizer SMCO. Indeed, the logic of our Phase II algorithm, coupled with any consistent warm start algorithm such as SMCO, easily extends beyond the baseline monotone index model. In Section 7, we discuss sample-selection settings in which observability is itself an economic choice (Heckman 1974; Abrevaya et al. 2010; Khan et al. 2024); multi-index and multi-alternative choice models; and monotone neural-network approximations to the unknown link. These extensions share the same architecture: a warm-start phase, such as SMCO (Chen et al. 2026), first locates a stable basin, after which a local joint sieve online learner produces averaged paths for estimation and inference.

Related Literature Our paper contributes to two literatures. The first is the large literature on identification, offline estimation and inference for semiparametric index models in economics. Well-known estimation methods include, among others, maximum rank correlation, semiparametric least squares, efficient binary-response estimation, censored and sample-selection estimators, and rank-based procedures (Han 1987; Powell et al. 1989; Härdle et al. 1993; Ichimura 1993; Klein and Spady 1993; Sherman 1993; Das et al. 2003; Ahn et al. 2018; Fan et al. 2020; Khan et al. 2025). These works are designed for fixed samples and often require optimization or nuisance-estimation steps that are computationally costly in massive-data settings: each evaluation of the maximum rank correlation criterion alone entails $n(n-1)$ pairwise comparisons (where n is the sample size), and the computational cost grows fast in the dimension of θ (i.e., the dimension of covariates X). To our knowledge, the present paper is among the first to give a joint online sieve-estimation and trajectory-inference theory for this class while retaining an interpretable index structure with many covariates.

The second is the vast literature on stochastic approximation (SA) in the tradition of Robbins and Monro (1951) and Kushner and Yin (2003), including abstract infinite-dimensional Hilbert-space online learning problems (Yin and Zhu 1990; Chen and White 2002). An important subclass of SA algorithms, stochastic gradient descent (SGD), is widely used in finite-dimensional optimization in machine learning and statistics (Ghadimi and Lan 2013; Toulis and Airoldi 2017; Bottou et al. 2018) and in nonparametric (kernel or sieve) density and conditional-mean estimation (Huang et al. 2013; Zhang and Simon 2022). Recent papers develop inference based on PR averaged SGD iterates in parametric strongly convex M-estimation problems (Chen et al. 2020; Lee et al. 2022, 2025) and in parametric over-identified GMM problems (Chen et al. 2023, 2025) for large data sets. Closest to our setting, Han et al. (2024) study online inference for θ_0 in high-dimensional single-index models with streaming data. However, they assume that (i) the covariate vector $(x_0, X^\top)$ satisfies a so-called linear expectation condition (i.e., for all parameter vector $(b_0, b^\top)$, $\mathbb{E}[x_0 b_0 + X^\top b \mid x_0 + X^\top \theta_0]$ is linear in the true index $x_0 + X^\top \theta_0$) and is statistically full independent of the latent error term); and (ii) there is a loss function $\mathbb{E}[\ell(Y, x_0 + X^\top \theta)]$ that is convex in the index, has a unique solution in θ_0 and does not depend on the unknown link function F_0. Under these and other assumptions, they study online SGD inference for θ_0 without estimating the unknown link F_0. In addition, their methods cannot perform online inference on policy functionals that depend on both θ_0 and the derivative of F_0 - like standard partial effects. Relative to this literature, our paper develops a joint sieve online estimation and trajectory-inference theory for semiparametric index models with finite- and infinite-dimensional unknowns, some discrete covariates, and policy effects.

Organization Section 2 introduces the model and reviews relevant offline estimators. Section 3 presents the two-phase online algorithm, random-scaling inference, and online estimation of policy functionals. Section 4 reports the two empirical illustrations. Section 5 establishes the statistical properties of the estimators. Section 6 reports Monte Carlo experiments. Section 7 presents conclusions and extensions. All proofs are collected in the Appendix.

Notation For nonrandom sequences $\{a_N\}$ and $\{b_N\}$ with $b_N > 0$, we write $a_N = o(b_N)$ if $\lim_{N\to\infty} a_N/b_N = 0$, and $a_N = O(b_N)$ if $\overline{\lim}_{N\to\infty}|a_N|/b_N < \infty$. We write $a_N \asymp b_N$ if both $a_N = O(b_N)$ and $b_N = O(|a_N|)$. Consider a sequence of random variables $\{a_N(w)\}$ defined on the underlying probability space $(A, \mathcal{A}, P_A)$. We say event $E \in \mathcal{A}$ occurs *almost surely* if $P_A(A \setminus E) = 0$. For any nonrandom sequence $b_N > 0$, we write $a_N = o(b_N)$ a.s. if $\lim_{N\to\infty} a_N/b_N = 0$ almost surely, and we write $a_N = O(b_N)$ a.s. if $\overline{\lim}_{N\to\infty}|a_N|/b_N < \infty$ almost surely. Specifically, if $a_N = o(b_N)$ a.s. with $b_N \equiv 1$, we write $a_N \to_{\text{a.s.}} 0$. We use $\to_d$ to denote convergence in distribution; we use $\Rightarrow$ to denote weak convergence in the metric space that will be clear from the context. For any scalar function $f(z)$ with domain $\mathcal{Z}$, $\|f\|_\infty \equiv \sup_{z\in\mathcal{Z}} |f(z)|$ denotes the sup norm. Finally, for any real symmetric matrix M, $\lambda_{\max}(M)$ and $\lambda_{\min}(M)$ denote its largest and smallest eigenvalues.

2. Model and Offline Estimation: A Selective Review

2.1. Semiparametric Monotone Index Models

We first consider the standard semiparametric monotone index model

$$Y = F_0(x_0 + X^\top\theta_0) + \varepsilon, \quad \mathbb{E}(\varepsilon|x_0, X) = 0, \tag{2.1}$$

where Y is an observed response, $F_0(\cdot)$ is an unknown monotonically increasing function, $(x_0, X^\top)^\top$ is a $(p+1)\times 1$ vector of regressors, θ_0 is a $p \times 1$ unknown parameter, and ε is an unobserved error term. The coefficient of x_0 is normalized to one for identification. Monotone index models nest binary choice, censored, hazard, and Box–Cox transformation models. Two examples illustrate the scope.

Example 1 (Binary economic choice). In a binary random-utility model (McFadden 2001), $Y = \mathbf{1}\{V_1(X) - V_0(X) - u \geq 0\}$ with systematic utility difference $V_1(X) - V_0(X) = x_0 + X^\top\theta_0$, so that $\Pr(Y = 1 \mid x_0, X) = F_0(x_0 + X^\top\theta_0)$, where F_0 is the distribution of u. Model (2.1) is therefore a distribution-free representation of an economic choice probability: it recovers the utility index and the link F_0 without imposing Logit or Probit. Semiparametric analysis of this model goes back to Manski (1975); Cosslett (1987); Han (1987); Ichimura (1993); Klein and Spady (1993).

Example 2 (Duration). In the proportional hazards model $\lambda(t \mid X) = \lambda_0(t)\exp(X^\top\theta_0)$ (Cox 1972; Kalbfleisch and Prentice 1980; Lancaster 1990), suppose the outcome records whether a spell terminates before a fixed horizon c (for example, whether an unemployment spell ends within a given period), $Y = \mathbf{1}\{T \le c\}$. Then $\mathbb{P}(Y = 1 \mid X) = 1 - \exp(-\varLambda_0(c)\exp(X^\top\theta_0)) = F_0(X^\top\theta_0)$, where $\varLambda_0(t) = \int_0^t \lambda_0(s)\,\mathrm{d}s$ and $F_0(z) = 1 - \exp(-\varLambda_0(c)e^z)$ is strictly increasing, with shape governed by the unknown cumulative baseline hazard. This is exactly a monotone index binary choice model.

2.2. Offline Estimators

In the classical offline setting, a sample $\mathcal{D}_n = \{(x_{0,i}, X_i, Y_i)\}_{i=1}^n$ of fixed size n is collected once, and different strategies apply depending on the parameter of interest. In the following we write $Z_i(\theta) = x_{0,i} + X_i^\top\theta$.

When θ_0 alone is of interest, the maximum rank correlation (MRC) estimator of Han (1987) maximizes $\widehat{L}_{\mathrm{MRC},n}(\theta) = \frac{1}{n(n-1)}\sum_{i\ne j}\mathbf{1}\{Z_i(\theta) > Z_j(\theta)\}\mathbf{1}\{Y_i > Y_j\}$ over $\theta \in \varTheta$. The MRC estimator requires no estimate of F_0, but its criterion is discontinuous in θ and is evaluated over all $n(n-1)$ pairs; even the smoothed version of Horowitz (1992) must be re-optimized over the full set of pairs at every iteration.

When F_0 is also of interest, one can adopt the method of sieves (Chen 2007). Let $J = J_n$ be a sample-size-dependent sieve dimension, let $\Psi_J(\cdot) = (\psi_{J,1}(\cdot), \ldots, \psi_{J,J}(\cdot))^\top$ collect the basis functions, and define the sieve space $\mathcal{S}_J = \{\sum_{j=1}^J b_j\psi_{J,j}(\cdot) : b_1, \ldots, b_J \in \mathbb{R}\}$. With θ_0 known (or replaced by a preliminary estimate), F_0 is estimated by least squares of Y_i on $\Psi_J(Z_i(\theta_0))$. When both components are of interest, the sieve nonlinear least squares (NLS) estimator solves

$$\left(\widehat{\theta}_n, \widehat{\beta}_n\right) \in \arg\min_{\theta\in\Theta,\ \beta\in\mathbb{R}^{J_n}} \frac{1}{2n}\sum_{i=1}^n \left(Y_i - \Psi_{J_n}(x_{0,i} + X_i^\top\theta)^\top\beta\right)^2, \tag{2.2}$$

with $\widehat{F}_{\text{sieve-NLS},n} = \Psi_J(\cdot)^\top\widehat{\beta}_n$ (Ai and Chen 2003; Chen 2007).

All of these estimators are offline by design. Each criterion evaluation requires a complete pass through the data (through all $n(n-1)$ pairs for MRC); the NLS criterion (2.2) is nonconvex in (θ, β), so reliable computation requires many such passes from multiple starting points; and the arrival of new observations forces re-estimation from scratch. These features make the offline toolkit impractical for large samples and unusable for streaming data, and they motivate the online algorithms of Section 3.

2.3. Sieve NLS as a Structured Neural Network

The sieve learner in (2.2) can be interpreted as a single-hidden-layer neural network with economically meaningful restrictions. Spline bases admit representations through truncated power functions, so a sieve approximation of J basis functions can be written as

$$F_J(x_0 + X^\top\theta) = \sum_{j=1}^{J} \beta_j \psi_j(x_0 + X^\top\theta) = \sum_{\ell=1}^{M_J} a_\ell \sigma_q(x_0 + X^\top\theta - \kappa_\ell), \tag{2.3}$$

where M_J is the number of hidden units, $\sigma_q(t) = t_+^q$ is a ReLU-power activation, and the κ_ℓ are spline knots. The conditional mean then takes the network form $\mathbb{E}(Y|x_0, X) \approx \sum_{\ell=1}^{M_J} a_\ell\, \sigma_q(b_{\ell,-1} + b_{\ell,0}x_0 + X^\top b_\ell)$ subject to $(b_{\ell,0}, b_\ell^\top)^\top \propto (1, \theta^\top)^\top$: all hidden units share a common index direction. The hidden layer learns nonlinear transformations of the scalar economic index $x_0 + X^\top\theta$, while the output layer approximates the unknown monotone response curve. The restrictions are economically meaningful: the common first-layer direction is the interpretable index parameter θ_0; monotonicity of F_0 can be imposed through monotone spline bases (Chen 2007); and inference concerns the low-dimensional structural parameter rather than unrestricted network weights.

This interpretation also clarifies the role of sieve-dimension growth in online learning: increasing the sieve dimension enlarges the hidden layer, so previously learned weights must be *re-embedded* into the expanded network in a way that preserves the learned shape of the response curve. The next section develops an online algorithm for (2.2) built around exactly this recursive re-embedding. Richer architectures with multiple index directions (multinomial choice) or selection gates (sample selection) extend the same design and are developed in Section 7.1.

3. The Online Learning Algorithms

In the online environment, data arrive sequentially in small batches. Let $k = 1, 2, \cdots$ index a discrete time sequence. In period k the researcher observes a batch of B i.i.d. realizations of (x_0, X, Y) from model (2.1), where the batch size B is a fixed integer. Define $W_{i,k} = (x_{0,i,k}, X_{i,k}, Y_{i,k})$ for $i = 1, \ldots, B$ and $\mathcal{W}_k = \{W_{i,k}\}_{i=1}^{B}$. Throughout, $\{\mathcal{W}_k\}_{k=1}^{\infty}$ is

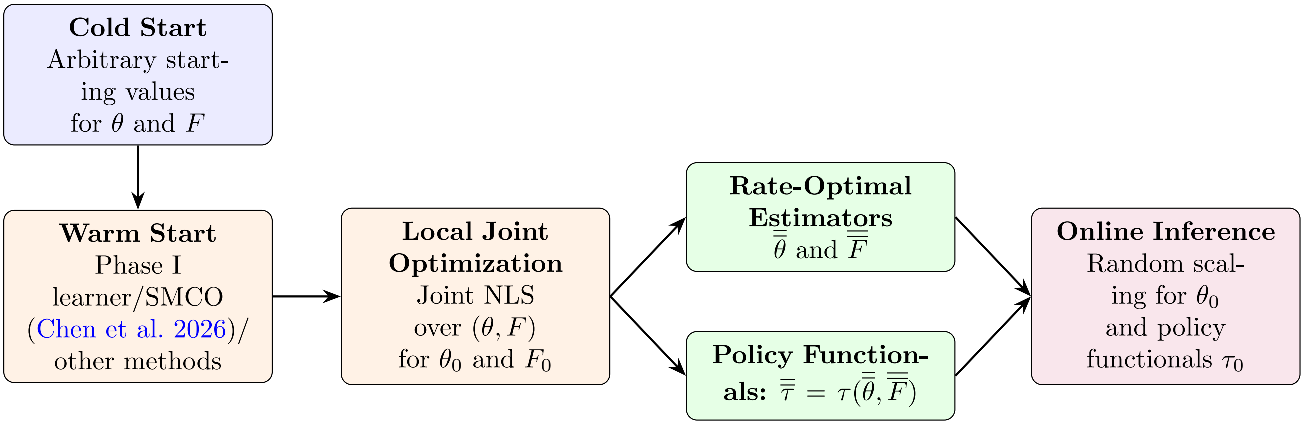


FIGURE 1. FLOW CHART OF SEMIPARAMETRIC ONLINE LEARNING

assumed i.i.d. across k; the analysis can be extended to dependent data under additional appropriate conditions.

Our online algorithm consists of two phases, summarized in Figure 1. The warm-start Phase I produces consistent estimators of (θ_0, F_0) from *arbitrary* initializations (Section 3.1 and Section 3.2). The refinement Phase II locally and jointly updates both components to attain optimal convergence rates (Section 3.3). The refinement trajectory is then the sole input to random-scaling inference (Section 3.4) and to online estimation and inference for policy functionals (Section 3.5).

A single averaging device is used throughout. For any sequence of iterates $a_1, a_2, \cdots$, the Polyak–Ruppert (PR) average (Ruppert 1988; Polyak and Juditsky 1992) is defined as $\overline{a}_N = N^{-1}\sum_{k=1}^{N} a_k$, computed by the recursion $\overline{a}_N = \frac{N-1}{N}\overline{a}_{N-1} + \frac{1}{N}a_N$. Using this recursion, only the current average requires storage, so PR averaging is itself fully online. All estimators reported by our algorithms are PR averages.

3.1. Online Warm Start for θ_0

Phase I estimates θ_0 without estimating F_0, through an online learner that is *globally stable*: it converges almost surely to θ_0 regardless of the starting point. Let $\widehat{\theta}_0$ denote an arbitrary initial guess and, for any $\theta \in \mathbb{R}^p$, define $Z_{i,k}(\theta) = x_{0,i,k} + X_{i,k}^{\top}\theta, \quad i = 1, \cdots, B$. For $k \geq 1$, the update from $\widehat{\theta}_{k-1}$ to $\widehat{\theta}_k$ is

$$\widehat{\theta}_k = \widehat{\theta}_{k-1} + \frac{\gamma_k}{h_k \cdot B(B-1)} \sum_{i_1 \neq i_2}^{B} \mathcal{K}\left(\frac{Z_{i_1,k}(\widehat{\theta}_{k-1}) - Z_{i_2,k}(\widehat{\theta}_{k-1})}{h_k}\right)(Y_{i_1,k} - Y_{i_2,k})(X_{i_1,k} - X_{i_2,k}), \tag{3.1}$$

where $\gamma_k > 0$ is the learning rate, $\mathcal{K}$ is a kernel function, and h_k is a bandwidth sequence.

Update (3.1) is related to the stochastic gradient of the MRC criterion of Han (1987) with one crucial modification: the ranking indicator $\mathbf{1}\{Y_{i_1,k} > Y_{i_2,k}\}$ is replaced by the response difference $Y_{i_1,k} - Y_{i_2,k}$. One of the core features of the algorithm is global stability: it almost surely converges to θ_0 from any initialization—which we establish formally in Theorem B.1 in Appendix B. The PR average estimator is defined as

$$\overline{\theta}_N = \frac{1}{N} \sum_{k=1}^{N} \widehat{\theta}_k. \tag{3.2}$$

The pseudo code of the warm-start estimation of θ_0 is given in Algorithm 1 in the Appendix.

3.2. Online Warm Start for F_0

We approximate F_0 by B-splines and estimate F_0 by the online sieve Least Squares. Throughout, we require that as the B-spline sieve dimension grows from J to $J+1$, one interior knot is inserted at the midpoint of the largest knot interval— preserving quasi-uniformity—while all existing knots are retained. The knot vectors are therefore nested, so the sieve spaces satisfy $\mathcal{S}_J \subseteq \mathcal{S}_{J+1}$ and the bases obey the exact refinement relation

$$\Psi_J(z) = R_{J,J+1}\Psi_{J+1}(z), \qquad \forall z, \tag{3.3}$$

where $R_{J,J+1} \in \mathbb{R}^{J\times(J+1)}$ is the transpose of the Boehm knot-insertion matrix (Boehm 1980); we adopt the convention $R_{J,J} = \mathbb{I}_J$.

We suppose that a sequence of index estimates $\check{\theta}_0, \check{\theta}_1, \cdots$ is available: $\check{\theta}_{k-1} = \theta_0$ if θ_0 were known, and otherwise any consistent online estimator, such as the PR average $\overline{\theta}_{k-1}$ of Section 3.1. Write $\check{Z}_{i,k} = Z_{i,k}(\check{\theta}_{k-1})$.

Given an initial coefficient $\widehat{\beta}_0$ of dimension J_0 and a pre-specified increasing sequence of sieve dimensions $\{J_k\}_{k\geq 1}$, the online sieve update is

$$\widehat{\beta}_k = R_{J_{k-1},J_k}^\top \widehat{\beta}_{k-1} + \frac{\eta_k}{B}\sum_{i=1}^{B}\Big(Y_{i,k} - \Psi_{J_k}(\check{Z}_{i,k})^\top R_{J_{k-1},J_k}^\top \widehat{\beta}_{k-1}\Big)\Psi_{J_k}(\check{Z}_{i,k}), \tag{3.4}$$

where $\eta_k > 0$ is a learning rate. When $J_k = J_{k-1}$, (3.4) is the mini-batch SGD step for the least squares projection of Y on $\mathcal{S}_{J_k}$. As in offline estimation, however, consistency requires the sieve dimension to grow, and enlarging a spline basis reconstructs the entire basis system: the coefficient vector for $\mathcal{S}_{J_{k-1}}$ is neither comparable to nor embedded in that for $\mathcal{S}_{J_k}$, so the update cannot simply be continued. Our solution is the *re-embedding* of $\widehat{\beta}_{k-1}$: the previous coefficient is re-initialized as $R_{J_{k-1},J_k}^\top \widehat{\beta}_{k-1}$, which, since $\Psi_{J_{k-1}}(z) = R_{J_{k-1},J_k}\Psi_{J_k}(z)$, satisfies

$$\Psi_{J_{k-1}}(z)^\top \widehat{\beta}_{k-1} = \Psi_{J_k}(z)^\top R_{J_{k-1},J_k}^\top \widehat{\beta}_{k-1}.$$

The previously fitted function is thus preserved *exactly* when embedded into the expanded sieve space, providing a stable starting point in the new space. The online estimator of F_0 and its PR average are

$$\widehat{F}_k(\cdot) = \Psi_{J_k}(\cdot)^\top \widehat{\beta}_k, \qquad \overline{F}_N(\cdot) = \frac{1}{N}\sum_{k=1}^{N}\widehat{F}_k(\cdot). \tag{3.5}$$

It is convenient to average at the coefficient level: setting $\overline{\beta}_0 = \widehat{\beta}_0$ and

$$\overline{\beta}_N = \frac{1}{N}\widehat{\beta}_N + \frac{N-1}{N}R_{J_{N-1},J_N}^\top \overline{\beta}_{N-1} \tag{3.6}$$

yields the closed form

$$\overline{\beta}_N = \frac{1}{N}\sum_{k=1}^{N} R_{J_k,J_N}^\top \widehat{\beta}_k, \qquad R_{J_k,J_N} = R_{J_k,J_{k+1}}R_{J_{k+1},J_{k+2}}\cdots R_{J_{N-1},J_N}, \tag{3.7}$$

so that $\overline{F}_N(\cdot) = \Psi_{J_N}(\cdot)^\top \overline{\beta}_N$. The procedure is summarized in Algorithm 2 in Appendix.

3.3. Phase II: Joint Local Refinement of θ_0 and F_0

Starting from the Phase I warm starts, Phase II jointly refines both components so that the resulting estimators attain optimal convergence rates. For sieve dimension J, define the sieve NLS loss

$$\mathcal{L}_J(\theta, \beta, W) = \frac{1}{2}\left(Y - \Psi_J\left(x_0 + X^\top\theta\right)^\top \beta\right)^2, \tag{3.8}$$

write $\omega = (\theta^\top, \beta^\top)^\top$ (whose dimension grows with the sieve dimension), and let $\mathcal{L}_J(\omega) \equiv \mathbb{E}[\mathcal{L}_J(\omega, W_{i,k})]$ denote the population loss. As in Section 3.2, the sieve dimension J_k grows along the path, and each update starts from the re-embedded previous iterate $\mathcal{R}^\top_{J_{k-1},J_k}\widetilde{\omega}_{k-1}$, where

$$\mathcal{R}_{J_1,J_2} = \begin{pmatrix} \mathbb{I}_p & 0 \\ 0 & R_{J_1,J_2} \end{pmatrix} \tag{3.9}$$

extends the re-embedding operator to the joint parameter. The natural stochastic gradient update is then

$$\widetilde{\omega}_k = \mathcal{R}^\top_{J_{k-1},J_k}\widetilde{\omega}_{k-1} - \frac{\xi_k \cdot \widehat{G}_k}{B}\sum_{i=1}^{B} \nabla_\omega \mathcal{L}_{J_k}\left(\mathcal{R}^\top_{J_{k-1},J_k}\widetilde{\omega}_{k-1}, W_{i,k}\right), \tag{3.10}$$

where $\xi_k > 0$ is the learning rate and $\widehat{G}_k$ is an (almost surely) positive definite conditioning matrix. For specific choices of the condition matrix in applications, see Section 4.

The critical complication is that $\mathcal{L}_J(\omega)$ is *not* globally convex in ω: unrestricted updates need not converge to the truth, which motivates confining (3.10) to a neighborhood over which the population loss has nondegenerate curvature. Let $Z_0 = x_0 + X^\top\theta_0$ and define

$$\omega_{J,0} = \left(\theta_0^\top, \beta_{J,0}^\top\right)^\top, \quad \beta_{J,0} = \left(\mathbb{E}\left[\Psi_J(Z_0)\Psi_J(Z_0)^\top\right]\right)^{-1}\mathbb{E}\left[\Psi_J(Z_0)F_0(Z_0)\right], \tag{3.11}$$

where $\beta_{J,0}$ is the $L_2(P_{Z_0})$ projection coefficient. We call $\omega_{J,0}$ the *sieve pseudo-true parameter* at dimension J. It does not need to be the exact joint finite-J population minimizer because the θ score at $\omega_{J,0}$ can contain a small approximation-bias term. We require

$$\inf_J \lambda_{\min}\left(\frac{\partial^2 \mathcal{L}_J(\omega_{J,0})}{\partial\omega\partial\omega^\top}\right) > 0, \tag{3.12}$$

so the population criterion has positive curvature at $\omega_{J,0}$ uniformly in J. As we formally show in Section D.3 in Appendix D, the curvature extends to the neighborhood

$$\Omega_J = \left\{\omega \in \mathbb{R}^{p+J} : \|\omega - \omega_{J,0}\|_2 \le C_\Omega J^{-2}\right\} \tag{3.13}$$

for a suitable constant $C_\Omega > 0$: the exponent 2 arises from the Lipschitz constant of the Hessian matrix within the local basin, so J^{-2} is the radius over which the Hessian of $\mathcal{L}_J$ moves by at most a small ϵ in operator norm. Since $\omega_{J_k,0}$ is unknown, we center a feasible neighborhood at the Phase I estimate $\widehat{\omega}_k = \left(\check{\theta}_{k-1}^\top, \{R_{J_{k-1},J_k}^\top \widehat{\beta}_{k-1}\}^\top\right)^\top$:

$$\widehat{\Omega}_k = \left\{\omega \in \mathbb{R}^{p+J_k} : \|\omega - \widehat{\omega}_k\|_2 \leq \frac{C_\Omega}{2} J_k^{-2}\right\}. \tag{3.14}$$

When $\widehat{\omega}_k$ converges to $\omega_{J_k,0}$ faster than J_k^{-2}, then $\widehat{\Omega}_k \subseteq \Omega_{J_k}$ eventually almost surely, so nondegenerate curvature holds on the feasible set. The feasible set is updated from the same incoming batch only after the center for that batch has been formed, and is retained for the next batch; hence the construction remains single-pass and $\widehat{\Omega}_k$ is measurable with respect to the past.

Letting $\Pi_A(a) = \arg\min_{a' \in A} \|a' - a\|_2$ denote the (well-defined) Euclidean projection onto a compact convex set A, the Phase II update is

$$\widetilde{\omega}_k = \Pi_{\widehat{\Omega}_k}\left[\mathcal{R}_{J_{k-1},J_k}^\top \widetilde{\omega}_{k-1} - \frac{\xi_k \cdot \widehat{G}_k}{B} \sum_{i=1}^{B} \nabla_\omega \mathcal{L}_{J_k}\left(\mathcal{R}_{J_{k-1},J_k}^\top \widetilde{\omega}_{k-1}, W_{i,k}\right)\right]. \tag{3.15}$$

The PR averages are computed along the way:

$$\overline{\overline{\theta}}_N = \frac{1}{N}\sum_{k=1}^{N} \widetilde{\theta}_k, \qquad \overline{\overline{\beta}}_N = \frac{1}{N}\sum_{k=1}^{N} R_{J_k,J_N}^\top \widetilde{\beta}_k, \qquad \overline{\overline{F}}_N(\cdot) = \Psi_{J_N}(\cdot)^\top \overline{\overline{\beta}}_N. \tag{3.16}$$

The pseudo code of Phase II procedure is summarized in Algorithm 3 in Appendix.

3.4. Random-Scaling Inference

Inference on θ_0 uses the random-scaling method of Lee et al. (2022), whose only input is the trajectory of Phase II PR averages $\overline{\overline{\theta}}_1, \overline{\overline{\theta}}_2, \cdots$. Define the self-normalizing matrix

$$V_N = \frac{1}{N^2}\sum_{k=1}^{N} k^2 \left(\overline{\overline{\theta}}_k - \overline{\overline{\theta}}_N\right)\left(\overline{\overline{\theta}}_k - \overline{\overline{\theta}}_N\right)^\top, \tag{3.17}$$

which is updated recursively from running sums. A 95% confidence interval for the j-th component of θ_0 is

$$\overline{\overline{\theta}}_{N,j} \pm 6.747 \cdot \sqrt{[V_N]_{jj}/N}, \tag{3.18}$$

where 6.747 is the two-sided 95% critical value of the associated mixed-normal limit. No variance component needs to be estimated, so inference adds essentially no computational cost beyond estimation itself. Although random scaling was proposed for parametric online least squares, its justification requires only the validity of an FCLT for the averaged trajectory; the FCLTs established in Section 5.2.1 therefore extend it to the present semiparametric setting, including the policy functionals of Section 3.5.

3.5. Online Estimation and Inference for Policy Interventions

The joint trajectory also supports online estimation and inference for policy-relevant functionals of (θ_0, F_0). As a leading example, consider the *average marginal effect*

$$\tau_0 = \mathbb{E}\left[\frac{\partial \mathbb{E}(Y|x_0, X)}{\partial X}\right] \equiv \mathbb{E}\left[\nabla_z F_0(Z_0)\, \theta_0\right], \tag{3.19}$$

whose j-th component measures the average change in the outcome induced by a one-unit shift in the j-th feature. Given the Phase II iterates $\{(\widetilde{\theta}_k, \widetilde{\beta}_k)\}$, we propose the plug-in estimator and its PR average

$$\widetilde{\tau}_k = \frac{1}{B}\sum_{i=1}^{B}\Big(\nabla_z \Psi_{J_{k-1}}(Z_{i,k}(\widetilde{\theta}_{k-1}))\Big)^{\top} \widetilde{\beta}_{k-1}\widetilde{\theta}_{k-1}, \quad \overline{\overline{\tau}}_N = \frac{1}{N}\sum_{k=1}^{N}\widetilde{\tau}_k. \tag{3.20}$$

Similar to the online estimators of θ_0 and F_0, the sequence $\overline{\overline{\tau}}_1, \overline{\overline{\tau}}_2, \cdots$ is also evaluated online, and we also show in Section 5.2.2 that it attains $1/\sqrt{N}$-consistency and satisfies an FCLT. Consequently, the random-scaling inference applies to it directly: we can simply replace $\overline{\overline{\theta}}$ with $\overline{\overline{\tau}}$ in (3.17) and (3.18). We point out that the FCLT-based fast inference is particularly attractive because the marginal effect is of the main concern in most applications, and most importantly, as we show in the appendix, the asymptotic variance of $\overline{\overline{\tau}}_N$ is complicated, making plug-in inference computationally demanding.

4. Two Empirical Illustrations

4.1. Application to a Synthetic Trade Data

We first illustrate our online estimation algorithm using the bilateral-trade panel synthesized from the data set of Helpman, Melitz, and Rubinstein (2008) (hereafter HMR). Section 4.1.1 describes the data and estimation design. Section 4.1.2 presents a specification check under a calibrated Probit data-generating process. Section 4.1.3 applies the estimator to the resampled panel outcomes.

4.1.1. Data, Variables, and Empirical Design

The original data set in Helpman et al. (2008) covers bilateral trade flows among 158 countries over 1980–1989, for a total of $n = 248{,}060$ country-pair-year observations. Positive trade is recorded for 44.6% of observations. The outcome is

$$Y_{ij,t} \;=\; \mathbf{1}\{\text{country } i \text{ records positive exports to country } j \text{ in year } t\}, \tag{4.1}$$

which is modeled as a binary choice process. The normalized regressor is $x_0 = -\ln(\text{distance}_{ij})$, the negative natural logarithm of physical distance between countries i and j. The remaining covariates are nine standard gravity variables—shared land border, island status, landlocked status, common legal origin, common language, colonial ties, currency union (CU), free trade agreement (FTA), and shared religion—together with 157 exporter, 157 importer, and 9 year fixed effects[2]. The island and landlock indicators take the value one if at least one of the two countries in the pair is, respectively, an island or landlocked, and zero otherwise.

We focus on the estimated coefficients of the nine gravity variables and the finite-difference average marginal effects on the conditional probability, which are defined as

$$\text{AME}_j \;=\; \mathbb{E}[F_0(Z_{0,-j} + \theta_{0,j}) - F_0(Z_{0,-j})]\,, \qquad j \in \{\text{CU}, \text{FTA}\}, \tag{4.2}$$

where $Z_{0,-j}$ is the index evaluated at $x_j = 0$, $\theta_{0,j}$ is the associated coefficient, and F_0 is the link: the Gaussian CDF for Probit, the logistic CDF for Logit, and the estimated link for the semiparametric estimator. We focus on the marginal impacts of two regressors: CU

[2] We drop 1 exporter dummy, 1 importer dummy, and 1 year dummy for identification.

and FTA. These two marginal impacts are particularly interesting because they quantify how belonging to the same currency union or the same free-trade agreement affects the probability of trade between two countries.

The empirical design comprises two streaming exercises. In the first exercise, synthetic outcomes are repeatedly generated from a Probit data-generating process whose parameters are calibrated to the full-sample Probit fit. In this setting, the true coefficient vector and marginal effect are both known, so this exercise serves as a robustness check: when Probit is the true data generating process, we show that online learning precisely recovers all the targets. In the second exercise, the estimator is applied to the panel outcomes directly. We randomly resample data from the original panel, and no parametric assumption is imposed on the link. In this case, we show that the semiparametric estimates differ from Probit or Logit for some targets, indicating potential model misspecification by Probit or Logit as well as gains from semiparametric modeling.

Implementation Details. *Number of updates.* For both exercises, we consider a total of $N = 10^7$ updates, with a Phase I global learner of $N_1 = 10^6$ updates and a Phase II rate-optimal stage of $N - N_1 = 9 \times 10^6$ updates. We consider batch size $B = 25$.

Choices of sieves. To simplify the implementation, we use degree 3 (i.e., order 4, or cubic) B-splines with uniformly spaced knots, and fix the sieve dimension at its initial value $J_0 = 50$ throughout updates. Since the true unknown link function F_0 is assumed to have approximation error rate s with $s > 7/2$[3], with cubic B-spline sieve dimension fixed at $J_0 = 50$ in this application, the approximation error rate is of order $50^{-4} \approx 1.6 \times 10^{-7}$. This is negligible compared with the width of the confidence intervals reported below, which are of order 10^{-3}.

Warm-start updates. The online warm-start learner of θ performs kernel-based updates, initialized at the origin $\mathbf{0}_p$. We use the sixth-order Epanechnikov kernel and choose bandwidth $h_k = c_k \cdot k^{-\alpha_h}$ with $\alpha_h = 0.1$ and $c_k = 1.414\,\widehat{\sigma}_{Z,k}$, where $\widehat{\sigma}_{Z,k}$ is the batch-based sample standard deviation of $\widehat{Z}_{i,k} = x_{0,i,k} + X_{i,k}^\top \widehat{\theta}_{k-1}$. We choose a constant learning rate $\gamma_k = 0.02$ for aggressive search in the first stage[4] and start tracking the PR average after $N_0 = 5 \times 10^5$

[3]By approximation error rate s, we mean that $\|F_0 - g\|_\infty \leq CJ^{-s}$ for some $g \in \mathcal{S}_J$, see Appendix C and Appendix D.

[4]A constant learning rate allows the algorithm to quickly locate a neighborhood of the true parameter, at the cost of nonvanishing noise in the iterates. Since Phase I only supplies the warm start in the application, this aggressive search does not affect the local refinement and inference in Phase II.

updates. The warm-start of sieve coefficients are constructed by

$$\widehat{\beta}_{N_1} = \left[\sum_{k=N_0+1}^{N_1}\sum_{i=1}^{B}\widetilde{\Psi}_{J_0}(\overline{Z}_{i,k})\widetilde{\Psi}_{J_0}(\overline{Z}_{i,k})^\top\right]^{-1}\left[\sum_{k=N_0+1}^{N_1}\sum_{i=1}^{B}\widetilde{\Psi}_{J_0}(\overline{Z}_{i,k})Y_{i,k}\right], \tag{4.3}$$

where J_0 is the initial sieve dimension, $\overline{Z}_{i,k} = x_{0,i,k} + X_{i,k}^\top\overline{\theta}_{k-1}$, and $\widetilde{\Psi}_J(\cdot) = \Psi_J(\frac{2}{\pi}\arctan(\cdot))$. Note that $\frac{2}{\pi}\arctan(\cdot)$ maps $\mathbb{R}$ to $(-1,1)$, so spline basis functions with bounded support can be used to approximate a function with unbounded support.

Local refinements by conservative learning rates. The local refinement starts from the last-step PR average of the first stage, i.e. $k = N_1 + 1$. In this stage, specifying the projection space would require an additional choice on the tuning parameter C_Ω. To avoid explicit construction of projection space (and choosing tuning parameter C_Ω), we consider Phase II learning rate $\xi_k = 2\cdot(k+10^5)^{-0.55}$. In this case, $\xi_1 = 0.0036$, and consequently the initial updates of second stage are conservative. This makes the second stage refinement local even without explicit projection. We also point out that this construction of learning rate only shifts its scale without changing its decaying rate.

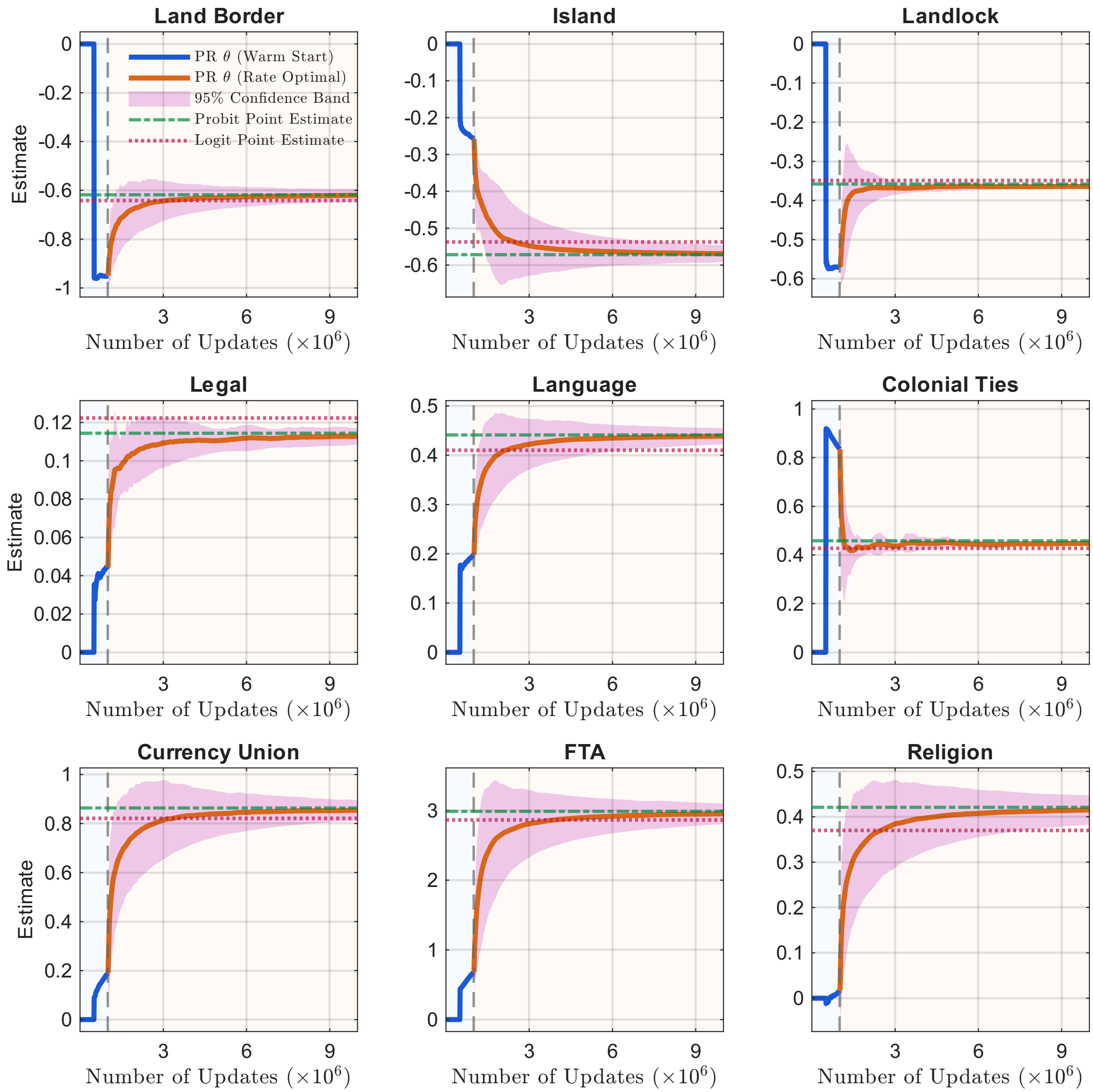
Choice of conditioning matrix $\hat{G}_k$. To stabilize the updates, we precondition the gradient with an inverse Hessian estimate that is updated recursively across Phase II, based on the cumulative sample Hessian evaluated along the estimated trajectory[5]. Since the parameter estimates are consistent, this matrix gain converges almost surely to the inverse population Hessian at the population truth; by standard results for averaged stochastic approximation with convergent matrix gains (Polyak and Juditsky 1992), the first-order asymptotic distribution of the PR average is invariant to such preconditioning; in particular, it neither affects the asymptotic properties of the PR average estimators nor the validity of the random-scaling-based inference.

4.1.2. Semiparametric Online Learning under a Calibrated Probit DGP

Let $(\beta^*_{-1}, \beta^*_0, \beta^*)$ denote the Probit maximum-likelihood estimates on the original data set. At each update k, a batch of size $B = 25$ of covariate vectors $(x_{0,k}, X_k)$ is drawn with replacement from the empirical distribution, and the outcome is generated as

$$Y_{i,k} \;=\; \mathbf{1}\{\beta^*_{-1} + \beta^*_0 x_{0,i,k} + X_{i,k}^\top\beta^* - u_{i,k} \geq 0\}, \qquad u_{i,k} \sim \mathcal{N}(0,1), \tag{4.4}$$

[5]We recursively update the Hessian matrix, update the inverse Hessian matrix every 1000 iterations. We also impose a small ridge term $10^{-4}\mathbb{I}$ to ensure invertibility.

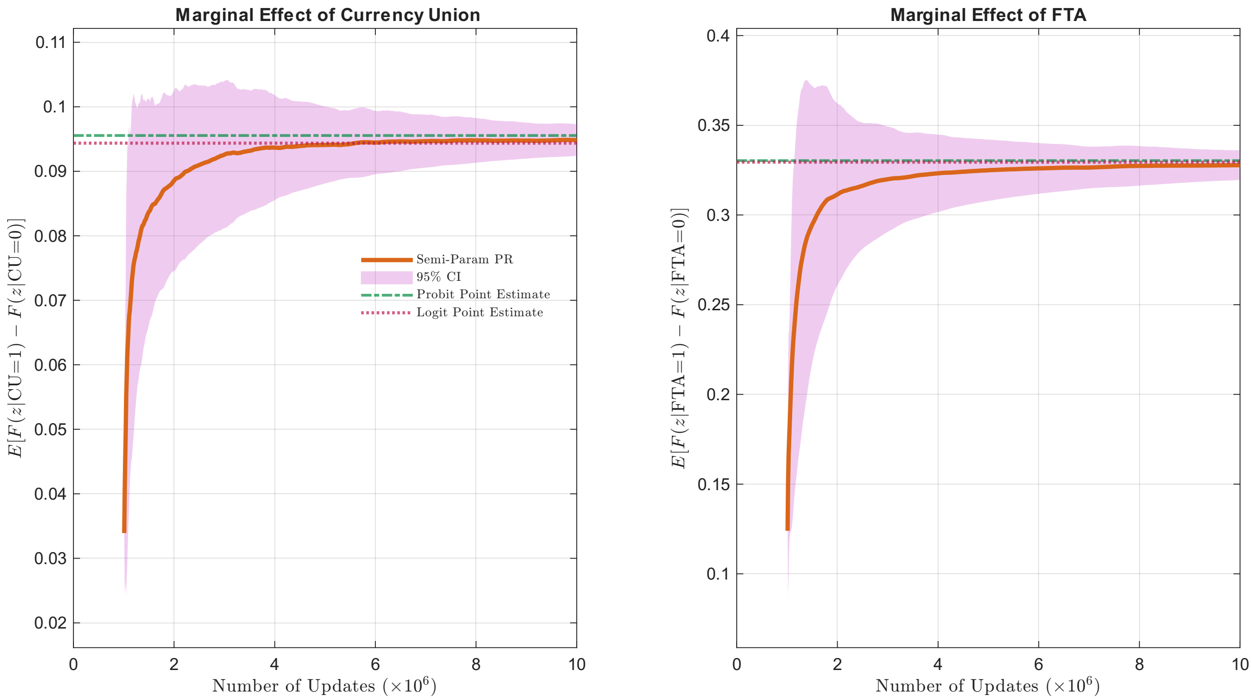


Note: The truth coefficients are fixed to the full-sample Probit estimators, so a valid semiparametric online estimation algorithm should recover the Probit benchmarks.

Figure 2. Synthetic Data: Parameters

with $u_{i,k}$ independent across i and k. The true normalized coefficient is $\theta^* = \beta^*/\beta_0^*$, and a correctly specified estimator should recover it.

Figure 2 reports the PR average coefficient paths across both phases, together with 95% random-scaling confidence bands and the Probit and Logit benchmarks. Phase I brings the



Note: The true AMEs are given by (4.2), where the link is Probit. A valid semiparametric online estimation algorithm should recover the Probit benchmarks.

FIGURE 3. SYNTHETIC DATA: MARGINAL EFFECTS

estimates close to their limits; Phase II delivers smooth convergence with rapidly shrinking bands. Across all nine gravity coefficients, the semiparametric paths track the Probit benchmark, which is the truth, to within 3% of the true value at the end of Phase II. The agreement is uniformly tight: the semiparametric coefficient on FTA is 2.953, against a Probit value of 2.988; on currency union, 0.854 against 0.864; on shared land border, -0.620 against -0.618; and on shared religion, 0.415 against 0.421. Figure 3 reports the corresponding marginal-effect paths. The semiparametric AMEs (0.0948 for CU and 0.3278 for FTA) coincide with the Probit benchmarks (0.0955 and 0.3303) to within sampling noise. The above results together confirm that, when the parametric specification is correct given the covariate distribution, the online procedure recovers both the true coefficient vector and the true marginal effects.

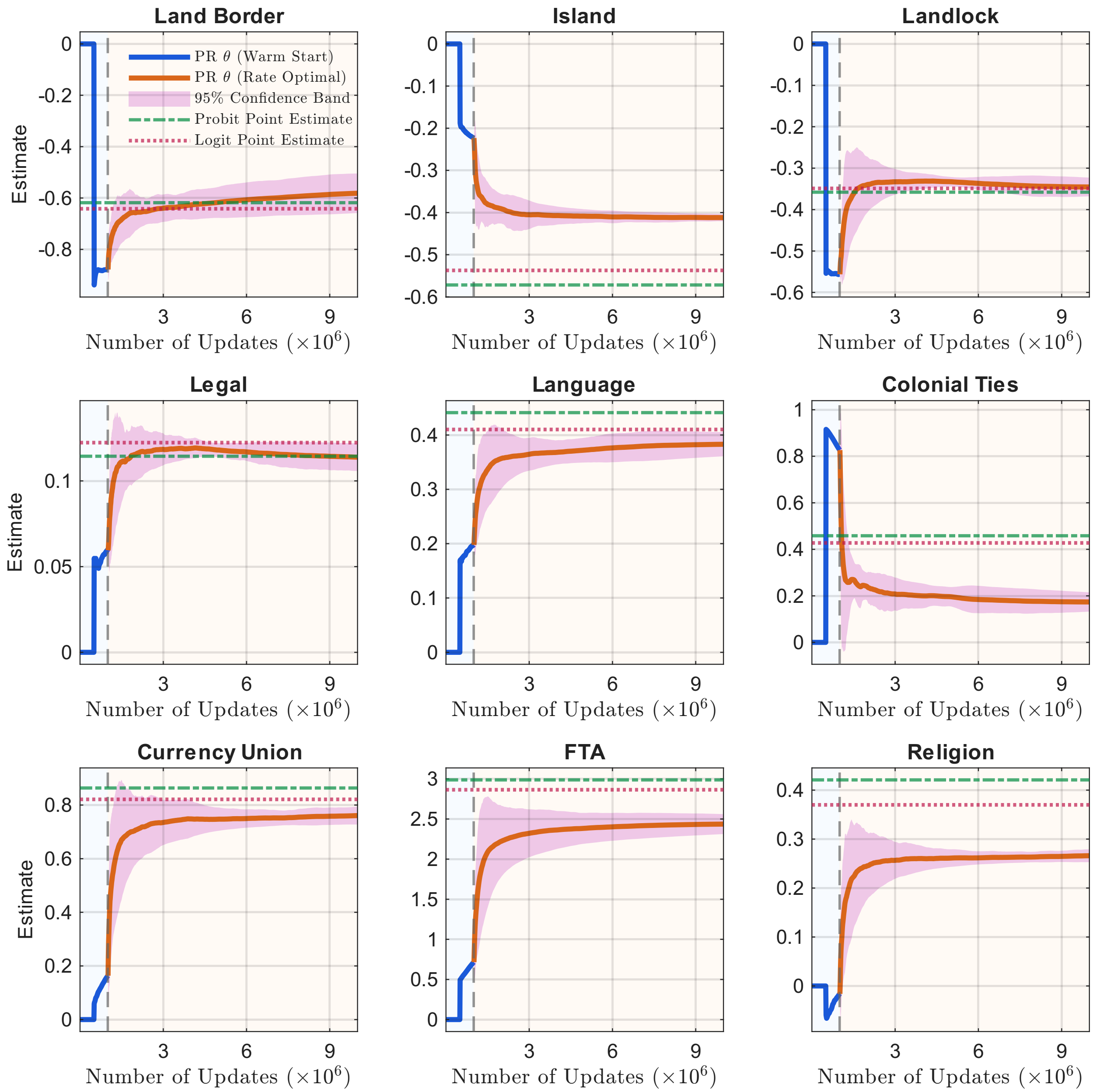


FIGURE 4. REAL DATA: PARAMETERS

4.1.3. Application to the HMR Panel

In this section, we treat the 248,060 observation points as the sample space[6]. At each update, a batch of size $B = 25$ is drawn with replacement from the joint empirical distribution of (x_0, X, Y), yielding an i.i.d. stream with respect to the empirical measure. In this setting,

[6]In this case, the online confidence band should be interpreted as the quantification of the uncertainty from random subsampling conditioned on the original 248,060 data points. A narrower band indicates a more accurate estimate of the full-sample counterpart.

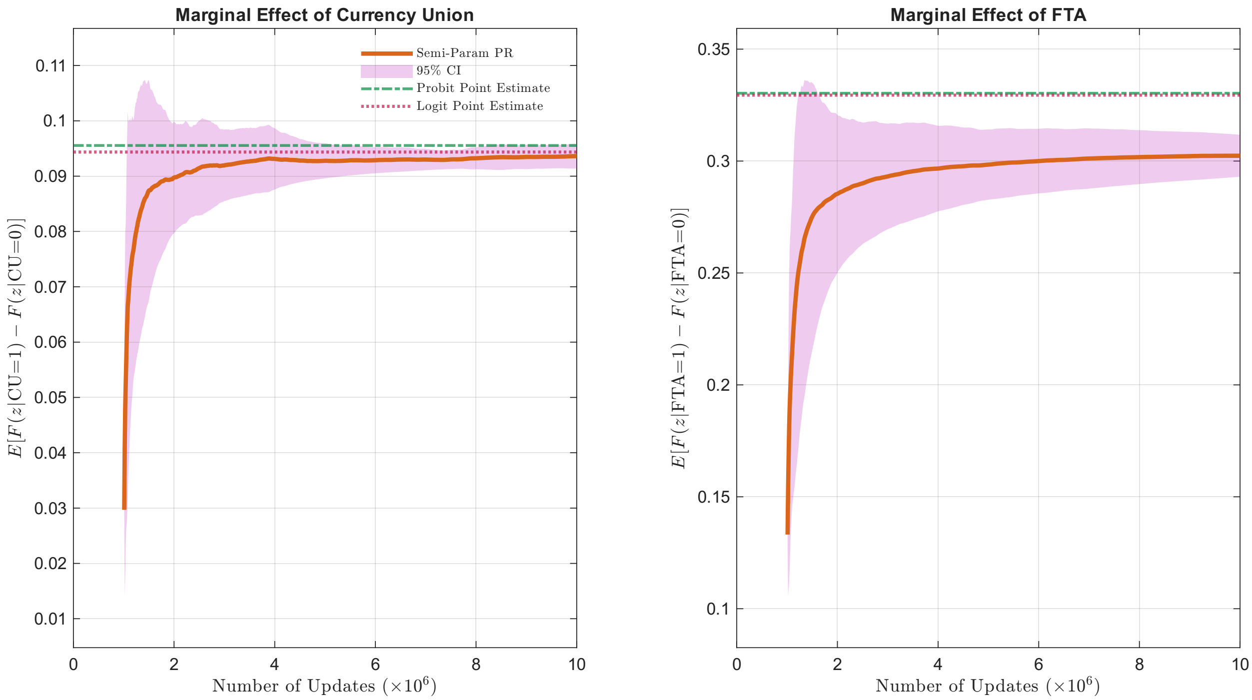


FIGURE 5. REAL DATA: MARGINAL EFFECTS

no parametric assumption is imposed on the link.

Figure 4 reports the coefficient paths with 95% random-scaling confidence bands. The convergence pattern mirrors that of Figure 2: Phase I lifts the estimates away from zero, and Phase II settles them at stable values with shrinking bands. The comparison with the parametric benchmarks, however, splits into two groups. For three of the nine coefficients—shared land border, landlocked status, and common legal origin—the semiparametric estimates agree closely with Probit and Logit, with deviations of less than 5%. For the remaining six—island status, common language, colonial ties, currency union, free trade agreement, and shared religion—the semiparametric estimates display sizeable departures from parametric ones. The largest proportional departure is on colonial ties, where the semiparametric estimate is 0.174 against a Probit value of 0.458—an attenuation of more than 60%. Sizeable departures also appear on shared religion (0.266 versus 0.421, a 37% attenuation), island status (-0.412 versus -0.572, a 28% attenuation), and FTA (2.437 versus 2.988, an 18% attenuation). In all six cases, the semiparametric coefficient is smaller in magnitude than the Probit benchmark, with the Logit value lying between them.

Figure 5 reports the marginal-effect paths for currency union and FTA. For currency union, the three estimators agree closely on the probability scale: the semiparametric AME is 0.094, compared with 0.096 for Probit and 0.094 for Logit. For FTA, however, the semiparametric

AME of 0.302 is significantly distinguishable from the Probit AME of 0.330, a departure of roughly 3 percentage points on the probability scale that lies well outside the bounds of sampling variability. While this gap may seem small in magnitude, it is economically meaningful: relative to the unconditional probability of a trade relationship of 44.6%, misspecification of this order can materially distort the implied policy conclusions.

Taken together, the two exercises demonstrate that the online semiparametric estimator is well-calibrated under correct parametric specification and is flexible enough to detect link misspecification when it arises. The second exercise reveals a systematic pattern: coefficient-level deviations from Probit are largely concentrated on the institutional and informational gravity variables—colonial ties, language, religion, currency union, FTA—rather than on the geographic ones—land border, landlock—although the island estimate by the Gaussian link is way off from the one estimated by our flexible sieve link as well. The probability-scale marginal effect of FTA inherits part of this distortion: a researcher relying on Probit to quantify the effect of FTA membership on the probability of trading would overstate the average partial derivative by roughly 3 percentage points, a discrepancy that is small in absolute terms but economically non-negligible at the trade-policy margins where such estimates are typically applied.

4.2. Application to the Stream of Stanford Open Policing Data

This section illustrates our estimation approach using the North Carolina State Patrol records from the Stanford Open Policing Project (Pierson et al. 2020)[7], a large administrative stream of traffic stops that is online in nature. The remainder of this section is arranged as follows: Section 4.2.1 describes the data and estimation design, and Section 4.2.2 applies the estimator to the stop outcomes directly, processing the records in their recorded arrival/online order.

4.2.1. Data, Variables, and Empirical Strategy

The Stanford Open Policing Project provides rich information on nationwide traffic stops. We choose the data set from North Carolina State Patrol between December 1999 and December 2015. The original data contains 20,286,645 stop-level observations. After data cleaning[8], we are left with n = 19,234,514 data points, which essentially constitute a large

[7]The data are available at https://openpolicing.stanford.edu/data/.

[8]For data cleaning, we drop the observations with missing values and also stops in which the driver's race is Asian or Pacific islander/unknown/other. These account for 1.11%, 1.20%, and 0.79% of full data set

online data stream.

The outcome of interest is whether a search is conducted following the stop by the officer, i.e.,

$$Y_i \;=\; \mathbf{1}\{\text{a search is conducted at stop } i\}, \tag{4.5}$$

which is modeled as a binary choice process. The regressors we consider include standardized driver's age (mean zero and variance 1), indicator for female drivers (male driver as benchmark), indicators for White and Black drivers (Hispanic driver as benchmark), and the recorded reasons for stop including violations of speed limit, vehicle regulatory, seat-belt, vehicle-equipment, investigation, safe-movement, and other-motor-vehicle violations—with stop-light/sign violation as the omitted category. This yields a total of $p = 11$ regressors. The normalized regressor x_0 is chosen as the negative standardized age; see the following rolling-window analysis for more discussion.

In this section, we mainly focus on the AME of three demographic regressors, White, Black, and Female, on search probabilities. These effects quantify how driver's demographics may shift the probability that a stop results in a search, relative to Hispanic, male, and stop-light violation baselines. The AME of these three dummy regressors can be computed following (4.2). Also, following the previous section, when we evaluate the estimation results, we also benchmark against Probit and Logit link estimates based on the full sample. However, we emphasize that our semiparametric estimation and the corresponding inference in this example are online, so that each new recorded data point is processed only once.

Data Stability and Estimation Sample Traffic stops can be plausibly regarded as independent across cases: individual stops are initiated by different officers encountering different drivers at dispersed locations across the state, and the search decision in one stop is not likely to affect the decision in another. However, what is far less plausible over the 1999–2015 sample window is distributional stability. The composition of stops and the propensity to search respond to slowly moving institutional forces including changes in enforcement priorities and departmental policy, personnel turnover, budget cycles, and the documented statewide decline in search rates over the 2000s. As a result, the joint distribution of stop characteristics and search outcomes may drift across the sample even while individual stops remain essentially independent draws from the prevailing regime.

respectively. We further drop the observations recorded as checkpoints and as driving-while-impaired investigations, whose proportions are 1.18% and 0.95% respectively. We drop these small-proportion categories because they are too sparse to support separately estimated coefficients.

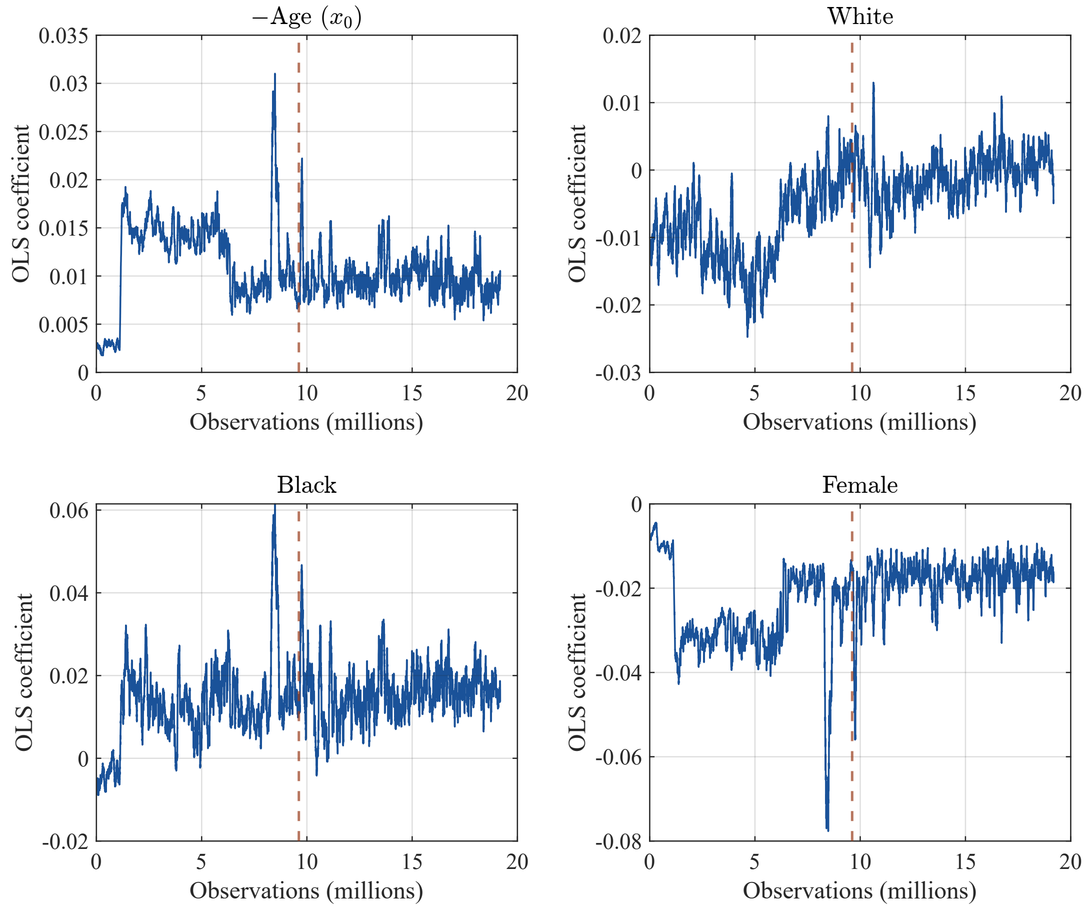


FIGURE 6. OLS ROLLING-WINDOW ESTIMATION RESULTS

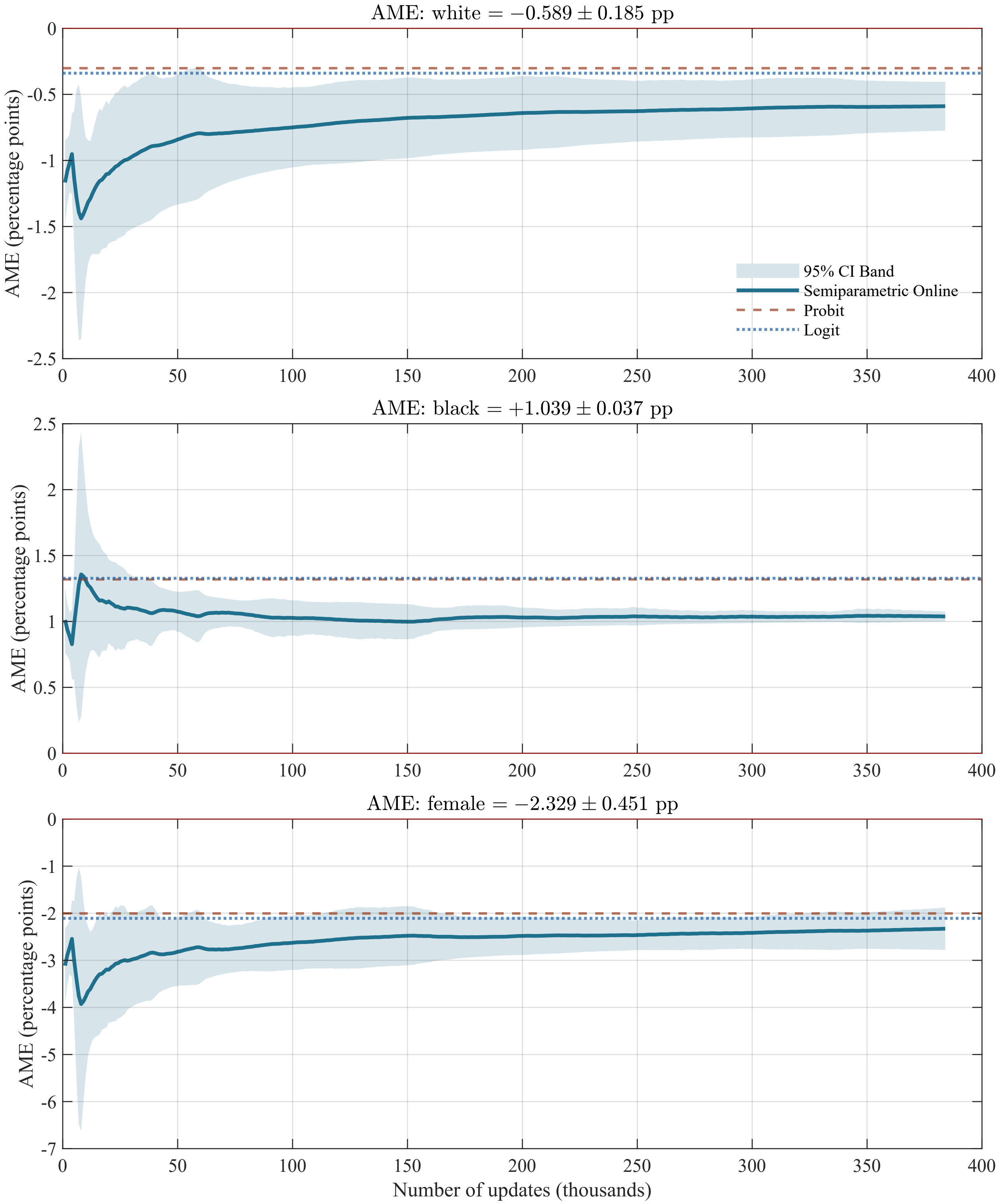
To provide a more detailed diagnostic analysis for the potential structural instability associated with the original data, we report in Figure 6 the rolling window least squares estimation results for the North Carolina sample in recorded order. Starting from the initial observation in 1999, we iteratively conduct OLS estimation based on regressors mentioned above including a constant term, using a sequence of rolling windows each with 10^5 observations. The window moves forward by 10^3 observations each time, so that the adjacent windows overlap heavily by 99% observations. This allows us to detect structural shifts that are even smooth across time. In Figure 6, we report the estimation results for four regressors: $-$age, white, black, and female.

Figure 6 clearly indicates that the first half of the sample exhibits pronounced structural instability. First of all, a discrete structural change happens to coefficients of negative age, black and female at around 1.1 million-th data point. Such shift has strong persistence and lifts up the magnitude of the above three coefficients. Another surge occurs to the three coefficients at around 8.4 million-th point, again inflating the coefficient magnitude, although with weak persistence. The coefficient of white features smooth structural breaks: it drifts gradually downward to about -0.025 by the 4.6 million-th observation and then climbs back to fluctuate around zero. Given all the observations, if we use the data set including a large proportion of the first half, the identical distribution condition may be violated, and the corresponding estimators and the confidence intervals may lose their statistical properties.

However, when we look at the second half of the data set, which starts from June 2009, we can clearly see that the rolling-window estimates behave differently: no coefficient path displays a sustained level shift. This result is consistent with local heterogeneity in the composition of stops, rather than with regime change.

In light of this evidence, and to align the estimation sample with the i.i.d. sampling framework we imposed before, we conduct online estimation and inference only on the second half of the sample. The first half serves solely as a training pool, supplying the kernel-based warm start and the bootstrap burn-in described below, which does not contribute to the reported estimators or confidence intervals. Note that in this case, the warm start is computed under a mixture of regimes, so its population target may differ from the truth governing the post-2009 stream. This is immaterial for our procedure provided that the shift between the two regimes is moderate. All reported estimates and confidence intervals are driven exclusively by the post-June-2009 stream. The full-sample Probit/Logit estimators are also conducted based on the second half of the data. Finally, note that the rolling-window estimation results of the coefficient of $-$age is strictly positive across all windows. This also supports the use of negative standardized age as x_0.

Implementation of the Algorithm. In this application, we also choose batch size $B = 25$. However, we slightly change the way of the warm-start construction. In particular, we fix the first half of the records and iteratively draw random samples (of batch size 25) from this offline data set with replacement. We then apply the online learning algorithms to these randomly generated batches with $N = 3 \times 10^6$, $N_1 = 10^6$ and $N_0 = 5 \times 10^5$. After the construction of the warm start, we shift to the online estimation with the second half of the data, where the rate-optimal phase proceeds with the remaining 9.6 million records strictly in their recorded order and each data point is processed only once. This roughly leads to 384 thousand updates. Given the smaller number of total updates, we choose a larger second phase learning rate $\xi_k = 5(k + 10^5)^{-0.55}$. Other choices of tuning parameters are identical to those in the previous section.



Note: Warm start estimation takes around 2 minutes, while online local refinement takes 17 seconds. The applications were implemented in MATLAB R2024b and executed on a laptop with an Intel Core Ultra 9 185H processor (16 cores, 2.5 GHz) and 64 GB of RAM.

Figure 7. Semiparametric Online Estimation Results for Average Marginal Effects: Stanford Policing Data

4.2.2. Online Estimation and Results for AME

The warm starts take roughly 2 minutes, while the subsequent online refinements take only 17 seconds – processing roughly half a million of data points per second. Figure 7 reports the PR averages of the AME paths on the probability scale from Phase II. The effects are economically meaningful against the low baseline unconditional search rate of 2.32% (second half of the data): relative to a baseline (hispanic male) driver, black drivers face an average marginal effect of +1.04 percentage points, women −2.33 percentage points, and white drivers −0.59 percentage points; all three random-scaling 95% confidence intervals exclude zero at the last update. Here we caution against a causal reading: differences in search rates alone do not directly identify discrimination. As emphasized in the policing literature (Simoiu et al. 2017; Pierson et al. 2020), such benchmark comparisons are subject to infra-marginality and may reflect non-discriminatory factors. Consequently, we should treat the estimated effects as descriptive.

The central finding of Figure 7 is that the semiparametric estimates differ systematically from the parametric benchmarks: for two of the three average marginal effects, the Probit and Logit point estimates lie outside the semiparametric 95% confidence intervals. The parametric models place the black–hispanic search gap at roughly 1.32 percentage points, overstating the semiparametric estimate of 1.04 (±0.037) by over 25 percent; for the white–hispanic gap the distortion runs the other way, with parametric estimates of about −0.3 percentage points amounting to roughly half of the semiparametric −0.59 (±0.19).

The above discussion implies that link misspecification does not merely rescale the estimated disparities, it compresses one racial gap while inflating another. Since all specifications are estimated based on the same post-June-2009 sample with the same covariates, these discrepancies should be attributed to the shape of the link function itself, the object that parametric practice fixes by assumption. This underscores the value of flexible link estimation for measuring disparities and, at this scale of sample size, our online procedure delivers it at essentially no additional computational cost.

5. Statistical Properties

5.1. Global Almost Sure Consistency of Phase I Algorithms

Define $\Delta\theta = \theta - \theta_0$ for any $\theta \in \mathbb{R}^p$. The following lemmas shows that our Phase I online algorithms generate paths that converge to the true values (θ_0, F_0) almost surely.

Lemma 5.1. *Let $\overline{\theta}_N$ be the PR averages of the Phase I algorithm for θ_0. If Condition 1–4 in Appendix B hold, then: $\|\Delta\overline{\theta}_N\|_2 \rightarrow_{a.s.} 0$, and see Theorem B.1 in Appendix B for its almost sure rate of convergence.*

Lemma 5.2. *Let $\overline{F}_N$ be the PR averages of the Phase I algorithm for F_0. If Condition 5–9 in Appendix C hold, then: $\|\overline{F}_N - F_0\|_\infty \rightarrow_{a.s.} 0$, and see Theorem C.1 in Appendix C for its almost sure rate of convergence.*

Two points are worth mentioning. First, the above convergence results are global in the sense that the results are valid regardless of the choice of the initial starting points. It is known to be very difficult to find a global solution to non-convex optimization problem over a high-dimensional parameter space (Chen et al. 2026). Consequently, the above global consistency result is not only theoretically important but also practically crucial for semiparametric index models with many covariates. Second, the global convergence is almost surely so is path-wise. This guarantees the validity of the warm starts for almost all paths of data, which is crucial for online estimation with single passage of data.

5.2. Rate-Optimal Statistical Properties of Phase II Algorithms

5.2.1. Automatic Orthogonalization and Rate-Optimality

This section formally studies the asymptotic properties of the local online learner $\overline{\overline{\omega}}_N$. Recall that $Z_0 = x_0 + X^\top\theta_0$. Define $\mu_0(\theta, z) = \mathbb{E}(X|x_0 + X^\top\theta = z)$,

$$\mathbb{M}_\theta = \mathbb{E}\left[(\nabla_z F_0(Z_0))^2 (X - \mu_0(\theta_0, Z_0))(X - \mu_0(\theta_0, Z_0))^\top\right], \tag{5.1}$$

$$P_J = \left(\mathbb{E}\left[\Psi_J(Z_0)\nabla_z F_0(Z_0)X^\top\right]\right)\left(\mathbb{E}\left[(\nabla_z F_0(Z_0))^2 XX^\top\right]\right)^{-1}, \tag{5.2}$$

and

$$\mathbb{M}_{\beta,J} = \mathbb{E}\left[\left(\Psi_J(Z_0) - P_J\nabla_z F_0(Z_0)X\right)\left(\Psi_J(Z_0) - P_J\nabla_z F_0(Z_0)X\right)^\top\right]. \tag{5.3}$$

We consider the learning rate ξ_k and the sieve dimension J_k to be chosen as

$$\xi_k = \xi_0 k^{-\alpha_\xi}, \quad J_k = [J_0 k^{\alpha_\dagger}], \tag{5.4}$$

where ξ_0, α_ξ and $\alpha_\dagger$ are all positive constants, and J_0 is a positive integer. Here $\frac{1}{2} < \alpha_\xi < 1$ controls the decreasing rate of the learning rate, while $\alpha_\dagger$ controls the increasing speed of sieve dimension J_k with the number of updates k. Detailed conditions on α_ξ and $\alpha_\dagger$ can be found in Appendix D.

Recall that $\Delta\theta = \theta - \theta_0$ for any $\theta \in \mathbb{R}^p$. Similarly, for any $\beta \in \mathbb{R}^J$, define $\Delta\beta = \beta - \beta_{J,0}$, where $\beta_{J,0}$ is the pseudo true sieve coefficient at dimension J. We have the following lemma that provides the asymptotic linear representation of the PR average estimators, with almost sure convergence rates for remainder terms.

Lemma 5.3. *Let Condition 5–7 in Appendix C and Condition 10–14 in Appendix D hold. Then:*

$$\Delta\overline{\overline{\theta}}_N = \frac{1}{N}\sum_{k=1}^{N}\frac{1}{B}\sum_{i=1}^{B}\varepsilon_{i,k}\mathbb{M}_\theta^{-1}\nabla_z F_0(Z_{0,i,k})\left(X_{i,k} - \mu_0(\theta_0, Z_{0,i,k})\right) + o\left(N^{-\frac{1}{2}}\right), \qquad \text{a.s.} \tag{5.5}$$

and

$$\begin{aligned}\Delta\overline{\overline{\beta}}_N = {} & \frac{1}{N}\sum_{k=1}^{N} R_{J_k,J_N}^\top \mathbb{M}_{\beta,J_k}^{-1}\frac{1}{B}\sum_{i=1}^{B}\varepsilon_{i,k}\left(\Psi_{J_k}\left(Z_{0,i,k}\right) - P_{J_k}\nabla_z F_0(Z_{0,i,k})X_{i,k}\right) \\ & + \frac{1}{N}\sum_{k=1}^{N}\left(R_{J_k,J_N}^\top\beta_{J_k,0} - \beta_{J_N,0}\right) + o\left(N^{-\frac{1}{2}}\right), \qquad \text{a.s.}\end{aligned} \tag{5.6}$$

Lemma 5.3 is a key result of the paper. Two features are worth emphasizing. First, the influence functions in the leading terms of both $\Delta\overline{\overline{\theta}}_N$ and $\Delta\overline{\overline{\beta}}_N$ are *automatically orthogonalized*. That is, orthogonality emerges from the joint optimization over (θ, β), without requiring any explicit orthogonalization step. This is important both conceptually and computationally. In particular, constructing an orthogonal score for θ would require estimating $\mu_0(\theta_0, z)$, a vector of p unknown functions, which becomes increasingly costly as p grows. The joint estimation procedure therefore avoids a potentially high-dimensional nuisance estimation step.

The second feature is that the sieve coefficient $\Delta\overline{\overline{\beta}}_N$ displays a clear online learning structure.

For example, $\Delta\overline{\overline{\beta}}_N$ contains the bias term $\frac{1}{N}\sum_{k=1}^{N}\left(R_{J_k,J_N}^{\top}\beta_{J_k,0}-\beta_{J_N,0}\right)$, which reflects the evolution of the re-embedded sieve target as the dimension J_k increases. Moreover, the presence of the transformation $R_{J_k,J_N}^{\top}$ in the leading term of $\Delta\overline{\overline{\beta}}_N$ has a useful implication for function estimation. For the pointwise estimator $\overline{\overline{F}}_N(z)=\Psi_{J_N}(z)^{\top}\overline{\overline{\beta}}_N$, the leading stochastic term can be rewritten as

$$\frac{1}{N}\sum_{k=1}^{N}\Psi_{J_k}(z)^{\top}\mathbb{M}_{\beta,J_k}^{-1}\frac{1}{B}\sum_{i=1}^{B}\varepsilon_{i,k}\left(\Psi_{J_k}(Z_{0,i,k})-P_{J_k}\nabla_z F_0(Z_{0,i,k})X_{i,k}\right),$$

where we use the fact that $\Psi_{J_N}(z)^{\top}R_{J_k,J_N}^{\top}=\Psi_{J_k}(z)^{\top}$. This shows that each update contributes to the variance through the *local sieve representation* at level J_k, rather than the final sieve J_N. This highlights that the variance accumulation is naturally aligned with the evolving sieve complexity.

Remark 5.1. *Note that when $\sigma^2(x_0,X)=\sigma_0^2$, where σ_0^2 is a positive constant, the online estimator $\overline{\overline{\theta}}_N$ attains semiparametric efficiency (Ai and Chen 2003). However, when there is heteroscedasticity, semiparametric efficiency no longer holds. In this case, we can reweight the update to attain efficiency.*

Lemma 5.3 yields a root-N law for $\overline{\overline{\theta}}_N$ and the explicit moving-sieve upper bound for $\overline{\overline{F}}_N$ stated next. Define

$$\mathbb{V}_\theta:=\frac{1}{B}\mathbb{E}\left[\varepsilon^2(\nabla_z F_0(Z_0))^2(X-\mu_0(\theta_0,Z_0))(X-\mu_0(\theta_0,Z_0))^{\top}\right],\quad \Omega_\theta:=\mathbb{M}_\theta^{-1}\mathbb{V}_\theta\mathbb{M}_\theta^{-1}.$$

Theorem 5.1. *Let all conditions in Lemma 5.3 hold and Ω_θ be positive definite. Then we have:*

1. *Almost surely, for every fixed $a\in\mathbb{R}^p\setminus\{0\}$,*

$$\limsup_{N\to\infty}\sqrt{\frac{N}{2\log\log N}}\frac{a^{\top}\Delta\overline{\overline{\theta}}_N}{\sqrt{a^{\top}\Omega_\theta a}}=1,\qquad \liminf_{N\to\infty}\sqrt{\frac{N}{2\log\log N}}\frac{a^{\top}\Delta\overline{\overline{\theta}}_N}{\sqrt{a^{\top}\Omega_\theta a}}=-1.$$

2. *Almost surely,*

$$\|\overline{\overline{F}}_N-F_0\|_\infty=O\left(J_N^{1/2}N^{-1/2}(\log N)^{1/2}+\frac{1}{N}\sum_{k=1}^{N}J_k^{-s}\right),\tag{5.7}$$

where J_k^{-s}, $s>0$, denotes the sup-norm approximation error rate of using J_k dimensional sieve for F_0 (see Appendix D). When $J_k\asymp k^{\alpha_\dagger}$, the approximation error term *is* $O(N^{-s\alpha_\dagger})$ *if* $s\alpha_\dagger < 1$, $O(N^{-1}\log N)$ *if* $s\alpha_\dagger = 1$, *and* $O(N^{-1})$ *if* $s\alpha_\dagger > 1$.

5.2.2. FCLTs for θ and Policy Functionals

Lemma 5.3 also gives the following FCLT of $\overline{\overline{\theta}}_N$.

Theorem 5.2. *Let all conditions in Lemma 5.3 hold. Then:*

$$(N\Omega_\theta)^{-\frac{1}{2}}\,[Nr]\Delta\overline{\overline{\theta}}_{[Nr]} \Longrightarrow \mathbb{W}_p(r),\;\; r\in[0,1], \tag{5.8}$$

where $\mathbb{W}_p(r)$ *is a* p*-dimensional standard Wiener process. Consequently,*

$$\Omega_\theta^{-\frac{1}{2}}N^{\frac{1}{2}}\Delta\overline{\overline{\theta}}_N \to_d \mathcal{N}(0,\mathbb{I}_p). \tag{5.9}$$

Based on the results of Lemma 5.3, we can also develop an FCLT result for the average marginal effect $\overline{\overline{\tau}}_N$ proposed in Section 3.5. To formally state the theorem, define

$$\begin{aligned}\Upsilon_{i,k} &= \varepsilon_{i,k}\mathbb{E}[\nabla_{zz}F_0(Z_0)\theta_0X^\top]\mathbb{M}_\theta^{-1}\nabla_zF_0(Z_{0,i,k})\left(X_{i,k}-\mu_0(\theta_0,Z_{0,i,k})\right)\\ &\quad+\varepsilon_{i,k}\mathbb{E}[\nabla_z\Psi_{J_k}(Z_0)^\top]\mathbb{M}_{\beta,J_k}^{-1}\left(\Psi_{J_k}(Z_{0,i,k})-P_{J_k}\nabla_zF_0(Z_{0,i,k})X_{i,k}\right)\theta_0\\ &\quad+\varepsilon_{i,k}\mathbb{E}[\nabla_zF_0(Z_0)]\mathbb{M}_\theta^{-1}\nabla_zF_0(Z_{0,i,k})\left(X_{i,k}-\mu_0(\theta_0,Z_{0,i,k})\right)\\ &\quad+\left(\nabla_zF_0(Z_{0,i,k})-\mathbb{E}[\nabla_zF_0(Z_0)]\right)\theta_0.\end{aligned}$$

In Appendix E, we show that $\frac{1}{N}\sum_{k=1}^N\frac{1}{B}\sum_{i=1}^B\Upsilon_{i,k}$ is the influence function of $\Delta\overline{\overline{\tau}}_N$. There we also impose a mild condition that $\mathbb{E}[\Upsilon_k\Upsilon_k^\top]\to V_\Upsilon$ for some positive definite matrix V_Υ. We have the following theorem.

Theorem 5.3. *Let all conditions in Lemma 5.3 and Condition 15 in Appendix E hold. Then*

$$(NV_\Upsilon)^{-\frac{1}{2}}[Nr]\Delta\overline{\overline{\tau}}_{[Nr]} \Longrightarrow \mathbb{W}_p(r),\quad r\in[0,1]. \tag{5.10}$$

Remark 5.2. *By using FCLT-based random-scaling inference, we avoid estimating complicated asymptotic variance matrix* V_τ*, which is also empirically attractive. We also point that, to save space and illustrate the main idea, we only report the FCLT for AME of continuous regressors; for formal results of FCLT of AME for binary regressors as we used in applications, see Section E.3.*

6. Monte Carlo Experiments

This section presents Monte Carlo studies to evaluate the performances of our online learning algorithm. The simulation data is generated according to the following binary-choice DGP:

$$Y = \mathbf{1}\big(x_0 + X^\top \theta_0 - 0.5\|\theta_0\|_2 \cdot u \geq 0\big)\,, \tag{6.1}$$

where Y is the observed response, $(x_0, X^\top)^\top$ is the regressor vector. Unless otherwise specified, we consider the following setups. $x_0 \sim \mathcal{N}(0,1)$, $x_j \sim \mathcal{N}(0,1)$ or $x_j \sim \frac{1}{\sqrt{5}}(\frac{1}{2}\mathcal{N}(2,1) + \frac{1}{2}\mathcal{N}(-2,1))$ for $1 \leq j \leq 0.4p$, and $x_j \sim$ Rademacher for $0.4p+1 \leq j \leq p$. The latent shock u is independent of (x_0, X). We consider two shock distributions: (i) $u \sim$ Cauchy and (ii) a skewed-normal–type distribution generated as $u = (v_1 + |v_2|)/\sqrt{2},\ v_1, v_2 \sim$ i.i.d. $\mathcal{N}(0,1)$. We specify θ_0 as $\theta_0 = (\theta_{10}^\top, -\theta_{20}^\top, \theta_{30}^\top, -\theta_{40}^\top)^\top$, where $\theta_{10}, \theta_{20} \in \mathbb{R}^{0.2p}$, $\theta_{30}, \theta_{40} \in \mathbb{R}^{0.3p}$, and their arguments decrease linearly from 1 to 0. We consider two dimensions $p \in \{50, 100\}$, two batch sizes $B \in \{25, 50\}$, and fixing the total number of online updates at $N = 10^7$. Unless otherwise specified, for all Monte Carlo results reported below, the details of the implementations of the online algorithm are identical to that in Section 4.

In Section 6.1 and Section 6.2 below, the reported simulation results on our online estimators are averaged over 200 independent Monte Carlo replications. Since it is very time-consuming to estimate a high-dimensional monotone index model using any offline procedure with a large sample size, the simulation results in Section 6.3 below are averaged over 20 independent Monte Carlo replications only.

6.1. Online Estimation and Inference of θ_0

This section focuses on Phase II's estimation and inference of parameter θ_0. We report the bias, root mean squared error (RMSE), confidence interval coverage rate (CR), and average confidence interval length in Table 1 and Table 2. The results demonstrate that the procedure achieves negligible bias, small RMSE, and reliable inference uniformly across all configurations, with coverage rates tightly centered around the nominal 0.95 level. One clear pattern is the effect of the mini-batch size B. Increasing B from 25 to 50 leads to uniform improvements in precision— the RMSE declines, and confidence interval becomes noticeably shorter. This is consistent with the effective sample size interpretation: larger batches reduce the variance of the stochastic updates and stabilize the online trajectory.

Table 1. Simulation Results (Normal + Rademacher Design)

Statistic	Target	Cauchy Errors				Skewed Normal Errors			
		$p = 50$		$p = 100$		$p = 50$		$p = 100$	
		$B = 25$	$B = 50$	$B = 25$	$B = 50$	$B = 25$	$B = 50$	$B = 25$	$B = 50$
Bias	$\theta_0(1)$	0.0001	0.0000	0.0003	0.0000	0.0000	0.0000	0.0001	0.0000
	Avg	0.0000	0.0000	0.0001	0.0000	0.0000	0.0000	0.0000	0.0000
RMSE	$\theta_0(1)$	0.0007	0.0004	0.0014	0.0006	0.0004	0.0003	0.0006	0.0003
	Avg	0.0005	0.0003	0.0010	0.0005	0.0003	0.0002	0.0005	0.0003
CR	$\theta_0(1)$	0.9400	0.9450	0.9250	0.9200	0.9550	0.9200	0.9550	0.9550
	Avg	0.9461	0.9424	0.9482	0.9403	0.9454	0.9407	0.9499	0.9412
CI Length	$\theta_0(1)$	0.0034	0.0020	0.0060	0.0030	0.0019	0.0013	0.0030	0.0018
	Avg	0.0028	0.0017	0.0050	0.0025	0.0016	0.0010	0.0025	0.0015

Note: Total number of updates is $N = 10^7$, and the effective sample size is $n = B \times N$. This also applies to Table 2 and Table 3. Results are Monte Carlo averages over 200 independent simulation replications. All online estimators/algorithms start from zero. The nominal coverage rate is 0.95. CR is constructed via random scaling. $\theta_0(1)$ refers to the first component of θ_0, whose value is 1. Avg reports the average of each summary statistic across all parameters.

The distribution of the shock u affects the level of difficulty but not the qualitative conclusions. Heavy-tailed Cauchy errors lead to larger RMSEs and longer confidence intervals relative to skewed-normal designs, reflecting the presence of extreme realizations. Nevertheless, the estimator remains stable, and coverage stays close to nominal even in these challenging settings.

Increasing the dimension from $p = 50$ to $p = 100$ results in only modest increases in RMSE and interval length, indicating that the procedure scales well with dimensionality. Importantly, inference remains well-calibrated across dimensions, with no systematic deterioration in coverage.

Finally, the results are robust to the distribution of regressors. Whether regressors are Gaussian, mixture-normal, or discrete, the same qualitative patterns emerge: accurate estimation, stable inference, and systematic gains from larger batch sizes.

6.2. Estimation and Inference of Marginal Effects

We now examine the finite-sample performance of the online estimators for average marginal effects defined in (3.19). To save space, we only report the simulation results for Normal

TABLE 2. SIMULATION RESULTS (MIXTURE NORMAL + RADEMACHER FEATURES)

Statistic	Target	Cauchy Errors				Skewed Normal Errors			
		$p = 50$		$p = 100$		$p = 50$		$p = 100$	
		$B = 25$	$B = 50$	$B = 25$	$B = 50$	$B = 25$	$B = 50$	$B = 25$	$B = 50$
Bias	$\theta_0(1)$	0.0000	0.0000	0.0003	0.0000	0.0000	0.0000	0.0000	0.0001
	Avg	0.0000	0.0000	0.0002	0.0000	0.0000	0.0000	0.0000	0.0000
RMSE	$\theta_0(1)$	0.0007	0.0004	0.0013	0.0006	0.0004	0.0003	0.0006	0.0004
	Avg	0.0006	0.0003	0.0010	0.0005	0.0003	0.0002	0.0005	0.0003
CR	$\theta_0(1)$	0.9550	0.9350	0.9500	0.9350	0.9550	0.9350	0.9600	0.9300
	Avg	0.9483	0.9388	0.9515	0.9458	0.9481	0.9379	0.9529	0.9406
CI Length	$\theta_0(1)$	0.0036	0.0021	0.0063	0.0030	0.0020	0.0013	0.0030	0.0019
	Avg	0.0028	0.0017	0.0051	0.0025	0.0016	0.0010	0.0025	0.0015

and Rademacher regressors in Table 3. We can see that the bias and RMSE of $\overline{\overline{\tau}}_N$ are uniformly small and close to 0. Moreover, the coverage rate of the random-scaling confidence interval is close to the nominal 0.95 level uniformly across dimensions, batch sizes, and shock distributions, confirming the validity of online inference for the average marginal effect.

In Appendix A, we also report additional results on estimation and inference for point-wise marginal effect.

6.3. Computation Efficiency: Insensitivity to Choice of Warm Start

Although our algorithm is designed for online estimation, it can also be applied to offline estimation with a fixed sample. In particular, we can draw random subsample from the fixed data set with replacement, and process these subsamples using our algorithm as if they were online. This section provides simulation results that demonstrate significant computational gains of our online Phase II estimation over the offline estimation using the same sieve NLS criterion function and the same warm start.

As an illustration, consider the DGP (6.1) with $u \sim$ Cauchy and $p = 100$. We consider two sample sizes: $n = 5 \times 10^5$ and $n = 10^6$, and correspondingly, two sieve dimensions $J_n = 30$ and $J_n = 40$. Two estimation methods are compared: full-sample estimation using MATLAB optimization package `lsqnonlin` and local online NLS with multiple passes. Precisely, the full-sample estimator solves the joint sieve NLS criterion via `lsqnonlin` using the trust-

Table 3. Average Marginal Effects: Normal and Rademacher Regressors

		Cauchy Errors				Skewed Normal Errors			
		$p = 50$		$p = 100$		$p = 50$		$p = 100$	
		$B = 25$	$B = 50$	$B = 25$	$B = 50$	$B = 25$	$B = 50$	$B = 25$	$B = 50$
Bias	$\tau_0(1)$	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
	Avg	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
RMSE	$\tau_0(1)$	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
	Avg	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
CR	$\tau_0(1)$	0.9550	0.9300	0.9500	0.9350	0.9550	0.9300	0.9500	0.9350
	Avg	0.9486	0.9393	0.9514	0.9428	0.9486	0.9393	0.9514	0.9428
CI Length	$\tau_0(1)$	0.0002	0.0001	0.0002	0.0001	0.0002	0.0001	0.0002	0.0001
	Avg	0.0002	0.0001	0.0002	0.0001	0.0002	0.0001	0.0002	0.0001

Note: Results are Monte Carlo averages over 200 simulation replications. $\tau_0(1)$ denotes the first component of the average marginal effect. Avg reports averages across all components.

region-reflective algorithm with an analytical sparse Jacobian for the joint (θ, β) block; the solver is capped at 100 iterations and 300 function evaluations, and the function, step, and first-order optimality tolerances are all set to 10^{-6}. For the local online sieve NLS, we randomly reshuffle the full sample, and sequentially draw data points from the reshuffled data with fixed batch size B and feed them to the online algorithm. A single pass is defined as one complete traversal of the dataset, consisting of n/B sequential updates with $B = 25$; *multiple passes* feature multiple data reshuffling and complete traversal. In the simulation, we consider 10 passes, with PR averaging over the last 5 passes only.

Table 4. Full-sample versus online estimation, Online Warm Start

	$n = 5 \times 10^5$, $J_n = 30$			$n = 10^6$, $J_n = 40$		
Estimator	Bias	RMSE	Time (min)	Bias	RMSE	Time (min)
Online Phase I (warm start)	0.1864	0.2180	0.1256	0.1868	0.2180	0.1349
Full-sample NLS (`lsqnonlin`)	0.0064	0.0193	9.5290	0.0043	0.0140	24.623
Online Phase II (multi-pass)	0.0022	0.0144	0.1217	0.0017	0.0109	0.2676
Speedup ratio (Full-sample time / Online time)			78.3			92.0

Notes. Results are averaged over 20 Monte Carlo replications with $p = 100$ for θ_0, J_n is sieve order for F_0. The online Phase I updates with $N_0 = 1 \times 10^5$ and $N_1 = 2 \times 10^5$ (see Section 4.1.1 for definitions of N_0 and N_1). Both methods are warm-started from the same Phase I output, so reported times exclude the shared Phase I cost. Full-sample estimator solves the joint sieve NLS criterion via `lsqnonlin`. The online estimator uses local joint sieve NLS refinement, multi-pass uses 10 passes through the data, with PR averaging over the last 5 passes only.

Table 5. Full-sample versus online estimation, SMCO warm start

	$n = 5 \times 10^5$, $J_n = 30$			$n = 10^6$, $J_n = 40$		
Estimator	Bias	RMSE	Time (min)	Bias	RMSE	Time (min)
SMCO Phase I (warm start)	0.1907	0.2324	3.7990	0.2021	0.2470	3.4410
Full-sample NLS (`lsqnonlin`)	0.0033	0.0192	9.6050	0.0031	0.0138	23.897
Online Phase II (multi-pass)	0.0023	0.0146	0.1440	0.0021	0.0104	0.2620
Speedup ratio (Full-sample time / Online time)						
			66.5			91.3

Notes. Results are averaged over 20 Monte Carlo replications with $p = 100$ for θ_0, J_n is sieve order for F_0. Both methods are warm-started from the same SMCO Phase I output, so reported times exclude the shared Phase I cost. SMCO Phase I uses a subsample of size 10^4 to maximize the negative sieve NLS criteria jointly over (θ, β) on the search box $[-2, 2]^p \times [-3, 3]^{J_n+1}$. SMCO is run with 10 Sobol-initialized starts; each start uses `iter_max` = 100 refinement iterations preceded by `iter_boost` = 200 boost iterations, forward partial differences, a fixed 5% bounds buffer, the run-max accumulator turned on, and convergence tolerance 10^{-8}. The best objective value across the 10 starts is taken as the warm start.

Two warm-start algorithms are considered: the online learner proposed in Section 3.1 and 3.2, and a generic warm start SMCO (Chen et al. 2026). Simulation results are reported in Table 4 and Table 5. Regardless of the choice of warm start, an attractive pattern emerges: initiating from the same warm start, full-sample estimation takes around 10 minutes to solve the optimization problem when sample size is 5×10^5 and 24 minutes when sample size is 1×10^6, whereas our online Phase II requires only about 8 and 16 seconds. The computational advantage is more compelling for larger sample sizes. More importantly, the bias and RMSE of package-based full-sample optimization are slightly larger than those of online estimators. This suggests that the full-sample objective function may be *under-optimized* given the tuning parameters chosen for `lsqnonlin` package. To close this gap, additional time will be necessary, making full-sample estimation even less favorable.

7. Conclusion and Extension

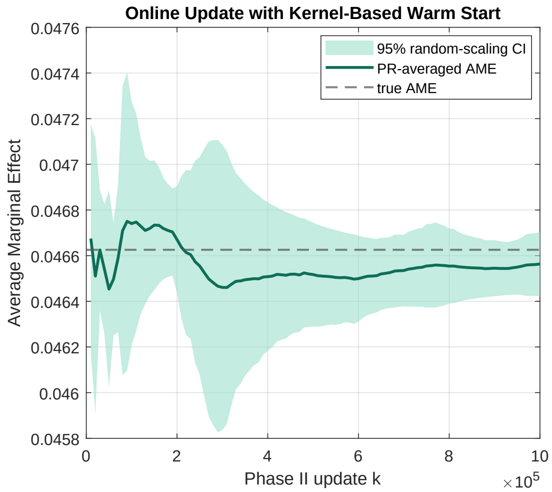
This paper develops a two-phase online estimation and inference algorithm for semiparametric monotone index models with many covariates (some of which are discrete). The online warm-start phase uses a globally stable update rule that consistently learns the finite-dimensional index parameter from arbitrary initialization, and also learn a sup-norm consistent link function via online sieve least squares using the warm start index as a consistent generated regressor. It then performs locally joint sieve nonlinear least squares updates in the second online phase. The PR averages of the Phase II updates attain the optimal convergence rates, are automatically orthogonalized and satisfy the FCLT. A key byproduct of the procedure is a sequence of parameter updates—learning trajectories—that can be used for online inference via random scaling with essentially no additional nonparametric estimation burden. The same trajectories also support online estimation and inference for policy-relevant functionals that depend on both the parametric and nonparametric components. Monte Carlo experiments show good finite-sample performance with coverage rates close to nominal. Two empirical illustrations demonstrate feasibility in a high-dimensional setting while leaving the link function unspecified. The procedure is ideal for streaming environments where continuously re-estimating on the full sample (or even storing it) is infeasible, but is also a computationally fast alternative to full-sample offline sieve M estimation and inference methods. Our method delivers a practical semiparametric toolkit for real-time estimation.

7.1. Extensions with Generic Warm Starts

The framework naturally invites many extensions, and we discuss a few here. Section 7.1.1 shows that our online estimation and inference framework can be extended to the case where the unknown function F_0 is approximated using deep neural network (DNN). Section 7.1.2 and 7.1.3 show that our framework can be extended to various semiparametric models. In particular, Section 7.1.2 sketches an online sample-selection model in which both the selection and outcome equations are learned semiparametrically. Section 7.1.3 outlines an extension to multi-alternative choice models, connecting the framework to interpretable multi-index neural architectures. In the extensions, the Phase I update can be replaced with any online updates that deliver consistent warm starts, such as SMCO (Chen et al. 2026).

7.1.1. Approximating Unknown Link Using Monotonic DNN

This section considers the extension where the monotonic link function F_0 is approximated using DNN instead of linear sieves. We follow Sartor, Sinigaglia, and Susto (2025)'s construction so that the DNN-based estimator of F_0 is monotonic, which preserves the shape. Then we replace the spline-based approximation of F_0 with the DNN-based estimator in the NLS criterion, and use Phase II update to jointly estimate the DNN weights and unknown parameter θ_0, based on which the average marginal effects can be estimated and online inference using random scaling can be conducted.



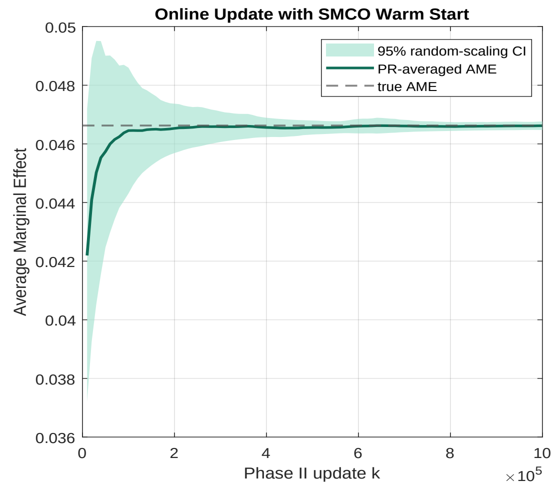


Figure 8. The PR Average Trajectory of Average Marginal Effect from One Monte Carlo Replicate, $p = 500$, DNN sieve

Table 6. Online Phase II AME estimator under different warm starts, $p = 500$, Monotone DNN sieve.

Warm start	Bias	RMSE	Coverage	Avg. CI length
Online Kernel	0.000006	0.000077	0.94	0.000371
SMCO	0.000021	0.000109	0.97	0.000701

Note: Note: Reported online Phase II results are averaged across 200 Monte Carlo replications, on the bias, root mean squared error (RMSE), empirical coverage rate, the average length of the 95% confidence interval for the average marginal effect (AME). Confidence intervals are constructed using random-scaling inference. "Online Kernel" and "SMCO" denote the two Phase I warm-start procedures used to initialize the online Phase II updates. The table shows that online Phase II random-scaling inference attains coverage close to the nominal level regardless of which warm-start procedure is used. The table is not to compare the two warm starts against each other since their effective warm-start sample sizes differ and thus their bias and RMSE are not directly comparable.

The DGP follows (6.1) in Section 6, except that here we consider a higher dimension $p = 500$. $x_0, x_1, \cdots, x_{40}$ are standard normal, and $x_{41}, \cdots, x_{100}$ are Rademacher. Their corresponding coefficients are constructed similarly as in $p = 100$ case of Section 6. The remaining 400 regressors are all standard normally distributed, and each of their coefficients is equal to 0.05 so that the total contribution of the additional 400 regressors to index variance is equal to 1. A deep neural network (DNN) with depth of 4 and width of 16 for each layer is used to approximate unknown monotone link function F_0.[9] The estimation and inference target is the average marginal effect. For illustration, we focus on the average marginal effect of the first regressor, which is given by $\mathbb{E}(\nabla_z F(Z_0))$ due to the fact that $\theta_{0,1} = 1$.

[9]Under Sartor et al. (2025)'s construction, 4 is the minimum depth of a DNN to guarantee its shape-preserving.

We consider two methods to construct the warm starts of θ_0 and DNN weights. The first method resembles our spline-based two-phase online learning algorithm: we use the kernel-based online algorithm to search for a warm start of θ_0, then use the estimated θ_0 to construct the index estimator, based on which the weights can be updated based on the NLS criterion. Finally, we jointly refine the weights and θ based on NLS criterion. For this method, we choose $N_0 = 1 \times 10^5$ and $N_1 = 2 \times 10^5$. The second method uses SMCO (Chen et al. 2026) to search for joint warm starts of θ_0 and DNN weights, where the warm-start sample size is chosen as 10^4. For both methods, we choose batch size $B = 25$, and consider 10^6 joint updates. Finally, 200 independent Monte Carlo replicates are performed.

Table 6 reports the simulation results across the 200 Monte Carlo replicates, and Figure 8 plots the PR average path of the AME and its random-scaling-based confidence band for one replicate. We can see that the confidence band constructed from the random scaling approach has good coverage rate close to the normal one, regardless of the choices of the warm start.

7.1.2. Sample Selection

Let $D \in \{0, 1\}$ indicate whether the outcome is observed, and suppose that the researcher observes $(D, DY^*, x_{10}, X_1, x_{20}, X_2)$. A semiparametric online sample-selection model is

$$D = 1\{S^* \geq 0\}, \quad \Pr(D = 1|x_{10}, X_1) = F_{10}(x_{10} + X_1^\top \theta_{10}), \quad (7.1)$$

$$Y^* = F_{20}(x_{20} + X_2^\top \theta_{20}) + \varepsilon, \quad (7.2)$$

$$\mathbb{E}[\varepsilon \mid x_{10}, X_1, x_{20}, X_2, D = 1] = \lambda_0\{F_{10}(x_{10} + X_1^\top \theta_{10})\}. \quad (7.3)$$

Here F_{10} is an unknown monotone selection link, F_{20} is the unknown outcome link, and λ_0 is a control function capturing selection on unobservables. The selected conditional mean is therefore

$$\mathbb{E}[Y^* \mid x_{10}, X_1, x_{20}, X_2, D = 1] = F_{20}(x_{20} + X_2^\top \theta_{20}) + \lambda_0\{F_{10}(x_{10} + X_1^\top \theta_{10})\}.$$

This nests the no-selection-bias case by setting $\lambda_0 \equiv 0$. It also separates the economics of outcome production, through $F_{20}(x_{20} + X_2^\top \theta_{20})$, from the economics of observability or participation, through $F_{10}(x_{10} + X_1^\top \theta_{10})$. For identification we need the usual economic exclusion or support condition: some component of (x_{10}, X_1) should move selection without entering the outcome index except through the control function.

The sample-selection model has a direct choice interpretation. The selection indicator D may represent labor-force participation, market entry, survey response, purchase, attention, or consideration. The outcome equation is observed only after that economic choice. The control function $\lambda_0\{F_{10}(x_{10} + X_1^\top \theta_{10})\}$ captures the unobserved factors that link the first-stage selection choice with the selected outcome. Thus selection correction is not merely a missing-data device. It is an online two-stage model of economic choice and realized outcomes.

Online Algorithm At batch k, with observations $W_{i,k} = (D_{i,k}, D_{i,k}Y^*_{i,k}, x_{10,i,k}, X_{1,i,k}, x_{20,i,k}, X_{2,i,k})$, we propose to use the following three online learners.

Step 1: selection learner. Estimate (θ_{10}, F_{10}) from the binary outcome D. One can use the same two-phase monotone-index learners as in the baseline paper.

Step 2: selected outcome learner. For selected observations, define the generated control of $W_{i,k}$ as $\widehat{q}_{i,k-1} = \widehat{F}_{10,k-1}(x_{10,i,k} + X_{1,i,k}^\top \widehat{\theta}_{1,k-1})$, where $\widehat{F}_{10,k-1}$ and $\widehat{\theta}_{1,k-1}$ are online estimators of F_{10} and θ_{10} in period $k-1$ from the first step algorithm. Then estimate $(\theta_{20}, F_{20}, \lambda_0)$ from the selected outcome loss

$$\mathcal{L}^y_{J,L}(\theta, \beta, \delta; W_{i,k}) = D_{i,k}\Big[Y^*_{i,k} - \Psi_J(x_{20,i,k} + X_{2,i,k}^\top \theta_2)^\top \beta - \Psi_L(\widehat{q}_{i,k-1})^\top \delta\Big]^2.$$

The online stochastic-gradient update is

$$\omega^y_k = \mathcal{R}_k^\top \omega^y_{k-1} - \frac{\xi_k}{B}\sum_{i=1}^{B} \nabla_{\omega^y} \mathcal{L}^y_{J_k, L_k}(\mathcal{R}_k^\top \omega^y_{k-1}; W_{i,k}),$$

where $\omega^y = (\theta^\top, \beta^\top, \delta^\top)^\top$, and $\mathcal{R}_k$ embeds the old coefficient vector into the enlarged sieve spaces for F_0 and λ_0, which is similarly defined as that in the baseline algorithm.

Step 3: random-scaling inference. The trajectory of PR averages can be used for random-scaling inference for θ_{10}, θ_{20}, average marginal effects, and selection-corrected means.

7.1.3. Extension to Multi-Alternative Choice

Let an agent choose $l \in \{0, 1, \ldots, L\}$ from utilities

$$U_l = V_l(x_{0l}, X_l; \theta_j) + u_l, \qquad V_l(x_{0l} + X_l; \theta_l) = x_{0l} + X_l^\top \theta_{l0}.$$

Let X collects all features vectors from the L utilities, the observable object is the vector of choice probabilities

$$P_l(X) = \Pr\{U_l \geq U_{l'} \text{ for all } 0 \leq l' \leq L \mid X\}, \qquad \sum_{l=0}^{L} P_l(X) = 1.$$

A direct extension of the monotone index model is the vector monotone-index model

$$P_l(X) = G_{l,0}(V_0, \ldots, V_L), \qquad l = 0, \ldots, L,$$

where for each l, $G_{l,0} : \mathbb{R}^{L+1} \to \Delta^L$ is an unknown vector-valued link. The loss is quasi-likelihood:

$$\mathcal{L}(\theta, \beta; W) = -\sum_{l=0}^{L} 1\{Y = l\} \log \left[\Psi_J(x_{00} + X_0^\top \theta_0, \cdots, x_{L0} + X_L^\top \theta_0)^\top \beta_l \right].$$

To preserve the economic restrictions of random utility, one can represent G_{0l} as the gradient of a convex surplus or choice-potential function:

$$G_{l,0}(v) = \frac{\partial \mathbf{U}_L(v)}{\partial v_j}, \qquad \mathbf{U}_L(v) = \sum_{r=1}^{R} a_r \phi(c_r + b_r' v), \qquad a_r \geq 0.$$

Convexity of $\mathbf{U}_L$ helps encode the shape restrictions associated with stochastic choice, while the gradient automatically gives nonnegative own-utility responses after normalization. The online update is again a single-hidden-layer neural network, now with the network output constrained to be a choice-probability vector.

A further extension that allows for attention or consideration can be written as

$$A_l = 1\{V_l(x_{0l} + X_l^\top \theta_{l0}) + u_l \geq 0\}, \qquad Y = \arg\max_{l: A_l = 1} \{x_{0l} + X_l^\top \theta_{0l} + u_l\}.$$

The first layer learns which alternatives enter the consideration set; the second layer learns choice probabilities conditional on consideration. This connects the sample selection model above with McFadden-style economic choice: observability, attention, and final choice are all economic decisions that can be learned online with interpretable semiparametric neural-index components.

References

J. Abrevaya, J. Hausman, and S. Khan. Testing for causal effects in a generalized regression model with endogenous regressors. *Econometrica*, 78:2043–2061, 2010.

Sumit Agarwal, Souphala Chomsisengphet, Neale Mahoney, and Johannes Stroebel. Do banks pass through credit expansions to consumers who want to borrow? *The Quarterly Journal of Economics*, 133(1):129–190, 2018.

Hyungtaik Ahn, Hidehiko Ichimura, James L Powell, and Paul A Ruud. Simple estimators for invertible index models. *Journal of Business & Economic Statistics*, 36(1):1–10, 2018.

Chunrong Ai and Xiaohong Chen. Efficient estimation of models with conditional moment restrictions containing unknown functions. *Econometrica*, 71(6):1795–1843, 2003.

David Atkin, Benjamin Faber, and Marco Gonzalez-Navarro. Retail globalization and household welfare: Evidence from mexico. *Journal of Political Economy*, 126(1):1–73, 2018.

Alexandre Belloni, Victor Chernozhukov, Denis Chetverikov, and Kengo Kato. Some new asymptotic theory for least squares series: Pointwise and uniform results. *Journal of Econometrics*, 186(2):345–366, 2015.

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

Wolfgang Boehm. Inserting new knots into b-spline curves. *Computer-Aided Design*, 12(4): 199–201, 1980.

Léon Bottou, Frank E Curtis, and Jorge Nocedal. Optimization methods for large-scale machine learning. *Siam Review*, 60(2):223–311, 2018.

Christian Broda and David E Weinstein. Product creation and destruction: Evidence and price implications. *American Economic Review*, 100(3):691–723, 2010.

Alberto Cavallo. Are online and offline prices similar? evidence from large multi-channel retailers. *American Economic Review*, 107(1):283–303, 2017.

X. Chen, J. Lee, X. Tong, and Y. Zhang. Statistical inference for model parameters in stochastic gradient descent. *The Annals of Statistics*, 48:251–273, 2020.

Xiaohong Chen. Large sample sieve estimation of semi-nonparametric models. *Handbook of Econometrics*, 6:5549–5632, 2007.

Xiaohong Chen and Timothy M Christensen. Optimal uniform convergence rates and asymptotic normality for series estimators under weak dependence and weak conditions. *Journal of Econometrics*, 188(2):447–465, 2015. doi: 10.1016/j.jeconom.2015.03.010.

Xiaohong Chen and Halbert White. Asymptotic properties of some projection-based robbins-monro procedures in a hilbert space. *Studies in Nonlinear Dynamics & Econometrics*, 6 (1), 2002.

Xiaohong Chen, Sokbae Lee, Yuan Liao, Myung Hwan Seo, Youngki Shin, and Myunghyun Song. Sgmm: Stochastic approximation to generalized method of moments. *Journal of Financial Econometrics*, page nbad027, 2023.

Xiaohong Chen, Min Seong Kim, Sokbae Lee, Myung Hwan Seo, and Myunghyun Song. Slim: Stochastic learning and inference in overidentified models. *arXiv preprint arXiv:2510.20996*, 2025.

Xiaohong Chen, Zengjing Chen, Wayne Yuan Gao, Xiaodong Yan, and Guodong Zhang. Optimization via the strategic law of large numbers. *Proceedings of the National Academy of Sciences*, 123(4):e2519845123, 2026.

Stephen R Cosslett. Efficiency bounds for distribution-free estimators of the binary choice and the censored regression models. *Econometrica: Journal of the Econometric Society*, pages 559–585, 1987.

D. R. Cox. Regression models and life-tables. *Journal of the Royal Statistical Society: Series B (Methodological)*, 34(2):187–220, 1972.

M. Das, W.K Newey, and F. Vella. Nonparametric estimation of sample selection models. *Review of Economic Studies*, 70(1):33–58, 2003.

Stefano DellaVigna and Matthew Gentzkow. Uniform pricing in us retail chains. *The Quarterly Journal of Economics*, 134(4):2011–2084, 2019.

Yanqin Fan, Fang Han, Wei Li, and Xiao-Hua Zhou. On rank estimators in increasing dimensions. *Journal of Econometrics*, 214(2):379–412, 2020.

David A Freedman. On tail probabilities for martingales. *the Annals of Probability*, pages 100–118, 1975.

Saeed Ghadimi and Guanghui Lan. Stochastic first-and zeroth-order methods for nonconvex stochastic programming. *SIAM journal on optimization*, 23(4):2341–2368, 2013.

Aaron K Han. Non-parametric analysis of a generalized regression model: the maximum rank correlation estimator. *Journal of Econometrics*, 35(2-3):303–316, 1987.

D. Han, J. Xie, J. Liu, L. Sun, J. Huang, B. Jiang, and L. Kong. Inference on high-dimensional single-index models with streaming data. *Journal of Machine Learning Research*, 25:1–68, 2024.

Wolfgang Härdle, Peter Hall, and Hidehiko Ichimura. Optimal smoothing in single-index models. *The annals of Statistics*, 21(1):157–178, 1993.

James Heckman. Shadow prices, market wages, and labor supply. *Econometrica*, 42(4): 679–694, 1974.

Elhanan Helpman, Marc Melitz, and Yona Rubinstein. Estimating trade flows: Trading partners and trading volumes. *The quarterly journal of economics*, 123(2):441–487, 2008.

Joel L Horowitz. A smoothed maximum score estimator for the binary response model. *Econometrica: journal of the Econometric Society*, pages 505–531, 1992.

Yinxiao Huang, Xiaohong Chen, and Wei Biao Wu. Recursive nonparametric estimation for time series. *IEEE Transactions on Information Theory*, 60(2):1301–1312, 2013.

Hidehiko Ichimura. Semiparametric least squares (sls) and weighted sls estimation of single-index models. *Journal of econometrics*, 58(1-2):71–120, 1993.

Gabriel Jiménez, Steven Ongena, José-Luis Peydró, and Jesús Saurina. Hazardous times for monetary policy: What do twenty-three million bank loans say about the effects of monetary policy on credit risk-taking? *Econometrica*, 82(2):463–505, 2014.

John D. Kalbfleisch and Ross L. Prentice. *The Statistical Analysis of Failure Time Data*. John Wiley & Sons, New York, 1980.

Shakeeb Khan, Elie Tamer, and Qingsong Yao. Inference on high dimensional selective labeling models. *arXiv preprint arXiv:2410.18381*, 2024.

Shakeeb Khan, Xiaoying Lan, Elie Tamer, and Qingsong Yao. Estimating high dimensional monotone index models by iterative convex optimization. *Journal of Econometrics*, 252: 105901, 2025.

Roger W Klein and Richard H Spady. An efficient semiparametric estimator for binary response models. *Econometrica: Journal of the Econometric Society*, pages 387–421, 1993.

H. J. Kushner and G. Yin. *Stochastic Approximation and Recursive Algorithms and Applications*. Springer, 2003.

Tony Lancaster. *The Econometric Analysis of Transition Data*. Cambridge University Press, Cambridge, 1990.

Sokbae Lee, Yuan Liao, Myung Hwan Seo, and Youngki Shin. Fast and robust online inference with stochastic gradient descent via random scaling. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 7381–7389, 2022.

Sokbae Lee, Yuan Liao, Myung Hwan Seo, and Youngki Shin. Fast inference for quantile regression with tens of millions of observations. *Journal of Econometrics*, 249:105673, 2025.

Charles F Manski. Maximum score estimation of the stochastic utility model of choice. *Journal of econometrics*, 3(3):205–228, 1975.

Daniel McFadden. Economic choices. *American economic review*, 91(3):351–378, 2001.

Emma Pierson, Camelia Simoiu, Jan Overgoor, Sam Corbett-Davies, Daniel Jenson, Amy Shoemaker, Vignesh Ramachandran, Phoebe Barghouty, Cheryl Phillips, Ravi Shroff, et al. A large-scale analysis of racial disparities in police stops across the united states. *Nature human behaviour*, 4(7):736–745, 2020.

Boris T Polyak and Anatoli B Juditsky. Acceleration of stochastic approximation by averaging. *SIAM journal on control and optimization*, 30(4):838–855, 1992.

James Powell. Estimation of semiparametric models. *Handbook of Econometrics*, 4:2444–2521, 1994.

James L Powell, James H Stock, and Thomas M Stoker. Semiparametric estimation of index coefficients. *Econometrica: Journal of the Econometric Society*, pages 1403–1430, 1989.

Herbert Robbins and Sutton Monro. A stochastic approximation method. *The annals of mathematical statistics*, pages 400–407, 1951.

Herbert Robbins and David Siegmund. A convergence theorem for non negative almost supermartingales and some applications. In *Optimizing methods in statistics*, pages 233–257. Elsevier, 1971.

David Ruppert. Efficient estimations from a slowly convergent robbins-monro process. Technical report, Cornell University Operations Research and Industrial Engineering, 1988.

Davide Sartor, Alberto Sinigaglia, and Gian Antonio Susto. Advancing constrained monotonic neural networks: Achieving universal approximation beyond bounded activations. *arXiv preprint arXiv:2505.02537*, 2025.

Robert P Sherman. The limiting distribution of the maximum rank correlation estimator. *Econometrica: Journal of the Econometric Society*, pages 123–137, 1993.

Camelia Simoiu, Sam Corbett-Davies, and Sharad Goel. The problem of infra-marginality in outcome tests for discrimination. *Annals of Applied Statistics*, 11:1193–1216, 2017.

Charles J Stone. Optimal global rates of convergence for nonparametric regression. *The annals of statistics*, pages 1040–1053, 1982.

RJ Tomkins. Lindeberg functions and the law of the iterated logarithm. *Zeitschrift für Wahrscheinlichkeitstheorie und Verwandte Gebiete*, 65(1):135–143, 1983.

Panos Toulis and Edoardo M Airoldi. Asymptotic and finite-sample properties of estimators based on stochastic gradients. *The Annals of Statistics*, 45(4):1694–1727, 2017.

Joel Tropp. Freedman's inequality for matrix martingales. *Electronic Communications in Probability*, 16(none):262 – 270, 2011. doi: 10.1214/ECP.v16-1624. URL https://doi.org/10.1214/ECP.v16-1624.

Joel A Tropp. User-friendly tail bounds for sums of random matrices. *Foundations of computational mathematics*, 12(4):389–434, 2012.

G. Yin and Y.M. Zhu. On h-valued robbins-monro processes. *Journal of Multivariate Analysis*, 34:116–140, 1990.

Tianyu Zhang and Noah Simon. A sieve stochastic gradient descent estimator for online nonparametric regression in sobolev ellipsoids. *Annals of statistics*, 50(5):2848, 2022.

Online Appendix

Throughout this Online Appendix, unless otherwise specified, C denotes a generic positive constant whose value may change even within a line. When we append "a.s." to an inequality, we mean that the inequality holds eventually almost surely; that is, for almost all paths of the data, the inequality holds for all sufficiently large indices. For any vector $A = (a_1, \ldots, a_n)^\top$, $\|A\|_2 = \sqrt{\sum_{i=1}^n a_i^2}$ denotes its Euclidean norm. For any matrix $A = (a_{ij})_{m\times n}$, $\|A\|_F = \sqrt{\sum_{i=1}^m \sum_{j=1}^n a_{ij}^2}$ denotes its Frobenius norm. When $f(z)$ is a vector of functions, $\|f\|_\infty = \sup_{z\in\mathcal{Z}} \|f(z)\|_2$.

The following provides a clear table of contents as the guidelines of the appendix.

Appendix Table of Contents

A. More Simulation Results for Point-Wise Marginal Effects

We can also consider point-wise marginal effect $\tau_0(x_0, X) = \nabla_z F_0(z)\theta_0(1) = \nabla_z F_0(z)$, which captures the effect of marginal change of x_1 when index value Z_0 is equal to z. We consider the same simulation design as in the main text, and consider $z \in [-2, 2]$, which is discretized into 20 grid points. Table 7 reports the corresponding statistics for the pointwise estimator. The pointwise estimator follows a non-standard convergence rate that depends on the rate at which the sieve dimension grows. However, since in the simulation we fix the sieve dimension, we use standard FCLT to provide confidence intervals. The point-wise

Table 7. Pointwise Marginal Effects: Normal and Rademacher Regressors

	Cauchy Errors				Skewed Normal Errors			
	$p = 50$		$p = 100$		$p = 50$		$p = 100$	
	$B = 25$	$B = 50$	$B = 25$	$B = 50$	$B = 25$	$B = 50$	$B = 25$	$B = 50$
Bias	0.0004	0.0002	0.0004	0.0003	0.0004	0.0001	0.0003	0.0003
RMSE	0.0076	0.0053	0.0098	0.0063	0.0067	0.0045	0.0083	0.0055
CR	0.9437	0.9373	0.9453	0.9403	0.9425	0.9380	0.9413	0.9425
CI Length	0.0303	0.0206	0.0379	0.0244	0.0258	0.0170	0.0321	0.0207

Note: Results are Monte Carlo averages over 200 simulation replications. All statistics for pointwise marginal effects are averaged across the evaluation grid points. CR denotes coverage rate, and CI Length is the average confidence interval length.

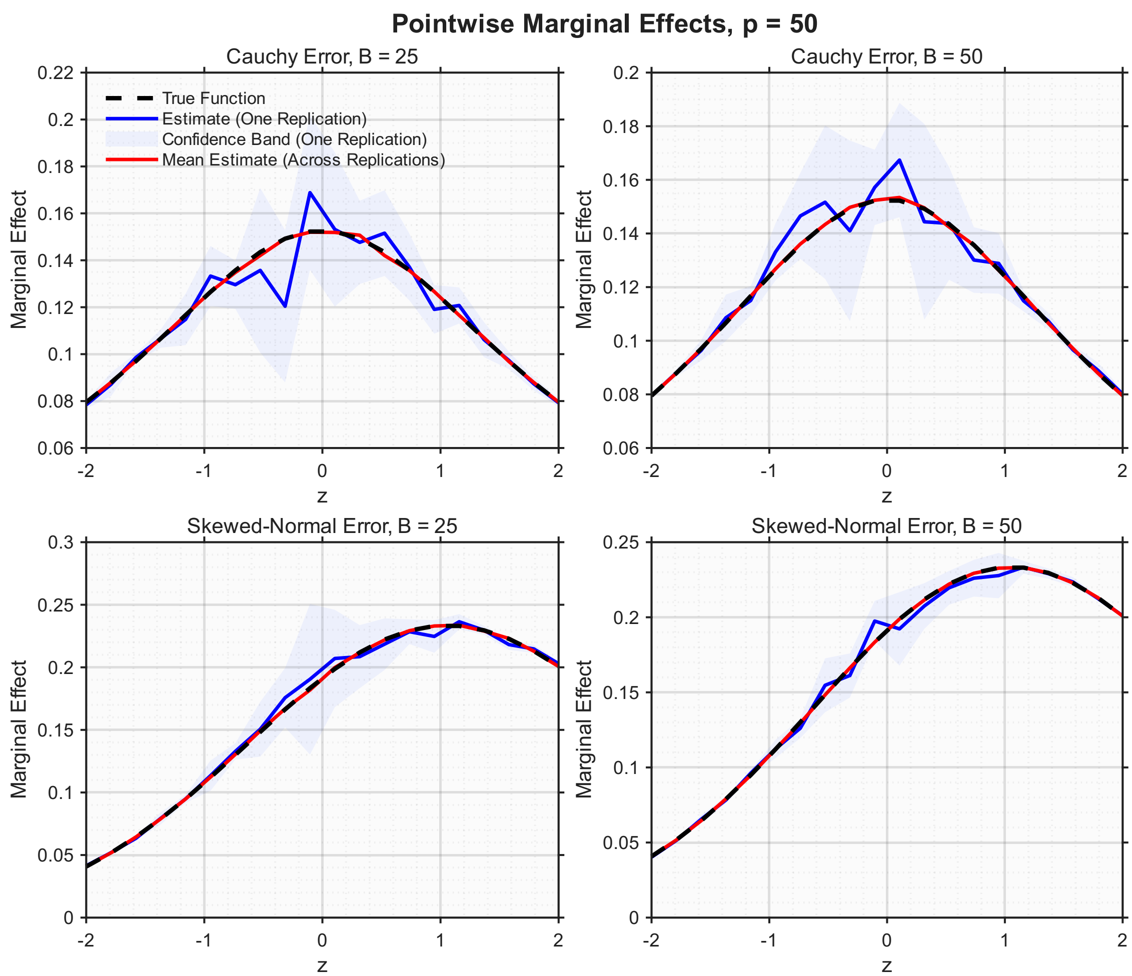


Figure 9. Simulation Results of Point-Wise Marginal Effect: $p = 50$ Case

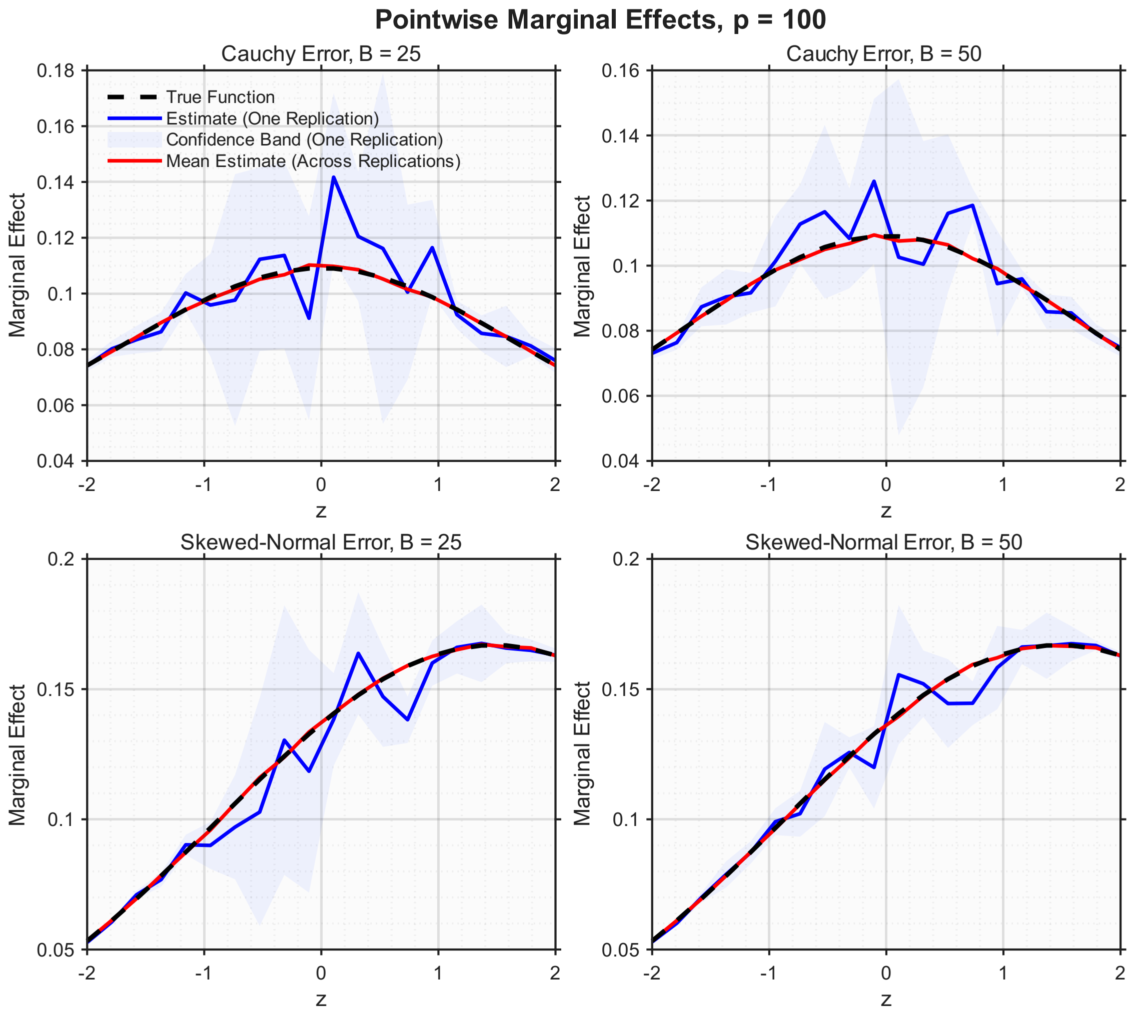


Figure 10. Simulation Results of Point-Wise Marginal Effect: $p = 100$ Case

estimator exhibits modestly larger bias and RMSE than the average estimator, reflecting the slower nonparametric rate, but coverage remains close to the nominal level across all configurations.

Figure 9 and Figure 10 display the estimated point-wise marginal effect trajectory across the grid of values of the index z, together with the 95% random-scaling confidence band from a single representative replication and the average estimate across replications. The semiparametric estimator tracks the true point-wise effect closely throughout the support of z, with confidence bands that are tight in the interior of the support and slightly wider near the boundaries, consistent with standard nonparametric behavior.

Across both functionals, the simulation evidence confirms that the online procedure delivers reliable estimation and valid inference for policy-relevant marginal effects with no additional nonparametric estimation burden beyond the trajectories already produced by the joint

Algorithm 1 Online Warm-Start Learner for θ_0 (Phase I)

Require: streaming batches $\mathcal{W}_k$, $k = 1, \ldots, N$; initial guess $\widehat{\theta}_0 \in \mathbb{R}^p$; kernel $\mathcal{K}$; learning rates $\{\gamma_k\}$; bandwidths $\{h_k\}$; batch size B. **Ensure:** PR average $\overline{\theta}_N$.

1: Initialize $\widehat{\theta}_0$ as supplied and set $\overline{\theta}_0 \leftarrow \widehat{\theta}_0$

2: **for** $k = 1, 2, \ldots, N$ **do**

3: Receive batch $\mathcal{W}_k$ and compute $Z_{i,k}(\widehat{\theta}_{k-1})$, $i = 1, \ldots, B$

4: $\widehat{\theta}_k \leftarrow \widehat{\theta}_{k-1} + \dfrac{\gamma_k}{h_k \cdot B(B-1)} \sum_{i_1 \neq i_2}^{B} \mathcal{K}\left(\dfrac{Z_{i_1,k}(\widehat{\theta}_{k-1}) - Z_{i_2,k}(\widehat{\theta}_{k-1})}{h_k}\right)(Y_{i_1,k} - Y_{i_2,k})(X_{i_1,k} - X_{i_2,k})$ ▷ online update (3.1)

5: $\overline{\theta}_k \leftarrow \dfrac{k-1}{k}\overline{\theta}_{k-1} + \dfrac{1}{k}\widehat{\theta}_k$ ▷ PR average

6: **return** $\overline{\theta}_N$

updating algorithm.

B. Phase I Kernel-Based Warm Start of θ_0

B.1. Pseudo Code

The pseudo code for constructing the warm-start estimator of θ_0 is provided in Algorithm 1.

B.2. Sufficient Conditions and Discussions

Condition 1. *Let $s_{\mathcal{K}} \in \mathbb{N}$ and let the batch size satisfy $B \geq 2$. (i) The link F_0 is bounded and has uniformly bounded derivatives through order $s_{\mathcal{K}}$; (ii) The pair (x_0, X) has a joint density $f(x_0, X)$ with partial derivatives through order $s_{\mathcal{K}}$ in x_0; (iii) There are nonnegative envelopes $g_0, \ldots, g_{s_{\mathcal{K}}}$ such that, for $0 \leq j \leq s_{\mathcal{K}}$,*

$$\sup_{x_0 \in \mathbb{R}} \left| \frac{\partial^j f(x_0, X)}{\partial x_0^j} \right| \leq g_j(X), \qquad \int (1 + \|X\|_2^4) g_j(X) dX < \infty,$$

and $\mathbb{E}\|X\|_2^4 < \infty$; (iv) $\sup_{x_0, X} \mathbb{E}(\varepsilon^4 \mid x_0, X) < \infty$, and $\sigma^2(x_0, X) = \mathbb{E}(\varepsilon^2 \mid x_0, X)$ has a uniformly bounded derivative in x_0; (v) The matrix $\mathcal{V}_{\mathcal{K},0}$ defined below is positive definite.

Condition 2. *The kernel $\mathcal{K}$ is even and satisfies: (i) $\int \mathcal{K}(t)dt = 1$, $\int t^j \mathcal{K}(t)dt = 0$ for $1 \leq j \leq s_{\mathcal{K}} - 1$, and $\int |\mathcal{K}(t)t^j|dt < \infty$ for every $j \geq s_{\mathcal{K}}$; (ii) $\int \mathcal{K}^4(t)|t|^j dt < \infty$ for every $j \geq 0$; and (iii) $\int (\nabla_t \mathcal{K}(t))^2 |t|^j dt < \infty$ for every $j \geq 0$.*

Condition 3. *There exist $-\infty < \underline{z} < \overline{z} < \infty$ and $r_X > 0$ such that for any $z \in [\underline{z}, \overline{z}]$ and X with $\|X\|_2 \leq r_X$, there holds $\nabla_z F_0(z) \geq \underline{c}_F$ and $f(z, X) \geq \underline{c}_f$, where $\underline{c}_F$ and $\underline{c}_f$ are two positive constants.*

Condition 4. *$\gamma_k = \gamma_0 k^{-\alpha_\gamma}, h_k = h_0 k^{-\alpha_h}$, where $\gamma_0, \alpha_\gamma, h_0, \alpha_h > 0$ and $\alpha_\gamma < 1$. Moreover, $s_{\mathcal{K}}\alpha_h > 1/2$, and $2\alpha_\gamma - 3\alpha_h > 1$.*

Remark B.1. *Condition 1 and Condition 2 impose regular conditions on data generating process and kernel functions. Condition 3 guarantees global stability of the algorithm. Note that for Condition 3, X can be normalized to have zero expectation, in which case such condition requires that X has nonvanishing density around its expectation. Moreover, such condition also requires that x_0 has sufficiently large support, which is required for point identification. Finally, Condition 3 requires that all regressors are continuous. For the case where there are discrete regressors, see Section F.2. Condition 4 regulates the bandwidth and learning rate. Specifically, there holds $\sum_{k=1}^{\infty} \gamma_k = \infty$, $\sum_{k=1}^{\infty} \gamma_k^2 h_k^{-3} < \infty$ and $\sum_{k=1}^{\infty} h_k^{2s_{\mathcal{K}}} < \infty$,*

B.3. Main Results: Theorem B.1

This section describes the global convergence properties of the online estimator $\widehat{\theta}_N$ and its PR average $\overline{\theta}_N$ proposed in Section 3.1 of main text. We will formally show that both $\widehat{\theta}_N$ and $\overline{\theta}_N$ almost surely converge to θ_0 regardless of the choice of the starting point, so for almost every path of the data stream, the algorithm locates a neighborhood of θ_0 with arbitrary precision.

To ease our exposition, in this section we continue with the assumption that x_0 and X are all continuous with joint density $f(x_0, X)$. Our results can also be extended to the case where there are discrete regressors, see Section F.2.

Define

$$\mathbb{H}_0 = \int \nabla_z F_0(z) f(z - X_1^\top \theta_0, X_1) f(z - X_2^\top \theta_0, X_2) (X_1 - X_2)(X_1 - X_2)^\top dz dX_1 dX_2,$$

and

$$\mathcal{V}_{\mathcal{K},0} = \int \mathcal{K}^2(t) dt \int \left(\sigma^2(z - X_1^\top \theta_0, X_1) + \sigma^2(z - X_2^\top \theta_0, X_2)\right) \times \\ (X_1 - X_2)(X_1 - X_2)^\top f(z - X_1^\top \theta_0, X_1) f(z - X_2^\top \theta_0, X_2) dz dX_1 dX_2,$$

where $\sigma^2(x_0, X) = \mathbb{E}(\varepsilon^2|x_0, X)$ is the conditional variance. Further, let learning rate γ_k and bandwidth parameter h_k be chosen as

$$\gamma_k = \gamma_0 k^{-\alpha_\gamma}, \quad h_k = h_0 k^{-\alpha_h}, \tag{B.1}$$

with γ_0, h_0, α_h, and α_γ chosen as in Condition 4. For any $\theta \in \mathbb{R}^p$, define $\Delta\theta \equiv \theta - \theta_0$. The following theorem states the theoretical properties of the online estimator $\widehat{\theta}_N$ and $\overline{\theta}_N$, which also proves the first part of Lemma 5.1 in them main text.

Theorem B.1. *Let Condition 1–4 hold. Then $\widehat{\theta}_N \to_{a.s.} \theta_0$ and $\overline{\theta}_N \to_{a.s.} \theta_0$. Further, for every fixed coordinate $j = 1, \ldots, p$,we have that*

$$\limsup_{N\to\infty} e_j^\top \left(\frac{4N^{1+\alpha_h} \log(\log(N))\, \mathcal{V}_{\mathcal{K},0}}{B(B-1)h_0(1+\alpha_h)} \right)^{-1/2} \mathbb{H}_0 N \Delta\overline{\theta}_N = 1, \qquad a.s. \tag{B.2}$$

$$\liminf_{N\to\infty} e_j^\top \left(\frac{4N^{1+\alpha_h} \log(\log(N))\, \mathcal{V}_{\mathcal{K},0}}{B(B-1)h_0(1+\alpha_h)} \right)^{-1/2} \mathbb{H}_0 N \Delta\overline{\theta}_N = -1, \qquad \text{a.s.} \tag{B.3}$$

Finally, there holds

$$\left(\frac{2N^{1+\alpha_h} \mathcal{V}_{\mathcal{K},0}}{B(B-1)h_0(1+\alpha_h)} \right)^{-\frac{1}{2}} \mathbb{H}_0 N \Delta\overline{\theta}_N \to_d \mathcal{N}(0, \mathbb{I}_p). \tag{B.4}$$

Theorem B.1 first states the a.s. convergence of $\widehat{\theta}_N$ and $\overline{\theta}_N$. Hence for almost *all paths* of the data stream, our algorithm will lead to consistent estimators for the unknown parameter θ_0 as long as the number of updates is sufficiently large, regardless of the choice of the initial guess $\widehat{\theta}_0$. Leveraging such global stability property, we can quickly locate a small neighborhood around the true parameter. Theorem B.1 also shows that

$$\|\Delta\overline{\theta}_N\|_2 = O(\sqrt{N^{-1+\alpha_h} \log(\log(N))}), \qquad \text{a.s.}$$

Theorem B.1 is the main result for the warm start estimator of θ_0. In the following, we further obtain the convergence rate of the non-averaged estimator $\widehat{\theta}_N$. In particular, we will show that $\|\Delta\widehat{\theta}_N\|_2 = O(\sqrt{N^{-\alpha_\gamma+\alpha_h} \log(\log(N))}$, a.s.. Since $\alpha_\gamma < 1$, the PR average $\overline{\theta}_N$ converges at faster rate than $\widehat{\theta}_N$. As a result, we recommend using the PR average $\overline{\theta}_N$ to construct the projection space when one needs to be constructed. The proof of Theorem B.1 is lengthy and hence postponed to Appendix F.

Algorithm 2 Online Warm-Start Learner for F_0 (Phase I)

Require: streaming batches $\mathcal{W}_k$, $k = 1, \dots, N$; sieve dimensions $\{J_k\}$; learning rates $\{\eta_k\}$; initial coefficient $\widehat{\beta}_0 \in \mathbb{R}^{J_0}$; index estimates $\{\check{\theta}_{k-1}\}$.

$\overline{\beta}_0 \leftarrow \widehat{\beta}_0$

for $k = 1, 2, \dots, N$ **do**

  Receive batch $\mathcal{W}_k$; $\check{Z}_{i,k} \leftarrow Z_{i,k}(\check{\theta}_{k-1})$, $i = 1, \dots, B$

  Update sieve coefficient ▷ embed via $R^\top_{J_{k-1},J_k}$, then online update (3.4)

$$\widehat{\beta}_k \leftarrow R^\top_{J_{k-1},J_k} \widehat{\beta}_{k-1} + \frac{\eta_k}{B} \sum_{i=1}^{B} \Big(Y_{i,k} - \Psi_{J_k}(\check{Z}_{i,k})^\top R^\top_{J_{k-1},J_k} \widehat{\beta}_{k-1} \Big) \Psi_{J_k}(\check{Z}_{i,k})$$

  Update PR-averaged coefficient ▷ recursive averaging in common sieve space

$$\overline{\beta}_k \leftarrow \frac{1}{k} \widehat{\beta}_k + \frac{k-1}{k} R^\top_{J_{k-1},J_k} \overline{\beta}_{k-1}$$

return $\overline{F}_N(\cdot) = \Psi_{J_N}(\cdot)^\top \overline{\beta}_N$

C. Phase I Sieve LS Warm Start Estimator of F_0

C.1. Pseudo Code

The pseudo code is provided in Algorithm 2.

C.2. Sufficient Conditions and Discussions

Following Chen and Christensen (2015), we define the L_2-projection of function F_0 onto $\mathcal{S}_J$ as

$$\mathbb{P}_J(F_0)(z) = \Psi_J(z)^\top \Gamma_J^{-1} \mathbb{E}\left[\Psi_J(Z_0) F_0(Z_0)\right] \equiv \Psi_J(z)^\top \beta_{J,0}, \tag{C.1}$$

where recall that $\mathcal{S}_J = \{\sum_{j=1}^J b_j \psi_{J,j}(\cdot) : b_1, \dots, b_J \in \mathbb{R}\}$ and $\Gamma_J = \mathbb{E}[\Psi_J(Z_0)\Psi_J(Z_0)^\top]$. By definition, $\beta_{J,0}$ is the $L^2(P_{Z_0})$ projection coefficient. Then for any function $g \in \mathcal{S}_J$, we have that $\|\mathbb{P}_J(F_0) - F_0\|_\infty = \|\mathbb{P}_J(F_0 - g) - (F_0 - g)\|_\infty$, so the sup-norm sieve approximation error of L_2-projection can be bounded by

$$\|\mathbb{P}_J(F_0) - F_0\|_\infty \le \inf_{g \in \mathcal{S}_J} \|F_0 - g\|_\infty \left(1 + \frac{\|\mathbb{P}_J(F_0 - g)\|_\infty}{\|F_0 - g\|_\infty}\right). \tag{C.2}$$

One implicit condition we impose is that $\sup_J \sup_{g\in\mathcal{S}_J:0<\|F_0-g\|_\infty<\infty} \|F_0 - g\|_\infty^{-1}\|\mathbb{P}_J(F_0 - g)\|_\infty < \infty$, which can be guaranteed when the operator norm of $\mathbb{P}_J$ is uniformly bounded for all J. This condition holds, by Chen and Christensen (2015), for compactly supported wavelets and splines. We also impose sup-norm approximation error rate as $\inf_{g\in\mathcal{S}_J} \|F_0 - g\|_\infty \le C_{F_0} J^{-s}$ where $s > 0$ is approximation error rate of F_0. Finally, we impose that the derivative function also guarantees uniform approximation error of order $1 - s$. The above is summarized in the following condition.

Condition 5. *There is some positive integer s and constant C_{F_0} such that for all J, $\|F_0(\cdot) - \mathbb{P}_J(F_0)(\cdot)\|_\infty \le C_{F_0} J^{-s}$ and $\|\nabla_z[F_0(\cdot) - \mathbb{P}_J(F_0)(\cdot)]\|_\infty \le C_{F_0} J^{1-s}$.*

Remark C.1. *Condition 5 does not impose any restriction on the support of $Z_0 \equiv x_0 + X^\top\theta_0$. However, when spline basis functions are used, the uniform approximation rates in Condition 5(ii) are typically established for functions defined on compact domains. To accommodate unbounded support, we can simply apply a transformation that maps $\mathbb{R}$ into a bounded interval (Bierens 2014). For instance, the transformation $\widetilde{Z}_0 = \arctan(Z_0)$ ensures $\widetilde{Z}_0 \in (-\pi/2, \pi/2)$ regardless of the support of Z_0. Let $\{\psi_j(\cdot)\}_{j\ge1}$ be spline basis functions defined on $(-\pi/2, \pi/2)$. We can approximate the transformed function $\widetilde{F}_0(u) := F_0(\tan(u)), \quad u \in (-\pi/2, \pi/2)$, using $\{\psi_j(u)\}_{j\ge1}$. Equivalently, this corresponds to using the sieve $\{\psi_j(\arctan(\cdot))\}$ to approximate $F_0(\cdot)$ on $\mathbb{R}$. Assume in addition that $\widetilde{F}_0$ extends to the closed transformed interval with the smoothness and boundary behavior required in Condition 5. Under this reparameterization, Condition 5 can be imposed on $\widetilde{F}_0$, allowing the framework to accommodate unbounded support of Z_0.*

Recall that $\mu_0(\theta_0, z) = \mathbb{E}(X|Z_0 = z)$. We choose

$$\eta_k = \eta_0 k^{-\alpha_\eta}, \qquad J_k = [J_0 k^{\alpha_\dagger}],$$

where $\eta_0, \alpha_\eta, \alpha_\dagger > 0$ and J_0 is a positive integer. We further impose the following conditions on the data-generating process, the sieve, and the predictable first-stage estimator of θ_0.

Condition 6. *(i) $\mathbb{E}\|X\|_2^4 < \infty$, $\max\{\|F_0\|_\infty, \|\nabla_z F_0\|_\infty, \|\nabla_{zz} F_0\|_\infty\} < \infty$, $\sup_{x_0,X} \mathbb{E}(\varepsilon^2|x_0, X) < \infty$ and $\mathbb{E}|\varepsilon|^\kappa < \infty$ for some $\kappa > 2 + \max\{\frac{2}{\alpha_\eta}, \frac{2(1+\alpha_\dagger)}{1-\alpha_\dagger}\}$; (ii) There holds $\sup_z \|(\nabla_z F_0(z))\mu_0(\theta_0, z)\|_2 < \infty$, and moreover, $\sup_z \|\mathbb{P}_J[(\nabla_z F_0(\cdot))\mu_0(\theta_0, \cdot)](z) - (\nabla_z F_0(z))\mu_0(\theta_0, z)\|_2 \le C_\mu J^{-s_\mu}$ for some $C_\mu, s_\mu > 0$ and all positive integers J, with the projection applied componentwise.*

Remark C.2. *Condition 6 regulates the tail behaviors of X and ε. The moment condition on ε is used to build almost sure convergence and sup-norm rate for the sieve estimator, and* *for showing the FCLT of the online marginal effect estimator in Theorem 5.3. We mention that if we only pursue the rate of convergence in probability for online sieve estimator, then $\kappa > 2 + \frac{2\alpha_\dagger}{1-\alpha_\dagger}$ alone as in Chen and Christensen (2015) will suffice (note that under optimal sieve rate $1/(2s+1)$, the condition reduces to $2 + 1/s$, which is exactly the dimension/smoothness condition in Chen and Christensen (2015)). When we further pursue optimal a.s. convergence rates of sieve estimator, we additionally need $\kappa > 2+\max\{\frac{2}{\alpha_\eta}, \frac{2(1+\alpha_\dagger)}{1-\alpha_\dagger}\}$.*

Condition 7. *(i) $\|\Psi_J\|_\infty \leq CJ^{1/2}$ and $\|\nabla_z \Psi_J\|_\infty \leq CJ^{3/2}$; (ii) $0 < \inf_J \lambda_{\min}(\Gamma_J) \leq \sup_J \lambda_{max}(\Gamma_J) < \infty$; (iii) Let $\mathcal{Z}_0$ be the support of Z_0. For each positive integer L, there exists $z_1, \cdots, z_L \in \mathcal{Z}_0$ such that $\sup_{z\in\mathcal{Z}_0} \min_{1\leq l\leq L} \|\Psi_J(z) - \Psi_J(z_l)\| \leq C_\Psi J^{3/2} L^{-1}$ for all J, where $C_\Psi > 0$ is a constant.*

Remark C.3. *Condition 7 (i) and (ii) are natural for normalized B-spline basis. Note that if $\mathcal{Z}_0$ is bounded, the Condition 7 (iii) automatically holds given Condition 7 (i). Otherwise, when $\mathcal{Z}_0$ is unbounded, we consider sieve functions of form $\{\psi_j(\arctan(\cdot))\}$ discussed in Remark C.1.*

Condition 8. *Let $\mathcal{F}_k = \sigma(\mathcal{W}_1, \ldots, \mathcal{W}_k)$. For specified constants $0 < \alpha_\theta \leq 1/2$ and $c_\theta \geq 0$, $\|\Delta\check{\theta}_k\|_2 = O_{\text{a.s.}}(k^{-\alpha_\theta}(\log k)^{c_\theta})$. For each k, $\check{\theta}_k$ is $\mathcal{F}_k$-measurable.*

Condition 9. *The exponents lie in the strict interior*

$$\frac{1}{2} < \alpha_\eta, \qquad \max\{1 - 2s\alpha_\dagger, 1 + 5\alpha_\dagger - 2\alpha_\theta\} < \alpha_\eta < \min\{2\alpha_\theta, 1 - 3\alpha_\dagger, (2s+1)\alpha_\dagger\}.$$

Throughout the rest of the Appendix, to avid complicated discussion on $s\alpha_\dagger$, we always assume that $s\alpha_\dagger < 1$ and $\alpha_\theta + s_\mu\alpha_\dagger < 1$. We note that violations of either assumptions will not affect any of the following results.

C.3. Main Results: Theorem C.1

This section develops the sup-norm almost-sure convergence rate of the warm-start estimator $\overline{F}_N$ defined in Section 3.2 of main text. The argument also applies to generated-regressor online sieve estimation with predictable first-stage estimators satisfying Condition 8.

Recall that $Z_0 = x_0 + X^\top\theta_0$ and $\Gamma_J = \mathbb{E}\left[\Psi_J(Z_0)\Psi_J(Z_0)^\top\right]$. For any $N \geq 1$, define $\Delta\overline{\beta}_N = \overline{\beta}_N - \beta_{J_N,0}$, where $\beta_{J_N,0}$ is the pseudo-true sieve coefficient and η_k, J_k are as defined above. Lemma 5.2 is implied by the following results.

Theorem C.1. *Let Condition 5–9 hold. Assume also that $\widehat{\beta}_0$ is $\mathcal{F}_0$-measurable and $\mathbb{E}\|\widehat{\beta}_0\|_2^2 < \infty$. Then*

$$\Delta\overline{\beta}_N = \frac{1}{N}\sum_{k=1}^{N} R_{J_k,J_N}^\top \Gamma_{J_k}^{-1}\frac{1}{B}\sum_{i=1}^{B}\varepsilon_{i,k}\Psi_{J_k}(Z_{0,i,k}) - \frac{1}{N}\sum_{k=1}^{N} R_{J_k,J_N}^\top \Gamma_{J_k}^{-1}\mathbb{E}[(\nabla_z F_0(Z_0))\Psi_{J_k}(Z_0)X^\top]\Delta\check{\theta}_{k-1}$$
$$+\frac{1}{N}\sum_{k=1}^{N}\left(R_{J_k,J_N}^\top\beta_{J_k,0} - \beta_{J_N,0}\right) + r_N, \qquad \|r_N\|_2 = o_{\text{a.s.}}(N^{-1/2})$$

Moreover, defining

$$A_{\mu,N} := \frac{1}{N}\sum_{k=1}^{N} J_k^{-s_\mu}\|\Delta\check{\theta}_{k-1}\|_2,$$

we have

$$\left\|\overline{F}_N - F_0\right\|_\infty = O_{\text{a.s.}}\left(\sqrt{\frac{J_N\log N}{N}} + \frac{1}{N}\sum_{k=1}^{N} J_k^{-s} + \left\|\frac{1}{N}\sum_{k=1}^{N}\Delta\check{\theta}_{k-1}\right\|_2 + A_{\mu,N}\right). \tag{C.3}$$

where $s > 0$ is the approximation error rate of F_0. In particular, with $J_k \asymp k^{\alpha_\dagger}$, $\frac{1}{N}\sum_{k=1}^{N} J_k^{-s} = O(N^{-s\alpha_\dagger})$ and $A_{\mu,N} = O_{\text{a.s.}}(N^{-\alpha_\theta - s_\mu\alpha_\dagger}(\log N)^{c_\theta})$.

Theorem C.1 is the key result of online sieve estimation with possibly generated regressor. It provides the first-order expansion of the average online estimator $\overline{\beta}_N$, which contains four terms. The first term is comparable to the variance term of the full-sample sieve estimator (e.g., Chen and Christensen (2015)), but since in the k-th round of update we only use J_k sieve functions, the effective variance component in the k-th round is $\varepsilon_{i,k}R_{J_k,J_N}^\top\Gamma_{J_k}^{-1}\Psi_{J_k}(Z_{0,i,k})$ instead of $\varepsilon_{i,k}\Gamma_{J_N}^{-1}\Psi_{J_N}(Z_{0,i,k})$. This highlights the online learning structure of the algorithm. The second term describes the first-order impacts of using the generated regressor $\check{Z}_{i,k}$ for sieve estimation. The third term describes the impacts of changing pseudo true sieve coefficients and re-embedding when the sieve dimension increases. The last term collects all the high-order remainders.

When we consider the sup-norm estimation error of F_0, apart from the last two terms capturing the average index error and the projection error $A_{\mu,N}$, the stochastic and approximation components match the familiar sieve benchmark under a compatible tuning choice. The usual minimax choice $\alpha_\dagger = 1/(2s+1)$ is admissible here when it lies in the strict region of Condition 9. With the fastest permitted first stage, $\alpha_\theta = 1/2$, this requires

$$s > \frac{7}{2}, \qquad \max\{1/2, 5\alpha_\dagger\} < \alpha_\eta < 1 - 3\alpha_\dagger.$$

Under these restrictions, we can choose $J_N \asymp N^{1/(2s+1)}$, then $J_N^{\frac{1}{2}} N^{-\frac{1}{2}} \log^{\frac{1}{2}}(N) + J_N^{-s} \asymp N^{-\frac{s}{2s+1}} \log^{\frac{1}{2}}(N)$, and this error rate matches the usual minimax benchmark up to poly-log terms (Stone 1982; Belloni et al. 2015; Chen and Christensen 2015).

C.4. Proof of Theorem C.1

C.4.1. Additional Lemma

We first state a useful lemma.

Lemma C.1. *Let nonnegative random sequence $Q_0, Q_1, \cdots$ satisfy $\mathbb{E}Q_0^2 < \infty$ and*

$$Q_N \leq (1 - CN^{-\alpha_1})Q_{N-1} + CN^{\alpha_2}T_N$$

where $\frac{1}{2} < \alpha_1 < 1$ and $\alpha_2 \in \mathbb{R}$, and $T_N \geq 0$ satisfies $\mathbb{E}_{N-1}T_N^2 \leq C_T < \infty$ for all N with $\mathbb{E}_{N-1}$ being the expectation conditioned on $Q_0, T_1, \cdots, T_{N-1}$. Then

$$Q_N = O(N^{\alpha_2+\alpha_1}), \qquad a.s$$

Under the same conditions, if instead

$$Q_N \leq Q_{N-1} + CN^{\alpha_2}T_N$$

then for any $r > 1$,

$$Q_N = \begin{cases} O_{\text{a.s.}}(N^{1+\alpha_2}), & \alpha_2 > -1, \\ O_{\text{a.s.}}(\log^r(N)), & \alpha_2 = -1, \\ O_{\text{a.s.}}(1), & \alpha_2 < -1. \end{cases}$$

Proof. Assume that $1 - CN^{-\alpha_1} \in [0, 1]$; otherwise we start from some sufficiently large k_0. Define $\varepsilon_{T,N} = T_N - \mathbb{E}_{N-1}T_N$ and let $\overline{Q}_N$ be defined as $\overline{Q}_N = (1 - CN^{-\alpha_1})\overline{Q}_{N-1} + CN^{\alpha_2}T_N$ with $\overline{Q}_0 = Q_0$. Recursion gives $0 \leq Q_N \leq \overline{Q}_N$. Decompose $\overline{Q}_N = Q_{1,N} + Q_{2,N}$, where

$$\begin{aligned} Q_{1,N} &= (1 - CN^{-\alpha_1})Q_{1,N-1} + CN^{\alpha_2}\mathbb{E}_{N-1}T_N, \\ Q_{2,N} &= (1 - CN^{-\alpha_1})Q_{2,N-1} + CN^{\alpha_2}\varepsilon_{T,N}, \end{aligned}$$

with $Q_{1,0} = \mathbb{E}Q_0$ and $Q_{2,0} = Q_0 - \mathbb{E}Q_0$. Let $r = \alpha_1 + \alpha_2$. Since $\mathbb{E}_{N-1}T_N \leq C_T^{1/2}$, iteration gives

$$|Q_{1,N}| \le Ce^{-CN^{1-\alpha_1}} + C\sum_{k=1}^{N} e^{-C(N^{1-\alpha_1}-k^{1-\alpha_1})}k^{\alpha_2}.$$

By Lemma F.5, we have that $|Q_{1,N}| = O(N^{\alpha_1+\alpha_2}) = O(N^r)$. For the martingale part, note that

$$N^{-r}Q_{2,N} = (1 - CN^{-\alpha_1})\left(\frac{N-1}{N}\right)^r (N-1)^{-r}Q_{2,N-1} + CN^{-\alpha_1}\varepsilon_{T,N},$$

Because $\alpha_1 < 1$, the $N^{-\alpha_1}$ contraction dominates the $O(N^{-1})$ drift, so

$$\mathbb{E}_{N-1}[N^{-r}Q_{2,N}]^2 \le (1 - CN^{-\alpha_1})[(N-1)^{-r}Q_{2,N-1}]^2 + CN^{-2\alpha_1}.$$

Now $\sum_N N^{-2\alpha_1} < \infty$ and $\sum_N N^{-\alpha_1} = \infty$. Robbins–Siegmund implies that $(N^{-r}Q_{2,N})^2$ converges and that $\sum_N N^{-\alpha_1}[N^{-r}Q_{2,N}]^2 < \infty$; hence its limit is zero. Therefore $Q_{2,N} = o_{\text{a.s.}}(N^r)$.

For the recursion without contraction, use the same domination and write

$$Q_{1,N} = Q_{1,0} + C\sum_{k=1}^{N} k^{\alpha_2}\mathbb{E}_{k-1}T_k, \qquad Q_{2,N} = Q_{2,0} + C\sum_{k=1}^{N} k^{\alpha_2}\varepsilon_{T,k}.$$

The deterministic part has the exact three orders

$$Q_{1,N} = \begin{cases} O(N^{1+\alpha_2}), & \alpha_2 > -1, \\ O(\log N), & \alpha_2 = -1, \\ O(1), & \alpha_2 < -1. \end{cases}$$

The martingale quadratic variation is bounded by $C\sum_{k\le N} k^{2\alpha_2}$. A dyadic martingale maximal inequality and Borel–Cantelli give, for every fixed $r_0 > 1$,

$$|Q_{2,N}| = \begin{cases} O_{\text{a.s.}}\{N^{1/2+\alpha_2}(\log N)^{r_0/2}\}, & \alpha_2 > -1/2, \\ O_{\text{a.s.}}\{(\log N)^{(r_0+1)/2}\}, & \alpha_2 = -1/2, \\ O_{\text{a.s.}}(1), & \alpha_2 < -1/2, \end{cases}$$

where in the last case the martingale converges in L_2 and almost surely. Combining $Q_N \le Q_{1,N} + |Q_{2,N}|$ proves the second assertion, including the logarithmic boundary $\alpha_2 = -1$. □

The following proof of Theorem C.1 will be divided into three steps.

C.4.2. Obtaining Initial Convergence Rate for $\|\Delta\widehat{\beta}_k\|_2$

Summary: This part will develop the following result:

$$\|\Delta\widehat{\beta}_k\|_2 = O\left(k^{\frac{1-\alpha_\eta-\alpha_\dagger\underline{\alpha}_F}{2}}\log^{\frac{r}{2}}(k)\right), \qquad \text{a.s.}$$

where $\underline{\alpha}_F = \min\{2s-1, 2\alpha_\theta/\alpha_\dagger - 3, \alpha_\eta/\alpha_\dagger - 1\} > 0$ *and* $r > 1 + 2c_\theta$.

To derive an initial a.s. convergence rate for $\|\Delta\widehat{\beta}_k\|_2$, we first provide a bound for $\mathbb{E}_{k-1}\|\Delta\widehat{\beta}_k\|_2^2$ based on $\|\Delta\widehat{\beta}_{k-1}\|_2^2$. Such bound will allow us to use the convergence result in Robbins and Siegmund (1971). Note that according to the definition of R_{J_1,J_2}, we have that

$$\widehat{\beta}_k = R_{J_{k-1},J_k}^\top\widehat{\beta}_{k-1} + \frac{\eta_k}{B}\sum_{i=1}^{B}\left(Y_{i,k} - \Psi_{J_k}(\check{Z}_{i,k})^\top R_{J_{k-1},J_k}^\top\widehat{\beta}_{k-1}\right)\Psi_{J_k}(\check{Z}_{i,k}),$$

for all k. So

$$\begin{aligned}
\widehat{\beta}_k - \beta_{J_k,0} = {} & R_{J_{k-1},J_k}^\top\left(\widehat{\beta}_{k-1} - \beta_{J_{k-1},0}\right) + \left(R_{J_{k-1},J_k}^\top\beta_{J_{k-1},0} - \beta_{J_k,0}\right) \\
& + \frac{\eta_k}{B}\sum_{i=1}^{B}\varepsilon_{i,k}\Psi_{J_k}\left(\check{Z}_{i,k}\right) + \frac{\eta_k}{B}\sum_{i=1}^{B}\left(F_0\left(Z_{0,i,k}\right) - F_0\left(\check{Z}_{i,k}\right)\right)\Psi_{J_k}\left(\check{Z}_{i,k}\right) \\
& + \frac{\eta_k}{B}\sum_{i=1}^{B}\left(F_0(\check{Z}_{i,k}) - \mathbb{P}_{J_k}(F_0)(\check{Z}_{i,k})\right)\Psi_{J_k}\left(\check{Z}_{i,k}\right) \\
& + \frac{\eta_k}{B}\sum_{i=1}^{B}\Psi_{J_k}\left(\check{Z}_{i,k}\right)\Psi_{J_k}\left(\check{Z}_{i,k}\right)^\top\left(\beta_{J_k,0} - R_{J_{k-1},J_k}^\top\beta_{J_{k-1},0}\right) \\
& + \frac{\eta_k}{B}\sum_{i=1}^{B}\Psi_{J_k}\left(\check{Z}_{i,k}\right)\Psi_{J_k}\left(\check{Z}_{i,k}\right)^\top R_{J_{k-1},J_k}^\top\left(\beta_{J_{k-1},0} - \widehat{\beta}_{k-1}\right)
\end{aligned}$$

Define $\check{\Gamma}_{J,k} = \frac{1}{B}\sum_{i=1}^{B} \Psi_J(\check{Z}_{i,k})\Psi_J(\check{Z}_{i,k})^\top$, we have that

$$
\begin{aligned}
\Delta\widehat{\beta}_k &= \left(\mathbb{I}_{J_k} - \eta_k \check{\Gamma}_{J_k,k}\right) R_{J_{k-1},J_k}^\top \Delta\widehat{\beta}_{k-1} + \left(\mathbb{I}_{J_k} - \eta_k \check{\Gamma}_{J_k,k}\right)\left(R_{J_{k-1},J_k}^\top \beta_{J_{k-1},0} - \beta_{J_k,0}\right) \\
&\quad + \frac{\eta_k}{B}\sum_{i=1}^{B} \varepsilon_{i,k}\Psi_{J_k}\left(Z_{0,i,k}\right) + \frac{\eta_k}{B}\sum_{i=1}^{B}\left(F_0(Z_{0,i,k}) - \mathbb{P}_{J_k}(F_0)(Z_{0,i,k})\right)\Psi_{J_k}\left(Z_{0,i,k}\right) \\
&\quad + \eta_k \left[\underbrace{\frac{1}{B}\sum_{i=1}^{B}\left(F_0\left(Z_{0,i,k}\right) - F_0\left(\check{Z}_{i,k}\right)\right)\Psi_{J_k}\left(\check{Z}_{i,k}\right)}_{\zeta_{1,k}} + \underbrace{\frac{1}{B}\sum_{i=1}^{B}\varepsilon_{i,k}(\Psi_{J_k}\left(\check{Z}_{i,k}\right) - \Psi_{J_k}\left(Z_{0,i,k}\right))}_{\zeta_{2,k}} \right. \\
&\quad \left. \underbrace{+\frac{1}{B}\sum_{i=1}^{B}\left(\left(F_0(\check{Z}_{i,k}) - \mathbb{P}_{J_k}(F_0)(\check{Z}_{i,k})\right)\Psi_{J_k}\left(\check{Z}_{i,k}\right) - \left(F_0(Z_{0,i,k}) - \mathbb{P}_{J_k}(F_0)(Z_{0,i,k})\right)\Psi_{J_k}\left(Z_{0,i,k}\right)\right)}_{\zeta_{3,k}} \right].
\end{aligned}
$$

Further define $\widehat{\Gamma}_{J,k} = \frac{1}{B}\sum_{i=1}^{B} \Psi_J(Z_{0,i,k})\Psi_J(Z_{0,i,k})^\top$. Since $\|\Psi_J\|_\infty \le C_\Psi J^{\frac{1}{2}}$ and $\|\nabla_z \Psi_J\|_\infty \le C_\Psi J^{\frac{3}{2}}$, we have that

$$
\left\|\widehat{\Gamma}_{J_k,k} - \check{\Gamma}_{J_k,k}\right\|_F \le \frac{CJ_k^2}{B}\sum_{i=1}^{B} \|X_{i,k}\|_2 \|\Delta\check{\theta}_{k-1}\|_2,
$$

$$
\|\zeta_{1,k}\|_2 \le \frac{CJ_k^{\frac{1}{2}}}{B}\sum_{i=1}^{B}\left|F_0\left(Z_{0,i,k}\right) - F_0\left(\check{Z}_{i,k}\right)\right| \le CJ_k^{\frac{1}{2}}\frac{1}{B}\sum_{i=1}^{B}\|X_{i,k}\|_2\|\Delta\check{\theta}_{k-1}\|_2,
$$

$$
\|\zeta_{2,k}\|_2 \le \frac{CJ_k^{\frac{3}{2}}}{B}\sum_{i=1}^{B}|\varepsilon_{i,k}|\|X_{i,k}\|_2\|\Delta\check{\theta}_{k-1}\|_2.
$$

Moreover, since

$$
\begin{aligned}
\zeta_{3,k} = \frac{1}{B}\sum_{i=1}^{B}\int_0^1 & \left(\nabla_z[F_0(z) - \mathbb{P}_{J_k}(F_0)(z)]\Psi_{J_k}(z)\right. \\
& \left. + [F_0(z) - \mathbb{P}_{J_k}(F_0)(z)]\nabla_z\Psi_{J_k}(z)\right)\big|_{z = Z_{0,i,k} + \tau(\check{Z}_{i,k} - Z_{0,i,k})}\, d\tau\,(\check{Z}_{i,k} - Z_{0,i,k}),
\end{aligned}
$$

and Condition 5 requires that $\|\nabla_z[F_0(z) - \mathbb{P}_J(F_0)(z)]\|_\infty \le C_{F_0}J^{1-s}$, we have that

$$
\|\zeta_{3,k}\|_2 \le \frac{CJ_k^{\frac{3}{2}-s}}{B}\sum_{i=1}^{B}\|X_{i,k}\|_2\|\Delta\check{\theta}_{k-1}\|_2.
$$

Now we obtain several useful results. First note that

$$\begin{aligned}
&\mathbb{E}_{k-1}\left[\left(\mathbb{I}_{J_k}-\eta_k\widehat{\Gamma}_{J_k,k}+\eta_k\left(\widehat{\Gamma}_{J_k,k}-\check{\Gamma}_{J_k,k}\right)\right)^\top\left(\mathbb{I}_{J_k}-\eta_k\widehat{\Gamma}_{J_k,k}+\eta_k\left(\widehat{\Gamma}_{J_k,k}-\check{\Gamma}_{J_k,k}\right)\right)\right]\\
&=\mathbb{I}_{J_k}-2\eta_k\Gamma_{J_k}+2\eta_k\mathbb{E}_{k-1}\left(\widehat{\Gamma}_{J_k,k}-\check{\Gamma}_{J_k,k}\right)+\eta_k^2\mathbb{E}_{k-1}\left(\check{\Gamma}^2_{J_k,k}\right).
\end{aligned}$$

We also show that

$$\|R^\top_{J_1,J_2}\|_{\text{op}}\le\sqrt{\frac{\sup_J\overline{\lambda}(\Gamma_J)}{\inf_J\underline{\lambda}(\Gamma_J)}}$$

holds for any $J_1<J_2$. This is because, note that $\Psi_{J_1}(Z_0)=R_{J_1,J_2}\Psi_{J_2}(Z_0)$. Then $\Gamma_{J_1}=\mathbb{E}[\Psi_{J_1}(Z_0)\Psi_{J_1}(Z_0)^\top]=R_{J_1,J_2}\mathbb{E}[\Psi_{J_2}(Z_0)\Psi_{J_2}(Z_0)^\top]R^\top_{J_1,J_2}=R_{J_1,J_2}\Gamma_{J_2}R^\top_{J_1,J_2}$. For any vector a, due to the lower and upper boundedness of the eigenvalues of Γ_J with respect to J, we have that

$$\sup_J\lambda_{\max}(\Gamma_J)\|a\|_2^2\ge a^\top\Gamma_{J_1}a=(R^\top_{J_1,J_2}a)^\top\Gamma_{J_2}(R^\top_{J_1,J_2}a)\ge\inf_J\lambda_{\min}(\Gamma_J)\|R^\top_{J_1,J_2}a\|_2^2$$

So

$$\|R^\top_{J_1,J_2}a\|_2^2\le\frac{\sup_J\lambda_{\max}(\Gamma_J)}{\inf_J\lambda_{\min}(\Gamma_J)}\|a\|_2^2$$

This proves the uniform boundedness of $\|R_{J_1,J_2}\|_{\text{op}}$.

Given the above results, due to the lower and upper boundedness of the eigenvalues of Γ_J with respect to J and $\mathbb{E}\|X\|_2<\infty$, for k sufficiently large we have that

$$\begin{aligned}
&\mathbb{E}_{k-1}\left\|\left(\mathbb{I}_{J_k}-\eta_k\check{\Gamma}_{J_k,k}\right)R^\top_{J_{k-1},J_k}\Delta\widehat{\beta}_{k-1}\right\|_2^2\\
&\le\|\mathbb{I}_{J_k}-2\eta_k\Gamma_{J_k}\|_{\text{op}}\|R^\top_{J_{k-1},J_k}\Delta\widehat{\beta}_{k-1}\|_2^2+2\eta_k\left(\mathbb{E}_{k-1}\|\widehat{\Gamma}_{J_k,k}-\check{\Gamma}_{J_k,k}\|_F\right)\|R^\top_{J_{k-1},J_k}\Delta\widehat{\beta}_{k-1}\|_2^2\\
&+\eta_k^2\left(\mathbb{E}_{k-1}\|\check{\Gamma}_{J_k,k}\|_F^2\right)\|R^\top_{J_{k-1},J_k}\Delta\widehat{\beta}_{k-1}\|_2^2\\
&\le\left(1-C\eta_k+C\eta_kJ_k^2\|\Delta\check{\theta}_{k-1}\|_2+C\eta_k^2J_k^2+C\mathbf{1}_k\right)\|\Delta\widehat{\beta}_{k-1}\|_2^2,
\end{aligned}$$

where $\mathbf{1}_k=\mathbf{1}(J_k>J_{k-1})$. When $\alpha_\theta>2\alpha_\dagger$ and $\alpha_\eta>2\alpha_\dagger$, as implied by the strict rate region, we have that

$$\mathbb{E}_{k-1}\left\|\left(\mathbb{I}_{J_k}-\eta_k\check{\Gamma}_{J_k,k}\right)R^\top_{J_{k-1},J_k}\Delta\widehat{\beta}_{k-1}\right\|_2^2\le\left(1-C\eta_k+C\mathbf{1}_k\right)\|\Delta\widehat{\beta}_{k-1}\|_2^2,\qquad\text{a.s.}$$

Moreover,

$$\mathbb{E}_{k-1}\left\|\frac{\eta_k}{B}\sum_{i=1}^{B}\left(F_0(Z_{0,i,k})-\mathbb{P}_{J_k}(F_0)(Z_{0,i,k})\right)\Psi_{J_k}\left(Z_{0,i,k}\right)+\frac{\eta_k}{B}\sum_{i=1}^{B}\varepsilon_{i,k}\Psi_{J_k}\left(Z_{0,i,k}\right)\right\|_2^2\leq C\eta_k^2 J_k,$$

and, because $\mathbb{E}\|X_{i,k}\|_2^2<\infty$,

$$\mathbb{E}_{k-1}\left\|\eta_k(\zeta_{1,k}+\zeta_{2,k}+\zeta_{3,k})\right\|_2^2\leq C\eta_k^2 J_{k-1}^3\|\Delta\check{\theta}_{k-1}\|_2^2.$$

Note that for any constants a,b, we have that $|ab|\leq\frac{a^2}{2c}+\frac{cb^2}{2}$ for arbitrary $c>0$. Using this result, we have that

$$\begin{aligned}
&\Bigg|\mathbb{E}_{k-1}\Bigg[2\left(\left(\mathbb{I}_{J_k}-\eta_k\check{\Gamma}_{J_k,k}\right)R_{J_{k-1},J_k}^{\top}\Delta\widehat{\beta}_{k-1}\right)^{\top}\times\\
&\qquad\frac{\eta_k}{B}\sum_{i=1}^{B}\left(F_0(Z_{0,i,k})-\mathbb{P}_{J_k}(F_0)(Z_{0,i,k})\right)\Psi_{J_k}\left(Z_{0,i,k}\right)+\frac{\eta_k}{B}\sum_{i=1}^{B}\varepsilon_{i,k}\Psi_{J_k}\left(Z_{0,i,k}\right)\Bigg]\Bigg|\\
&\leq c\eta_k\left(1-C\eta_k+C\mathbf{1}_k\right)\|\Delta\widehat{\beta}_{k-1}\|_2^2+c^{-1}C\eta_k J_k^{-2s+1},
\end{aligned}$$

where c is a positive constant and can be arbitrarily chosen. We also have that, because, $\mathbb{E}\|X_{i,k}\|_2^2<\infty$,

$$\begin{aligned}
&\left|\mathbb{E}_{k-1}\left[2\left(\left(\mathbb{I}_{J_k}-\eta_k\check{\Gamma}_{J_k,k}\right)R_{J_{k-1},J_k}^{\top}\Delta\widehat{\beta}_{k-1}\right)^{\top}\eta_k\left(\zeta_{1,k}+\zeta_{2,k}+\zeta_{3,k}\right)\right]\right|\\
&\leq c\eta_k\left(1-C\eta_k+C\mathbf{1}_k\right)\|\Delta\widehat{\beta}_{k-1}\|_2^2+c^{-1}C\eta_k J_k^3\|\Delta\check{\theta}_{k-1}\|_2^2,
\end{aligned}$$

and

$$\begin{aligned}
&\Bigg|\mathbb{E}_{k-1}\Bigg[\left(\frac{\eta_k}{B}\sum_{i=1}^{B}\left(F_0(Z_{0,i,k})-\mathbb{P}_{J_k}(F_0)(Z_{0,i,k})\right)\Psi_{J_k}\left(Z_{0,i,k}\right)+\frac{\eta_k}{B}\sum_{i=1}^{B}\varepsilon_{i,k}\Psi_{J_k}\left(Z_{0,i,k}\right)\right)^{\top}\\
&\times\ \eta_k\left(\zeta_{1,k}+\zeta_{2,k}+\zeta_{3,k}\right)\Bigg]\Bigg|\leq C\eta_k^2\left(J_{k-1}+J_{k-1}^3\|\Delta\check{\theta}_{k-1}\|^2\right).
\end{aligned}$$

Since $\eta_k\to 0$ according to our choice, then we can choose c sufficiently small such that for k large, $(1+2c\eta_k)(1-C\eta_k)\leq 1-C'\eta_k$, where $0<c<C/2$ is fixed and $C'>0$. Finally, we show that

$$\|\beta_{J_k,0}-R_{J_{k-1},J_k}^{\top}\beta_{J_{k-1},0}\|_2\leq C\mathbf{1}_k J_k^{-s},$$

where C does not depend on J_k. To show this, note that

$$\begin{aligned}&\left|\Psi_{J_k}(Z_0)^\top(\beta_{J_k,0} - R_{J_{k-1},J_k}^\top \beta_{J_{k-1},0})\right| \\ &\leq \left|\Psi_{J_k}(Z_0)^\top \beta_{J_k,0} - F_0(Z_0)\right| + \left|F_0(Z_0) - \Psi_{J_{k-1}}(Z_0)^\top \beta_{J_{k-1},0}\right| \leq C J_k^{-s}\end{aligned}$$

according to Condition 5 and the fact that $\sup_k J_k/J_{k-1} < \infty$. Moreover, using Condition 7,

$$\begin{aligned}\mathbb{E}|\Psi_{J_k}(Z_0)^\top(\beta_{J_k,0} - R_{J_{k-1},J_k}^\top \beta_{J_{k-1},0})|^2 &= (\beta_{J_k,0} - R_{J_{k-1},J_k}^\top \beta_{J_{k-1},0})^\top \Gamma_{J_k}(\beta_{J_k,0} - R_{J_{k-1},J_k}^\top \beta_{J_{k-1},0}) \\ &\geq C\|\beta_{J_k,0} - R_{J_{k-1},J_k}^\top \beta_{J_{k-1},0}\|_2^2.\end{aligned}$$

This shows the desired result.

So together we have that

$$\mathbb{E}_{k-1}\|\Delta\widehat{\beta}_k\|_2^2 \leq (1 - C\eta_k + C\mathbf{1}_k)\,\|\Delta\widehat{\beta}_{k-1}\|_2^2 + C\left(\eta_k J_k^{1-2s} + \eta_k J_k^3\|\Delta\check{\theta}_{k-1}\|_2^2 + \eta_k^2 J_k + \mathbf{1}_k J_k^{-2s}\right).$$

Now define

$$b(k) = k^{-\alpha_\eta-(2s-1)\alpha_\dagger} + k^{-\alpha_\eta+3\alpha_\dagger-2\alpha_\theta}\log^{2c_\theta}(k) + k^{-2\alpha_\eta+\alpha_\dagger}.$$

The logarithm in the middle term is required by Condition 8. The preceding inequality then becomes

$$\mathbb{E}_{k-1}\|\Delta\widehat{\beta}_k\|_2^2 \leq \left(1 - Ck^{-\alpha_\eta} + C\mathbf{1}_k\right)\|\Delta\widehat{\beta}_{k-1}\|_2^2 + C\left(b(k) + \mathbf{1}_k k^{-2s\alpha_\dagger}\right), \qquad \text{a.s}$$

Now using the above dynamics, we can analyze the behavior of $\Delta\widehat{\beta}_k$. Note that Theorem 1 of Robbins and Siegmund (1971) can not be directly applied here due to non-contraction of $\mathbb{E}_{k-1}\|\Delta\widehat{\beta}_k\|_2^2$ when sieve dimension changes. To deal with this issue, for any $m \geq J_0$, we define $\mathcal{J}(m) = \min\{k : J_k \geq m\}$, which is the starting period of sieve dimension m. Then we look at dynamics between $\mathbb{E}_{\mathcal{J}(m-1)-1}\|\Delta\widehat{\beta}_{\mathcal{J}(m)-1}\|_2^2$ and $\|\Delta\widehat{\beta}_{\mathcal{J}(m-1)-1}\|_2^2$. Obviously,

$$\begin{aligned}
\mathbb{E}_{\mathcal{J}(m-1)}\|\Delta\widehat{\beta}_{\mathcal{J}(m)-1}\|_2^2 &\le \left(1 - C(\mathcal{J}(m)-1)^{-\alpha_\eta}\right)\mathbb{E}_{\mathcal{J}(m-1)}\|\Delta\widehat{\beta}_{\mathcal{J}(m)-2}\|_2^2 + Cb(\mathcal{J}(m)-1) \qquad \text{a.s} \\
&\le \left(1 - C(\mathcal{J}(m)-1)^{-\alpha_\eta}\right)\left(1 - C(\mathcal{J}(m)-2)^{-\alpha_\eta}\right)\mathbb{E}_{\mathcal{J}(m-1)}\|\Delta\widehat{\beta}_{\mathcal{J}(m)-3}\|_2^2 \\
&+ C\left(b(\mathcal{J}(m)-1) + \left(1 - C(\mathcal{J}(m)-1)^{-\alpha_\eta}\right)b(\mathcal{J}(m)-2)\right), \qquad \text{a.s} \\
&\le \left[\prod_{j=\mathcal{J}(m-1)+1}^{\mathcal{J}(m)-1}(1 - Cj^{-\alpha_\eta})\right]\|\Delta\widehat{\beta}_{\mathcal{J}(m-1)}\|_2^2 \\
&\quad + C\left[\prod_{j=\mathcal{J}(m-1)+1}^{\mathcal{J}(m)-1}(1 - Cj^{-\alpha_\eta})\right]\sum_{j=\mathcal{J}(m-1)+1}^{\mathcal{J}(m)-1} b(j)\left[\prod_{l=\mathcal{J}(m-1)+1}^{j}(1 - Cl^{-\alpha_\eta})\right]^{-1}.
\end{aligned}$$

Using Lemma F.5, when $\frac{1}{2} < \alpha_\eta < 1$, we have that for each m,

$$\left[\prod_{j=\mathcal{J}(m-1)+1}^{\mathcal{J}(m)-1}(1 - Cj^{-\alpha_\eta})\right] \asymp \exp\left(-\frac{C}{1-\alpha_\eta}\left(\mathcal{J}(m)^{1-\alpha_\eta} - \mathcal{J}(m-1)^{1-\alpha_\eta}\right)\right).$$

We can also show that

$$\begin{aligned}
&\sum_{j=\mathcal{J}(m-1)+1}^{\mathcal{J}(m)-1} b(j)\left[\prod_{l=\mathcal{J}(m-1)+1}^{j}(1 - Cl^{-\alpha_\eta})\right]^{-1} \\
&\le C\,\mathcal{J}(m)^{\alpha_\eta}\, b(\mathcal{J}(m))\exp\left(\tfrac{C}{1-\alpha_\eta}\left(\mathcal{J}(m)^{1-\alpha_\eta} - \mathcal{J}(m-1)^{1-\alpha_\eta}\right)\right).
\end{aligned}$$

To see this, write $Q_j = \prod_{l=1}^{j}(1 - Cl^{-\alpha_\eta})$, so that the summand equals $b(j)Q_{\mathcal{J}(m-1)}\,Q_j^{-1}$. Extending the (positive) summation range and applying Lemma F.5(ii) and then Lemma F.5(i),

$$\begin{aligned}
\sum_{j=\mathcal{J}(m-1)+1}^{\mathcal{J}(m)-1} b(j)\,Q_{\mathcal{J}(m-1)}\,Q_j^{-1} &\le Q_{\mathcal{J}(m-1)}\sum_{j=1}^{\mathcal{J}(m)-1} b(j)\,Q_j^{-1} \\
&\le C\,Q_{\mathcal{J}(m-1)}\cdot\mathcal{J}(m)^{\alpha_\eta}\,b(\mathcal{J}(m))\exp\left(\tfrac{C}{1-\alpha_\eta}\mathcal{J}(m)^{1-\alpha_\eta}\right) \\
&\le C\,\mathcal{J}(m)^{\alpha_\eta}\,b(\mathcal{J}(m))\exp\left(\tfrac{C}{1-\alpha_\eta}\left(\mathcal{J}(m)^{1-\alpha_\eta} - \mathcal{J}(m-1)^{1-\alpha_\eta}\right)\right),
\end{aligned}$$

where the last step uses Lemma F.5(i) to bound $Q_{\mathcal{J}(m-1)} \le C\exp(-\frac{C}{1-\alpha_\eta}\mathcal{J}(m-1)^{1-\alpha_\eta})$. By the mean value theorem and $\mathcal{J}(m) \asymp m^{1/\alpha_\dagger}$, when $1 - \alpha_\eta - \alpha_\dagger > 0$, we have that

$$\mathcal{J}(m)^{1-\alpha_\eta} - \mathcal{J}(m-1)^{1-\alpha_\eta} \asymp \mathcal{J}(m)^{-\alpha_\eta}\left(\mathcal{J}(m) - \mathcal{J}(m-1)\right) \asymp m^{-\frac{\alpha_\eta}{\alpha_\dagger}}\cdot m^{\frac{1-\alpha_\dagger}{\alpha_\dagger}} = m^{\frac{1-\alpha_\eta-\alpha_\dagger}{\alpha_\dagger}},$$

which yields the upper bound. This leads to that

$$\begin{aligned}\mathbb{E}_{\mathcal{J}(m-1)}\|\Delta\widehat{\beta}_{\mathcal{J}(m)-1}\|_2^2 &\le C\exp\left(-C_\beta m^{\frac{1-\alpha_\eta-\alpha_\dagger}{\alpha_\dagger}}\right)\|\Delta\widehat{\beta}_{\mathcal{J}(m-1)}\|_2^2\\ &\quad + C\left(m^{1-2s} + m^{3-\frac{2\alpha_\theta}{\alpha_\dagger}}\log^{2c_\theta}(m) + m^{1-\frac{\alpha_\eta}{\alpha_\dagger}}\right), \quad \text{a.s.}\end{aligned}$$

and because the changing pseudo-true sieve coefficients only has impacts of magnitude m^{-2s} at sieve dimension m, which is effectively dominated by the m^{1-2s} term, we eventually have that

$$\begin{aligned}\mathbb{E}_{\mathcal{J}(m-1)-1}\|\Delta\widehat{\beta}_{\mathcal{J}(m)-1}\|_2^2 &\le C\exp\left(-C_\beta m^{\frac{1-\alpha_\eta-\alpha_\dagger}{\alpha_\dagger}}\right)\|\Delta\widehat{\beta}_{\mathcal{J}(m-1)-1}\|_2^2\\ &\quad + C\left(m^{1-2s} + m^{3-\frac{2\alpha_\theta}{\alpha_\dagger}}\log^{2c_\theta}(m) + m^{1-\frac{\alpha_\eta}{\alpha_\dagger}}\right), \qquad \text{a.s.}\end{aligned}$$

Define $\underline{\alpha}_F = \min\{2s-1, 2\alpha_\theta/\alpha_\dagger - 3, \alpha_\eta/\alpha_\dagger - 1\} > 0$, then

$$\mathbb{E}_{\mathcal{J}(m-1)-1}\|\Delta\widehat{\beta}_{\mathcal{J}(m)-1}\|_2^2 \le C\exp\left(-C_\beta m^{\frac{1-\alpha_\eta-\alpha_\dagger}{\alpha_\dagger}}\right)\|\Delta\widehat{\beta}_{\mathcal{J}(m-1)-1}\|_2^2 + Cm^{-\underline{\alpha}_F}\log^{2c_\theta}(m), \qquad \text{a.s.}$$

Fix $r > 2c_\theta + 1$. Then

$$\begin{aligned}&\mathbb{E}_{\mathcal{J}(m-1)-1}\left[m^{-1+\underline{\alpha}_F}\log^{-r}(m)\|\Delta\widehat{\beta}_{\mathcal{J}(m)-1}\|_2^2\right]\\ &\le \frac{Cm^{-1+\underline{\alpha}_F}\log^{-r}(m)}{(m-1)^{-1+\underline{\alpha}_F}\log^{-r}(m-1)}\exp\left(-C_\beta m^{\frac{1-\alpha_\eta-\alpha_\dagger}{\alpha_\dagger}}\right)\left[(m-1)^{-1+\underline{\alpha}_F}\log^{-r}(m-1)\|\Delta\widehat{\beta}_{\mathcal{J}(m-1)-1}\|_2^2\right]\\ &\quad + Cm^{-1}\log^{-r+2c_\theta}(m), \quad \text{a.s.}\\ &\le C\exp\left(-C_\beta m^{\frac{1-\alpha_\eta-\alpha_\dagger}{\alpha_\dagger}}\right)\left[(m-1)^{-1+\underline{\alpha}_F}\log^{-r}(m-1)\|\Delta\widehat{\beta}_{\mathcal{J}(m-1)-1}\|_2^2\right] + Cm^{-1}\log^{-r+2c_\theta}(m), \quad \text{a.s.}\end{aligned}$$

Since $r > 2c_\theta + 1$ makes $\sum_{m\ge 2} m^{-1}\log^{-r+2c_\theta}(m) < \infty$, Theorem 1 of Robbins and Siegmund (1971) shows that $m^{-1+\underline{\alpha}_F}\log^{-r}(m)\|\Delta\widehat{\beta}_{\mathcal{J}(m)-1}\|_2^2$ converges almost surely to a finite random variable. Thus

$$\|\Delta\widehat{\beta}_{\mathcal{J}(m)-1}\|_2^2 = O\left(m^{1-\underline{\alpha}_F}\log^r(m)\right), \qquad \text{a.s.}$$

Then we have that

$$\begin{aligned}\mathbb{E}_{k-1}\|\Delta\widehat{\beta}_k\|_2^2 &\le \left(1 - Ck^{-\alpha_\eta}\right)\|\Delta\widehat{\beta}_{k-1}\|_2^2 + +Ck^{-\alpha_\eta-\alpha_\dagger\underline{\alpha}_F}\log^{2c_\theta}(k)\\ &\quad + C\mathbf{1}_k k^{\alpha_\dagger(1-\underline{\alpha}_F)}\log^r(k) + C\mathbf{1}_k k^{-2\alpha_\dagger s}, \qquad \text{a.s}\end{aligned}$$

where recall that $r > 2c_\theta + 1$. Note that $\frac{k^{\alpha_\eta+\alpha_\dagger\underline{\alpha}_F-1}\log^{-r}(k)}{(k-1)^{\alpha_\eta+\alpha_\dagger\underline{\alpha}_F-1}\log^{-r}(k-1)} \le 1 + Ck^{-1}$. Then

$$
\begin{aligned}
&\mathbb{E}_{k-1}\left[k^{\alpha_\eta+\alpha_\dagger\underline{\alpha}_F-1}\log^{-r}(k)\|\Delta\widehat{\beta}_k\|_2^2\right] \\
&\le \left(1 - Ck^{-\alpha_\eta}\right)\left(1 + Ck^{-1}\right)\left[(k-1)^{\alpha_\eta+\alpha_\dagger\underline{\alpha}_F-1}\log^{-r}(k-1)\|\Delta\widehat{\beta}_{k-1}\|_2^2\right] \\
&+ Ck^{-1}\log^{-r+2c_\theta}(k) + C\mathbf{1}_k k^{\alpha_\eta+\alpha_\dagger-1} + C\mathbf{1}_k k^{-2\alpha_\dagger s+\alpha_\eta+\alpha_\dagger\underline{\alpha}_F-1}\log^{-r}(k), \qquad \text{a.s.}
\end{aligned}
$$

$\sum_{k=1}^\infty k^{-1}\log^{-r+2c_\theta}(k) < \infty$ for any $r > 1+2c_\theta$. Moreover, $\sum_k \mathbf{1}_k k^{\alpha_\eta+\alpha_\dagger-1} = \sum_m m^{\frac{\alpha_\eta+\alpha_\dagger-1}{\alpha_\dagger}} < \infty$ because $(\alpha_\eta+\alpha_\dagger-1) < -2\alpha_\dagger$. Also, since $\underline{\alpha}_F \le 2s-1$, $-2\alpha_\dagger s+\alpha_\eta+\alpha_\dagger\underline{\alpha}_F-1 \le \alpha_\eta-\alpha_\dagger-1$, so $\sum_k \mathbf{1}_k k^{-2\alpha_\dagger s+\alpha_\eta+\alpha_\dagger\underline{\alpha}_F-1}\log^{-r}(k) < \infty$. This, by the Theorem 1 of Robbins and Siegmund (1971), shows that $k^{\alpha_\eta+\alpha_\dagger\underline{\alpha}_F-1}\log^{-r}(k)\|\Delta\widehat{\beta}_k\|_2^2$ converges almost surely to a finite random variable and we prove that

$$
\|\Delta\widehat{\beta}_k\|_2 = O\left(k^{\frac{1-\alpha_\eta-\alpha_\dagger\underline{\alpha}_F}{2}}\log^{\frac{r}{2}}(k)\right), \qquad \text{a.s}
$$

C.4.3. Refine the Convergence Rate of $\|\Delta\widehat{\beta}_k\|_2$

Summary: This part shows that

$$
\|\Delta\widehat{\beta}_N\|_2 = O\left(N^{-\frac{\alpha_\eta-\alpha_\dagger}{2}}\log^{\frac{1}{2}}(N)\right), \qquad \text{a.s.}
$$

Recall that in the previous step, we have derived the following dynamics of $\Delta\widehat{\beta}_k$:

$$
\begin{aligned}
\Delta\widehat{\beta}_k &= \left(\mathbb{I}_{J_k} - \eta_k\check{\Gamma}_{J_k,k}\right)R_{J_{k-1},J_k}^\top\Delta\widehat{\beta}_{k-1} + \left(\mathbb{I}_{J_k} - \eta_k\check{\Gamma}_{J_k,k}\right)\left(R_{J_{k-1},J_k}^\top\beta_{J_{k-1},0} - \beta_{J_k,0}\right) \\
&+ \frac{\eta_k}{B}\sum_{i=1}^B \varepsilon_{i,k}\Psi_{J_k}\left(Z_{0,i,k}\right) + \frac{\eta_k}{B}\sum_{i=1}^B \left(F_0(Z_{0,i,k}) - \mathbb{P}_{J_k}(F_0)(Z_{0,i,k})\right)\Psi_{J_k}\left(Z_{0,i,k}\right) \\
&+ \eta_k\left[\underbrace{\frac{1}{B}\sum_{i=1}^B\left(F_0\left(Z_{0,i,k}\right) - F_0\left(\check{Z}_{i,k}\right)\right)\Psi_{J_k}\left(\check{Z}_{i,k}\right)}_{\zeta_{1,k}} + \underbrace{\frac{1}{B}\sum_{i=1}^B\varepsilon_{i,k}(\Psi_{J_k}\left(\check{Z}_{i,k}\right) - \Psi_{J_k}\left(Z_{0,i,k}\right))}_{\zeta_{2,k}}\right. \\
&\left.+\underbrace{\frac{1}{B}\sum_{i=1}^B\left(\left(F_0(\check{Z}_{i,k}) - \mathbb{P}_{J_k}(F_0)(\check{Z}_{i,k})\right)\Psi_{J_k}\left(\check{Z}_{i,k}\right) - \left(F_0(Z_{0,i,k}) - \mathbb{P}_{J_k}(F_0)(Z_{0,i,k})\right)\Psi_{J_k}\left(Z_{0,i,k}\right)\right)}_{\zeta_{3,k}}\right].
\end{aligned}
$$

For notational simplicity, further denote

$$A_k = \left(\mathbb{I}_{J_k} - \eta_k \Gamma_{J_k}\right) R^\top_{J_{k-1},J_k}$$

$$\zeta_{4,k} = \frac{1}{B}\sum_{i=1}^{B} \varepsilon_{i,k} \Psi_{J_k}\left(Z_{0,i,k}\right) + \frac{1}{B}\sum_{i=1}^{B} \left(F_0(Z_{0,i,k}) - \mathbb{P}_{J_k}(F_0)(Z_{0,i,k})\right) \Psi_{J_k}\left(Z_{0,i,k}\right),$$

$$\zeta_{5,k} = \left(\Gamma_{J_k} - \check{\Gamma}_{J_k,k}\right) R^\top_{J_{k-1},J_k} \Delta\widehat{\beta}_{k-1},$$

and

$$\zeta_{6,k} = \left(\mathbb{I}_{J_k} - \eta_k \check{\Gamma}_{J_k,k}\right) \left(R^\top_{J_{k-1},J_k} \beta_{J_{k-1},0} - \beta_{J_k,0}\right).$$

We then have that

$$\Delta\widehat{\beta}_k = A_k \Delta\widehat{\beta}_{k-1} + \eta_k \sum_{l=1}^{5} \zeta_{l,k} + \zeta_{6,k}.$$

This leads to

$$\Delta\widehat{\beta}_N = \left(\prod_{k=1}^{N} A_k\right) \Delta\widehat{\beta}_0 + \sum_{k=1}^{N} \left(\prod_{j=k+1}^{N} A_j\right) \left(\eta_k \sum_{l=1}^{5} \zeta_{l,k} + \zeta_{6,k}\right),$$

where $\prod_{k=1}^{N} A_k = A_N A_{N-1} \cdots A_1$, $\prod_{j=k+1}^{N} A_j = A_N A_{N-1} \cdots A_{k+1}$, and $\prod_{j=N+1}^{N} A_j = \mathbb{I}_{J_N}$. Now we analyze the a.s. order of the above terms one by one.

For the first term $\left(\prod_{j=1}^{N} A_j\right) \Delta\widehat{\beta}_0$, we have that $\|A_k\|_{\text{op}} \le 1 - C\eta_k$ when $J_k = J_{k-1}$ and $\|A_k\|_{\text{op}} \le C$ when $J_k > J_{k-1}$. So

$$\left\| \left(\prod_{k=1}^{N} A_k\right) \Delta\widehat{\beta}_0 \right\|_2 \le \exp\left(-C \sum_{k=1}^{N} \eta_k + C J_N\right) \|\Delta\widehat{\beta}_0\|_2 \le \exp\left(-C N^{1-\alpha_\eta}\right) \|\Delta\widehat{\beta}_0\|_2$$

when $1 - \alpha_\eta > \alpha_\dagger$, which decreases exponentially and is thus negligible. In the following we only need to focus on ζ-terms.

For $\sum_{k=1}^{N}\left(\prod_{j=k+1}^{N}A_j\right)\eta_k\zeta_{1,k}$, again using Lemma F.5, when $1-\alpha_\eta > \alpha_\dagger$, we have that

$$\begin{aligned}
&\left\|\sum_{k=1}^{N}\left(\prod_{j=k+1}^{N}A_j\right)\eta_k\zeta_{1,k}\right\|_2 \\
&\leq C\sum_{k=1}^{N}\exp\left(-C\sum_{j=k+1}^{N}\eta_j + C\left(J_N - J_k\right)\right)\eta_k J_k^{\frac{1}{2}}\frac{1}{B}\sum_{i=1}^{B}\|X_{i,k}\|_2\|\Delta\check{\theta}_{k-1}\|_2 \\
&\leq C\exp\left(-CN^{1-\alpha_\eta}\right)\sum_{k=1}^{N}\exp\left(Ck^{1-\alpha_\eta}\right)k^{\frac{\alpha_\dagger}{2}-\alpha_\eta-\alpha_\theta}\log^{c_\theta}(k)\frac{1}{B}\sum_{i=1}^{B}\|X_{i,k}\|_2, \qquad \text{a.s} \\
&= O\left(N^{\frac{\alpha_\dagger}{2}-\alpha_\theta}\log^{c_\theta}(N)\right) \qquad \text{a.s.}
\end{aligned}$$

The last line order is confirmed by Lemma C.1.

To show the rate of $\left\|\sum_{k=1}^{N}\left(\prod_{j=k+1}^{N}A_j\right)\eta_k\zeta_{2,k}\right\|_2$, we denote $Q_N = \sum_{k=1}^{N}\left(\prod_{j=k+1}^{N}A_j\right)\eta_k\zeta_{2,k}$. Then

$$\mathbb{E}_{N-1}\|Q_N\|_2^2 \leq \left(1 - CN^{-\alpha_\eta} + C\mathbf{1}_k\right)\|Q_{N-1}\|_2^2 + CN^{-2\alpha_\eta+3\alpha_\dagger-2\alpha_\theta}\log^{2c_\theta}(N), \qquad \text{a.s.}$$

Here the cross term vanishes conditionally because $\mathbb{E}_{k-1}\varepsilon_{i,k} = 0$. The innovation exponent is $2\alpha_\eta + 2\alpha_\theta - 3\alpha_\dagger$ because $\mathbb{E}\|X\|_2^2 < \infty$. Applying the proof of Section C.4.2, we easily have that for any $r > 1 + 2c_\theta$,

$$\left\|\sum_{k=1}^{N}\left(\prod_{j=k+1}^{N}A_j\right)\eta_k\zeta_{2,k}\right\|_2 = O\Big(N^{-\{\alpha_\theta+\alpha_\eta-(3\alpha_\dagger+1)/2\}}\log^{\frac{r}{2}}(N)\Big), \qquad \text{a.s.}$$

Following the proof of $\zeta_{1,k}$ term, we can also verify that

$$\left\|\sum_{k=1}^{N}\left(\prod_{j=k+1}^{N}A_j\right)\eta_k\zeta_{3,k}\right\|_2 \leq CN^{(\frac{3}{2}-s)\alpha_\dagger-\alpha_\theta}\log^{c_\theta}(N), \qquad \text{a.s}$$

We next control $\zeta_{5,k}$, postponing the discussion of $\zeta_{4,k}$ to the end. We show that this term does not determine the convergence rate of $\|\Delta\widehat{\beta}_k\|_2$. To show this, we will show that for any $\alpha_\beta > 0$, if $\|\Delta\widehat{\beta}_k\|_2 = O_{\text{a.s.}}(k^{-\alpha_\beta}\text{PolyLog})$, then

$$\left\|\sum_{k=1}^{N}\left(\prod_{j=k+1}^{N}A_j\right)\eta_k\zeta_{5,k}\right\|_2 = O\left(N^{-\alpha_\beta-\delta}\text{PolyLog}\right), \qquad \text{a.s.}$$

for some $\delta > 0$. To show this, first let

$$Q_N = \sum_{k=1}^{N} \left(\prod_{j=k+1}^{N} A_j \right) \eta_k \left(\Gamma_{J_k} - \widehat{\Gamma}_{J_k,k} \right) R_{J_{k-1},J_k}^{\top} \Delta\widehat{\beta}_{k-1} \mathbf{1} \left(\|\Delta\widehat{\beta}_{k-1}\|_2 \le k^{-\alpha_\beta} \text{PolyLog} \right).$$

Then

$$\text{Var}(Q_N) \le CN^{-\alpha_\eta + 2\alpha_\dagger - 2\alpha_\beta} \text{PolyLog}.$$

For N sufficiently large, the largest increment of the above partial sum is bounded by $CN^{-\alpha_\eta+\alpha_\dagger-\alpha_\beta}$PolyLog. By Freedman (1975); Tropp (2011), we have that

$$\|Q_N\|_2 = O\left(N^{\frac{-\alpha_\eta + 2\alpha_\dagger}{2} - \alpha_\beta} \text{PolyLog} + N^{-\alpha_\eta+\alpha_\dagger-\alpha_\beta} \text{PolyLog} \right), \qquad \text{a.s.}$$

This shows that $\|Q_N\|_2 = O_{a.s.}(N^{\frac{-\alpha_\eta+2\alpha_\dagger}{2} - \alpha_\beta}\text{PolyLog})$. And by appropriately choosing the poly-log terms, we can guarantee that the difference between Q_N and

$$\sum_{k=1}^{N} \left(\prod_{j=k+1}^{N} A_j \right) \eta_k \left(\Gamma_{J_k} - \widehat{\Gamma}_{J_k,k} \right) R_{J_{k-1},J_k}^{\top} \Delta\widehat{\beta}_{k-1}$$

decays almost surely exponentially quickly. This shows that

$$\left\| \sum_{k=1}^{N} \left(\prod_{j=k+1}^{N} A_j \right) \eta_k \left(\Gamma_{J_k} - \widehat{\Gamma}_{J_k,k} \right) R_{J_{k-1},J_k}^{\top} \Delta\widehat{\beta}_{k-1} \right\|_2 = O\left(N^{-\frac{\alpha_\eta - 2\alpha_\dagger}{2} - \alpha_\beta} \text{PolyLog} \right), \qquad \text{a.s.}$$

On the other side, let $Q_N = \sum_{k=1}^{N} \left(\prod_{j=k+1}^{N} A_j \right) \eta_k \left(\check{\Gamma}_{J_k,k} - \widehat{\Gamma}_{J_k,k} \right) R_{J_{k-1},J_k}^{\top} \Delta\widehat{\beta}_{k-1}$, it is direct to verify that

$$\|Q_N\|_2 \le (1 - CN^{-\alpha_\eta} + C\mathbf{1}_N)\|Q_{N-1}\|_2 + C\eta_N J_N^2 \|\Delta\check{\theta}_{N-1}\|_2 \|\Delta\widehat{\beta}_{N-1}\|_2 \sum_{i=1}^{B} \|X_{i,k}\|_2$$

By Lemma C.1, we have

$$\|Q_N\|_2 = O\left(N^{2\alpha_\dagger - \alpha_\theta - \alpha_\beta} \text{PolyLog} \right), \qquad \text{a.s.}$$

Define $\delta = \min\{\frac{\alpha_\eta - 2\alpha_\dagger}{2}, \alpha_\theta - 2\alpha_\dagger\} > 0$ under Condition 9, we obviously have that

$$\left\| \sum_{k=1}^{N} \left(\prod_{j=k+1}^{N} A_j \right) \eta_k \zeta_{5,k} \right\|_2 = O\left(N^{-\alpha_\beta - \delta} \text{PolyLog} \right), \qquad \text{a.s.}$$

This shows the result.

For $\sum_{k=1}^{N}\left(\prod_{j=k+1}^{N} A_j\right)\zeta_{6,k}$, we obviously have that

$$\left\|\sum_{k=1}^{N}\left(\prod_{j=k+1}^{N} A_j\right)\zeta_{6,k}\right\|_2 \leq C\exp\left(-CN^{1-\alpha_\eta}\right)\sum_{k=1}^{N}\exp\left(Ck^{1-\alpha_\eta}\right)\mathbf{1}_k J_k^{-s} \leq CN^{-s\alpha_\dagger},$$

which is effectively the bias of sieve approximation.

We finally look at $\zeta_{4,k}$. We first look at the order of the second term

$$\frac{\eta_k}{B}\sum_{i=1}^{B}\left(F_0(Z_{0,i,k}) - \mathbb{P}_{J_k}(F_0)(Z_{0,i,k})\right)\Psi_{J_k}\left(Z_{0,i,k}\right)$$

Denote the weighted sum as Q_N, then

$$\mathbb{E}_{N-1}\|Q_N\|_2^2 \leq (1-C\eta_N+C\mathbf{1}_N)\|Q_{N-1}\|_2^2+C\eta_N^2 J_N^{1-2s} \leq (1-CN^{-\alpha_\eta}+C\mathbf{1}_N)\|Q_{N-1}\|_2^2+CN^{-2\alpha_\eta+(1-2s)\alpha_\dagger}.$$

Here $2\alpha_\eta+(2s-1)\alpha_\dagger > 1$ under Condition 9. Thus

$$\|Q_N\|_2 = O\Big(N^{-\alpha_\eta-(s-\frac{1}{2})\alpha_\dagger+\frac{1}{2}}\mathrm{PolyLog}\Big), \qquad \text{a.s.}$$

It remains to analyze the order of the following sum

$$\sum_{k=1}^{N}\left(\prod_{j=k+1}^{N} A_j\right)\frac{\eta_k}{B}\sum_{i=1}^{B}\varepsilon_{i,k}\Psi_{J_k}(Z_{0,i,k}) = \sum_{k=1}^{N}\frac{1}{B}\sum_{i=1}^{B}\varepsilon_{i,k}\eta_k\left(\prod_{j=k+1}^{N} A_j\right)\Psi_{J_k}(Z_{0,i,k}).$$

which is of the main focus here in this proof section. We will adopt the proof idea following Chen and Christensen (2015). Note that in the following proofs, since B is a fixed batch size, we sometimes absorb it into constant without explicit notification.

First of all, it's direct to verify that when $1-\alpha_\eta > \alpha_\dagger$,

$$\left\|\eta_k\left(\prod_{j=k+1}^{N} A_j\right)\Psi_{J_k}(Z_{0,i,k})\right\|_2 \leq CJ_k^{\frac{1}{2}}\eta_k\exp\left(-C(N^{1-\alpha_\eta}-k^{1-\alpha_\eta})\right)$$

so

$$\max_{1\leq i\leq B}\max_{1\leq k\leq N}\left\|\eta_k\left(\prod_{j=k+1}^{N} A_j\right)\Psi_{J_k}(Z_{0,i,k})\right\|_2 \leq CN^{-\alpha_\eta+\frac{\alpha_\dagger}{2}}.$$

To analyze the behavior of $\sum_{k=1}^N \frac{1}{B}\sum_{i=1}^B \varepsilon_{i,k}\eta_k \left(\prod_{j=k+1}^N A_j\right)\Psi_{J_k}(Z_{0,i,k})$, let M_N be a positive number to be chosen later, and $\mathbf{1}_{i,k}^{M_N}$ as the indicator of whether $|\varepsilon_{i,k}| \leq M_N$ occurs and $\mathbf{1}_{i,k}^{M_N,c} = 1 - \mathbf{1}_{i,k}^{M_N}$. Following Chen and Christensen (2015), we define

$$\varepsilon_{1,N,i,k} = \varepsilon_{i,k}\mathbf{1}_{i,k}^{M_N} - \mathbb{E}(\varepsilon_{i,k}\mathbf{1}_{i,k}^{M_N}|Z_{0,i,k}), \quad \varepsilon_{2,N,i,k} = \varepsilon_{i,k} - \varepsilon_{1,N,i,k}. \tag{C.4}$$

We obviously have that $|\varepsilon_{1,N,i,k}| \leq 2M_N$. Also, we have that, under Condition 6 about the conditional variance of ε, there holds

$$\begin{aligned}
\mathbb{E}\left(\varepsilon_{1,N,i,k}^2\middle| Z_{0,i,k}\right) &= \mathbb{E}\left(\varepsilon_{i,k}^2\mathbf{1}_{i,k}^{M_N} + \left(\mathbb{E}(\varepsilon_{i,k}\mathbf{1}_{i,k}^{M_N}|Z_{0,i,k})\right)^2 - 2\varepsilon_{i,k}\mathbf{1}_{i,k}^{M_N}\mathbb{E}(\varepsilon_{i,k}\mathbf{1}_{i,k}^{M_N}|Z_{0,i,k})\middle| Z_{0,i,k}\right)\\
&\leq \mathbb{E}\left(\varepsilon_{i,k}^2\mathbf{1}_{i,k}^{M_N}\middle| Z_{0,i,k}\right) - \left(\mathbb{E}(\varepsilon_{i,k}\mathbf{1}_{i,k}^{M_N}|Z_{0,i,k})\right)^2 \leq \sup_z \mathbb{E}\left(\varepsilon^2\middle| Z_0 = z\right) < \infty,
\end{aligned}$$

and

$$\begin{aligned}
\mathbb{E}\left|\varepsilon_{2,N,i,k}\right| &\leq \mathbb{E}\left(\left|\varepsilon_{i,k}\right|\mathbf{1}_{i,k}^{M_N,c}\right) + \mathbb{E}\left|\mathbb{E}\left(\varepsilon_{i,k}\mathbf{1}_{i,k}^{M_N,c}\middle| Z_{0,i,k}\right)\right|\\
&\leq 2\mathbb{E}\left(|\varepsilon_{i,k}|\mathbf{1}_{i,k}^{M_N,c}\right) \leq 2\mathbb{E}(|\varepsilon_{i,k}|^\kappa\mathbf{1}_{i,k}^{M_N,c})M_N^{1-\kappa} \leq CM_N^{1-\kappa}.
\end{aligned}$$

Note that

$$\begin{aligned}
\sum_{k=1}^N \frac{1}{B}\sum_{i=1}^B \varepsilon_{i,k}\eta_k\left(\prod_{j=k+1}^N A_j\right)\Psi_{J_k}(Z_{0,i,k}) &= \sum_{k=1}^N \frac{1}{B}\sum_{i=1}^B \varepsilon_{1,N,i,k}\eta_k\left(\prod_{j=k+1}^N A_j\right)\Psi_{J_k}(Z_{0,i,k})\\
&+ \sum_{k=1}^N \frac{1}{B}\sum_{i=1}^B \varepsilon_{2,N,i,k}\eta_k\left(\prod_{j=k+1}^N A_j\right)\Psi_{J_k}(Z_{0,i,k}).
\end{aligned}$$

For any $\varrho > 0$, Markov's inequality together Lemma F.5 lead to

$$\begin{aligned}
&P\left(\left\|\sum_{k=1}^N \frac{1}{B}\sum_{i=1}^B \varepsilon_{2,N,i,k}\eta_k\left(\prod_{j=k+1}^N A_j\right)\Psi_{J_k}(Z_{0,i,k})\right\|_2 > \varrho\right)\\
&\leq \varrho^{-1}\sum_{k=1}^N \eta_k \exp\left(-C(N^{1-\alpha_\eta} - k^{1-\alpha_\eta})\right) J_k^{\frac{1}{2}}\mathbb{E}|\varepsilon_{2,N,i,k}| \leq C\varrho^{-1}N^{\frac{\alpha_\dagger}{2}}M_N^{1-\kappa}.
\end{aligned}$$

On the other side, noting that

$$
\begin{aligned}
&\sum_{k=1}^{N}\sum_{i=1}^{B}\mathbb{E}\left[\varepsilon_{1,N,i,k}^2\left\|\eta_k\left(\prod_{j=k+1}^{N}A_j\right)\Psi_{J_k}(Z_{0,i,k})\right\|_2^2\right]\\
&=\sum_{k=1}^{N}\sum_{i=1}^{B}\mathbb{E}\left[\mathbb{E}(\varepsilon_{1,N,i,k}^2|Z_{0,i,k})\left\|\eta_k\left(\prod_{j=k+1}^{N}A_j\right)\Psi_{J_k}(Z_{0,i,k})\right\|_2^2\right]\\
&\leq C\sum_{k=1}^{N}\sum_{i=1}^{B}J_k\eta_k^2\exp\left(-C(N^{1-\alpha_\eta}-k^{1-\alpha_\eta})\right)\leq CN^{-\alpha_\eta+\alpha_\dagger}.
\end{aligned}
$$

Then by using the results from Tropp (2012), we have that

$$
\begin{aligned}
P\left(\left\|\sum_{k=1}^{N}\frac{1}{B}\sum_{i=1}^{B}\varepsilon_{1,N,i,k}\eta_k\left(\prod_{j=k+1}^{N}A_j\right)\Psi_{J_k}(Z_{0,i,k})\right\|_2>\varrho\right)&\leq C\exp\left(C_1\log(N)-C_2\varrho^2N^{\alpha_\eta-\alpha_\dagger}\right)\\
&+C\exp\left(C_3\log(N)-C_4\varrho M_N^{-1}N^{\alpha_\eta-\frac{\alpha_\dagger}{2}}\right)
\end{aligned}
$$

As a result, if we choose $\varrho_N=\sqrt{(C_1+2)N^{\alpha_\dagger-\alpha_\eta}\log(N)/C_2}$, we have that

$$
\sum_{N=1}^{\infty}C\exp\left(C_1\log(N)-C_2\varrho_N^2N^{\alpha_\eta-\alpha_\dagger}\right)\leq C\sum_{N=1}^{\infty}N^{-2}<\infty.
$$

Furthermore, if we choose $M_N=\sqrt{C_4^2(C_1+2)N^{\alpha_\eta}/C_2(C_3+2)^2\log(N)}$, we have that

$$
\sum_{N=1}^{\infty}C\exp\left(C_3\log(N)-C_4\varrho_NM_N^{-1}N^{\alpha_\eta-\frac{\alpha_\dagger}{2}}\right)\leq C\sum_{N=1}^{\infty}N^{-2}<\infty,
$$

and finally when $\frac{\alpha_\eta}{2}-\frac{\alpha_\eta(\kappa-1)}{2}<-1$, or equivalently, $\kappa>2+2/\alpha_\eta$, which is imposed directly by Condition 6, we have that

$$
\sum_{N=1}^{\infty}\varrho_N^{-1}N^{\frac{\alpha_\dagger}{2}}M_N^{1-\kappa}\leq\sum_{N=1}^{\infty}\frac{CN^{\frac{\alpha_\eta}{2}}\log^{-\frac{1}{2}}(N)}{N^{\frac{\alpha_\eta(\kappa-1)}{2}}\log^{-\frac{\kappa-1}{2}}(N)}<\infty
$$

This shows that

$$
\left\|\sum_{k=1}^{N}\frac{1}{B}\sum_{i=1}^{B}\varepsilon_{i,k}\eta_k\left(\prod_{j=k+1}^{N}A_j\right)\Psi_{J_k}(Z_{0,i,k})\right\|_2=O\left(N^{-\frac{\alpha_\eta-\alpha_\dagger}{2}}\log^{\frac{1}{2}}(N)\right),\qquad\text{a.s.}
$$

As a result, when $\alpha_\eta<2\alpha_\theta, 2\alpha_\theta+\alpha_\eta>1+2\alpha_\dagger, \alpha_\eta+2s\alpha_\dagger>1, \alpha_\eta<(2s+1)\alpha_\dagger$ all hold as implied by Condition 9, we have that

$$\frac{\alpha_\eta - \alpha_\dagger}{2} < \min\{\alpha_\theta - \frac{\alpha_\dagger}{2}, \frac{2\alpha_\theta + 2\alpha_\eta - 3\alpha_\dagger - 1}{2}, \frac{2\alpha_\eta + (2s-1)\alpha_\dagger - 1}{2}, s\alpha_\dagger\}.$$

Then

$$\|\Delta\widehat{\beta}_N\|_2 = O\left(N^{-\frac{\alpha_\eta - \alpha_\dagger}{2}} \log^{\frac{1}{2}}(N)\right), \qquad \text{a.s.}$$

This shows the refined rate.

C.4.4. Obtain the Sharp Convergence Rate of PR Average Estimator

Note that the PR average of β is given by

$$\Delta\overline{\beta}_N = \frac{1}{N}\sum_{k=1}^N R_{J_k,J_N}^\top \Delta\widehat{\beta}_k + \frac{1}{N}\sum_{k=1}^N \left(R_{J_k,J_N}^\top \beta_{J_k,0} - \beta_{J_N,0}\right).$$

So

$$\Delta\widehat{\beta}_k = A_k \Delta\widehat{\beta}_{k-1} + \eta_k \sum_{l=1}^5 \zeta_{l,k} + \zeta_{6,k}.$$

Because $\alpha_\dagger < 1$, the integer sequence $J_k = [J_0 k^{\alpha_\dagger}]$ satisfies $J_k - J_{k-1} \in \{0,1\}$ for k large. Fix a deterministic $m_0 \geq J_0$ so large that such property holds from the block indexed by m_0 onward and every such constant-dimension block has length at least two (because $\mathcal{J}(m+1) - \mathcal{J}(m) \to \infty$). For N with $J_N \geq m_0$, define the exact remainder $\mathcal{R}_N^{(0)} := \frac{1}{N}\sum_{k<\mathcal{J}(m_0)} R_{J_k,J_N}^\top \Delta\widehat{\beta}_k$. The prefix contains a fixed number of terms and the nested re-embedding operators are uniformly bounded, so $\|\mathcal{R}_N^{(0)}\|_2 = O_{\text{a.s.}}(N^{-1})$.

For any $m_0 \leq m \leq J_N$, we have that $A_k = \mathbb{I}_m - \eta_k \Gamma_m$ and $\zeta_{6,k} = 0$ when $\mathcal{J}(m) + 1 \leq k \leq \mathcal{J}(m+1) - 1$ because the sieve dimension does not change across $k-1$ to k at these k. Then

$$\eta_k \Gamma_m \Delta\widehat{\beta}_{k-1} = \Delta\widehat{\beta}_{k-1} - \Delta\widehat{\beta}_k + \eta_k \sum_{l=1}^5 \zeta_{l,k}, \quad \mathcal{J}(m) + 1 \leq k \leq \mathcal{J}(m+1) - 1.$$

So

$$\Delta\widehat{\beta}_{k-1} = \Gamma_m^{-1}\eta_k^{-1}(\Delta\widehat{\beta}_{k-1} - \Delta\widehat{\beta}_k) + \Gamma_m^{-1}\sum_{l=1}^5 \zeta_{l,k}, \quad \mathcal{J}(m) + 1 \leq k \leq \mathcal{J}(m+1) - 1,$$

which, by Abel transformation, leads to

$$\sum_{k=\mathcal{J}(m)}^{\mathcal{J}(m+1)-2} \Delta\widehat{\beta}_k = \Gamma_m^{-1} \sum_{k=\mathcal{J}(m)+1}^{\mathcal{J}(m+1)-2} \Delta\widehat{\beta}_k \left(\eta_{k+1}^{-1} - \eta_k^{-1}\right) + \Gamma_m^{-1} \Delta\widehat{\beta}_{\mathcal{J}(m)} \eta_{\mathcal{J}(m)+1}^{-1}$$
$$- \Gamma_m^{-1} \Delta\widehat{\beta}_{\mathcal{J}(m+1)-1} \eta_{\mathcal{J}(m+1)-1}^{-1} + \Gamma_m^{-1} \sum_{k=\mathcal{J}(m)+1}^{\mathcal{J}(m+1)-1} \sum_{l=1}^{5} \zeta_{l,k}.$$

We then have the following exact expansion

$$\frac{1}{N} \sum_{k=1}^{N} R_{J_k,J_N}^\top \Delta\widehat{\beta}_k = \mathcal{R}_N^{(0)} + \frac{1}{N} \sum_{m=m_0}^{J_N-1} R_{m,J_N}^\top \sum_{k=\mathcal{J}(m)}^{\mathcal{J}(m+1)-2} \Delta\widehat{\beta}_k + \frac{1}{N} \sum_{k=\mathcal{J}(J_N)}^{N-1} \Delta\widehat{\beta}_k$$
$$+ \frac{1}{N} \sum_{m=m_0}^{J_N-1} R_{m,J_N}^\top \Delta\widehat{\beta}_{\mathcal{J}(m+1)-1} + \frac{1}{N} \Delta\widehat{\beta}_N.$$

For the prefix and the terms on the second line of the equality, the results from Section C.4.3 give

$$\begin{aligned}
&\left\| \mathcal{R}_N^{(0)} + \frac{1}{N} \sum_{m=m_0}^{J_N-1} R_{m,J_N}^\top \Delta\widehat{\beta}_{\mathcal{J}(m+1)-1} + \frac{1}{N} \Delta\widehat{\beta}_N \right\|_2 \\
&\le \frac{C(\omega)}{N} + \frac{C}{N} \sum_{m=m_0}^{J_N-1} \|\Delta\widehat{\beta}_{\mathcal{J}(m+1)-1}\|_2 + \frac{C}{N} \|\Delta\widehat{\beta}_N\|_2 \\
&\le \frac{C(\omega)}{N} + \frac{C}{N} \sum_{m=m_0}^{J_N} m^{-\frac{\alpha_\eta - \alpha_\dagger}{2\alpha_\dagger}} \log^{\frac{1}{2}}(m) + \frac{C}{N} J_N^{-\frac{\alpha_\eta - \alpha_\dagger}{2\alpha_\dagger}} \log^{\frac{1}{2}}(N) = o\left(N^{-\frac{1}{2}}\right), \qquad \text{a.s.}
\end{aligned}$$

when $1 + \alpha_\eta > 3\alpha_\dagger$. So we only need to look at

$$\frac{1}{N} \sum_{m=m_0}^{J_N-1} R_{m,J_N}^\top \sum_{k=\mathcal{J}(m)}^{\mathcal{J}(m+1)-2} \Delta\widehat{\beta}_k + \frac{1}{N} \sum_{k=\mathcal{J}(J_N)}^{N-1} \Delta\widehat{\beta}_k.$$

Define $\iota_N = \mathbf{1}\{\mathcal{J}(J_N) < N\}$. This indicator is needed because the terminal block is empty when N itself is an expansion checkpoint. Then

$$\begin{aligned}
&\frac{1}{N}\sum_{m=m_0}^{J_N-1} R_{m,J_N}^\top \sum_{k=\mathcal{J}(m)}^{\mathcal{J}(m+1)-2} \Delta\widehat{\beta}_k + \frac{1}{N}\sum_{k=\mathcal{J}(J_N)}^{N-1} \Delta\widehat{\beta}_k \\
&= \frac{1}{N}\sum_{m=m_0}^{J_N-1} R_{m,J_N}^\top \Gamma_m^{-1} \sum_{k=\mathcal{J}(m)+1}^{\mathcal{J}(m+1)-2} \Delta\widehat{\beta}_k \left(\eta_{k+1}^{-1} - \eta_k^{-1}\right) + \frac{\iota_N}{N}\Gamma_{J_N}^{-1} \sum_{k=a_N+1}^{N-1} \Delta\widehat{\beta}_k \left(\eta_{k+1}^{-1} - \eta_k^{-1}\right) \\
&+ \frac{1}{N}\sum_{m=m_0}^{J_N-1} R_{m,J_N}^\top \Gamma_m^{-1} \sum_{k=\mathcal{J}(m)+1}^{\mathcal{J}(m+1)-1} \sum_{l=1}^{5} \zeta_{l,k} + \frac{\iota_N}{N}\Gamma_{J_N}^{-1} \sum_{k=a_N+1}^{N} \sum_{l=1}^{5} \zeta_{l,k} \\
&+ \frac{1}{N}\sum_{m=m_0}^{J_N-1} R_{m,J_N}^\top \left(\Gamma_m^{-1}\Delta\widehat{\beta}_{\mathcal{J}(m)}\eta_{\mathcal{J}(m)+1}^{-1} - \Gamma_m^{-1}\Delta\widehat{\beta}_{\mathcal{J}(m+1)-1}\eta_{\mathcal{J}(m+1)-1}^{-1}\right) \\
&+ \frac{\iota_N}{N}\left(\Gamma_{J_N}^{-1}\Delta\widehat{\beta}_{\mathcal{J}(J_N)}\eta_{\mathcal{J}(J_N)+1}^{-1} - \Gamma_{J_N}^{-1}\Delta\widehat{\beta}_N\eta_N^{-1}\right).
\end{aligned}$$

Since in Section C.4.3 we have shown that $\|\Delta\widehat{\beta}_N\|_2 = O_{\text{a.s.}}(N^{-\frac{\alpha_\eta-\alpha_\dagger}{2}}\log^{\frac{1}{2}}(N))$, for N sufficiently large, the first line on RHS is bounded by

$$\frac{C}{N}\sum_{k=1}^{N} k^{\alpha_\eta-1}\|\Delta\widehat{\beta}_k\|_2 \le \frac{C}{N}\sum_{k=1}^{N} k^{\alpha_\eta-1-\frac{\alpha_\eta-\alpha_\dagger}{2}}\log^{\frac{1}{2}}(k)) \le CN^{\frac{\alpha_\eta+\alpha_\dagger-2}{2}}\log^{\frac{1}{2}}(N)), \qquad \text{a.s.}$$

The last two lines are together bounded by

$$\frac{C}{N}\sum_{m=m_0}^{J_N} m^{\frac{\alpha_\eta}{\alpha_\dagger}-\frac{\alpha_\eta-\alpha_\dagger}{2\alpha_\dagger}}\log^{\frac{1}{2}}(m) + \frac{C}{N}N^{\alpha_\eta-\frac{\alpha_\eta-\alpha_\dagger}{2}}\log^{\frac{1}{2}}(N) \le CN^{\frac{\alpha_\eta+3\alpha_\dagger-2}{2}}\log^{\frac{1}{2}}(N), \qquad \text{a.s.}$$

When $\alpha_\eta + 3\alpha_\dagger < 1$, the above two terms are both of order $o(N^{-1/2})$. So we only need to look at the summations of ζ-terms. Recall from the proof of Section C.4.2 and Section C.4.3, we have $\|\zeta_{1,k}\|_2 \le CJ_k^{\frac{1}{2}}\sum_{i=1}^B \|X_{i,k}\|_2\|\Delta\check{\theta}_{k-1}\|_2$, $\|\zeta_{2,k}\|_2 \le CJ_k^{\frac{3}{2}}\sum_{i=1}^B |\varepsilon_{i,k}|\|X_{i,k}\|_2\|\Delta\check{\theta}_{k-1}\|_2$, $\|\zeta_{3,k}\|_2 \le CJ_k^{\frac{3}{2}-s}\sum_{i=1}^B \|X_{i,k}\|_2\|\Delta\check{\theta}_{k-1}\|_2$, $\|\zeta_{4,k}\|_2 \le CJ_k^{\frac{1}{2}}(\sum_{i=1}^B |\varepsilon_{i,k}| + J_k^{-s})$, and $\|\zeta_{5,k}\|_2 \le$

$CJ_k\|\Delta\widehat{\beta}_{k-1}\|_2$. When $s>1$, applying Lemma C.1

$$
\begin{aligned}
&\left\|\frac{1}{N}\sum_{m=m_0}^{J_N-1} R_{m,J_N}^\top \Gamma_m^{-1}\sum_{k=\mathcal{J}(m)+1}^{\mathcal{J}(m+1)-1}\sum_{l=1}^{5}\zeta_{l,k} + \frac{1}{N}\Gamma_{J_N}^{-1}\sum_{k=\mathcal{J}(J_N)+1}^{N}\sum_{l=1}^{5}\zeta_{l,k} - \frac{1}{N}\sum_{k=1}^{N}R_{J_k,J_N}^\top\Gamma_{J_k}^{-1}\sum_{l=1}^{5}\zeta_{l,k}\right\|_2 \\
&\leq \frac{C(\omega)}{N} + \frac{C}{N}\sum_{m=m_0}^{J_N}\sum_{l=1}^{5}\|\zeta_{l,\mathcal{J}(m)}\|_2 \leq \frac{C(\omega)}{N} + \frac{C}{N}\sum_{m=m_0}^{J_N}\left(m^{\frac{1}{2}}\sum_{i=1}^{B}\|X_{i,\mathcal{J}(m)}\|_2\|\Delta\check{\theta}_{\mathcal{J}(m)-1}\|_2\right) \\
&+ \frac{C}{N}\sum_{m=J_0}^{J_N}\left(m^{\frac{3}{2}}\sum_{i=1}^{B}\|X_{i,\mathcal{J}(m)}\|_2|\varepsilon_{i,\mathcal{J}(m)}|\|\Delta\check{\theta}_{\mathcal{J}(m)-1}\|_2\right) + \frac{C}{N}\sum_{m=J_0}^{J_N}\left(m^{\frac{1}{2}}\sum_{i=1}^{B}|\varepsilon_{i,\mathcal{J}(m)}| + m^{\frac{1}{2}-s}\right) \\
&+ \frac{C}{N}\sum_{m=J_0}^{J_N} m\|\Delta\widehat{\beta}_{\mathcal{J}(m)-1}\|_2 = O\left(N^{\frac{5\alpha_\dagger}{2}-\alpha_\theta-1} + N^{\frac{3\alpha_\dagger}{2}-1} + N^{\frac{5\alpha_\dagger}{2}-\frac{\alpha_\eta}{2}-1}\right), \qquad \text{a.s.}
\end{aligned}
$$

up to some poly-log terms. When $1+2\alpha_\theta > 5\alpha_\dagger$, $3\alpha_\dagger < 1$, and $1+\alpha_\eta > 5\alpha_\dagger$, we have that the above terms are all $o(N^{-\frac{1}{2}})$ a.s. Then we only need to look at $\frac{1}{N}\sum_{k=1}^{N} R_{J_k,J_N}^\top\Gamma_{J_k}^{-1}\sum_{l=1}^{5}\zeta_{l,k}$. We look at each $\zeta_{l,k}$ separately.

For $l=1$, put $d_{i,k} = Z_{0,i,k} - \check{Z}_{i,k} = -X_{i,k}^\top\Delta\check{\theta}_{k-1}$. Taylor's theorem, applied once to F_0 and once to Ψ_{J_k}, gives the expansion

$$
\big(F_0(Z_{0,i,k}) - F_0(\check{Z}_{i,k})\big)\Psi_{J_k}(\check{Z}_{i,k}) = \nabla_z F_0(Z_{0,i,k})d_{i,k}\Psi_{J_k}(Z_{0,i,k}) + \mathcal{R}_{i,k}, \qquad \|\mathcal{R}_{i,k}\|_2 \leq CJ_k^{3/2}d_{i,k}^2.
$$

The remainder bound follows from the bounded first two derivatives of F_0 and the sieve derivative envelopes. Hence

$$
\begin{aligned}
\frac{1}{N}\sum_{k=1}^{N}R_{J_k,J_N}^\top\Gamma_{J_k}^{-1}\zeta_{1,k} &= \frac{1}{N}\sum_{k=1}^{N}R_{J_k,J_N}^\top\Gamma_{J_k}^{-1}\frac{1}{B}\sum_{i=1}^{B}\left(F_0\left(Z_{0,i,k}\right) - F_0\left(\check{Z}_{i,k}\right)\right)\Psi_{J_k}\left(\check{Z}_{i,k}\right) \\
&= -\frac{1}{N}\sum_{k=1}^{N}R_{J_k,J_N}^\top\Gamma_{J_k}^{-1}\frac{1}{B}\sum_{i=1}^{B}\left(\nabla_z F_0\left(Z_{0,i,k}\right)\right)\Psi_{J_k}(Z_{0,i,k})X_{i,k}^\top\Delta\check{\theta}_{k-1} \\
&+ O\left(\frac{1}{N}\sum_{k=1}^{N}\sum_{i=1}^{B}J_k^{\frac{3}{2}}\|X_{i,k}\|_2^2\|\Delta\check{\theta}_{k-1}\|_2^2\right).
\end{aligned}
$$

Let $T_k = B^{-1}\sum_{i=1}^{B}\|X_{i,k}\|_2^2$; then $\mathbb{E}_{k-1}T_k^2 \leq C$. The exact power-sum bound and Lemma C.1 give

$$
\frac{1}{N}\sum_{k=1}^{N}J_k^{3/2}T_k\|\Delta\check{\theta}_{k-1}\|_2^2 = O\left(N^{3\alpha_\dagger/2-2\alpha_\theta}\mathrm{PolyLog}\right), \qquad \text{a.s.}
$$

The strict inequality $\alpha_\theta > 1/4 + 3\alpha_\dagger/4$ makes the above $o_{\text{a.s.}}(N^{-1/2})$. Further more, define

$Q_k = \frac{1}{B}\sum_{i=1}^{B}\left(\nabla_z F_0\left(Z_{0,i,k}\right)\right)\Psi_{J_k}(Z_{0,i,k})X_{i,k}^\top - \mathbb{E}[(\nabla_z F_0\left(Z_0\right))\Psi_{J_k}(Z_0)X^\top]$, we have that

$$\begin{aligned}
&\frac{1}{N}\sum_{k=1}^{N} R_{J_k,J_N}^\top \Gamma_{J_k}^{-1}\frac{1}{B}\sum_{i=1}^{B}\left(\nabla_z F_0\left(Z_{0,i,k}\right)\right)\Psi_{J_k}(Z_{0,i,k})X_{i,k}^\top \Delta\check{\theta}_{k-1}\\
&= \frac{1}{N}\sum_{k=1}^{N} R_{J_k,J_N}^\top \Gamma_{J_k}^{-1}\mathbb{E}[(\nabla_z F_0\left(Z_0\right))\Psi_{J_k}(Z_0)X^\top]\Delta\check{\theta}_{k-1} + \frac{1}{N}\sum_{k=1}^{N} R_{J_k,J_N}^\top \Gamma_{J_k}^{-1} Q_k \Delta\check{\theta}_{k-1}
\end{aligned}$$

Let $S_N = \sum_{k=1}^{N} R_{J_k,J_N}^\top \Gamma_{J_k}^{-1} Q_k \Delta\check{\theta}_{k-1}$, we have that $S_N = R_{J_{N-1},J_N}^\top S_{N-1} + \Gamma_{J_N}^{-1} Q_N \Delta\check{\theta}_{N-1}$. For each N, define

$$\|S_N\|_{\Gamma_{J_N}} = \left(S_N^\top \Gamma_{J_N} S_N\right)^{\frac{1}{2}},$$

We have that

$$\begin{aligned}
\|S_N\|_{\Gamma_{J_N}}^2 &= S_{N-1}^\top R_{J_{N-1},J_N}\Gamma_{J_N} R_{J_{N-1},J_N}^\top S_{N-1} + \Delta\check{\theta}_{N-1}^\top Q_N^\top \Gamma_{J_N}^{-1} Q_N \Delta\check{\theta}_{N-1}\\
&\quad + 2S_{N-1}^\top R_{J_{N-1},J_N} Q_N \Delta\check{\theta}_{N-1}.
\end{aligned}$$

Since

$$\begin{aligned}
R_{J_{N-1},J_N}\Gamma_{J_N} R_{J_{N-1},J_N}^\top &= \mathbb{E}\left(R_{J_{N-1},J_N}\Psi_{J_N}(Z_0)\Psi_{J_N}(Z_0)^\top R_{J_{N-1},J_N}^\top\right)\\
&= \mathbb{E}\left(\Psi_{J_{N-1}}(Z_0)\Psi_{J_{N-1}}(Z_0)^\top\right) = \Gamma_{J_{N-1}},
\end{aligned}$$

we have that

$$\|S_N\|_{\Gamma_{J_N}}^2 = \|S_{N-1}\|_{\Gamma_{J_{N-1}}}^2 + \Delta\check{\theta}_{N-1}^\top Q_N^\top \Gamma_{J_N}^{-1} Q_N \Delta\check{\theta}_{N-1} + 2S_{N-1}^\top R_{J_{N-1},J_N} Q_N \Delta\check{\theta}_{N-1}.$$

So

$$\mathbb{E}_{N-1}\|S_N\|_{\Gamma_{J_N}}^2 \le \|S_{N-1}\|_{\Gamma_{J_{N-1}}}^2 + CJ_N\|\Delta\check{\theta}_{N-1}\|_2^2 \le \|S_{N-1}\|_{\Gamma_{J_{N-1}}}^2 + CN^{\alpha_\dagger - 2\alpha_\theta}\log^{2c_\theta}(N), \quad \text{a.s.}$$

This immediately leads to $\|S_N\|_{\Gamma_{J_N}}^2 = O_{\text{a.s.}}\left(N^{(1+\alpha_\dagger-2\alpha_\theta)_+}\text{PolyLog}\right)$. Since Γ_J has uniformly bounded eigenvalues, we have that

$$\left\|\frac{1}{N}\sum_{k=1}^{N} R_{J_k,J_N}^\top \Gamma_{J_k}^{-1} Q_k \Delta\check{\theta}_{k-1}\right\|_2 = \frac{1}{N}\|S_N\|_2 \le \frac{C}{N}\|S_N\|_{\Gamma_{J_N}} = O\left(N^{\frac{(1+\alpha_\dagger-2\alpha_\theta)_+}{2}-1}\text{PolyLog}\right), \qquad \text{a.s.}$$

which is $o(N^{-\frac{1}{2}})$ a.s. when $\alpha_\theta > \frac{1}{2}\alpha_\dagger$. The above analysis together shows that

$$\left\| \frac{1}{N}\sum_{k=1}^N R_{J_k,J_N}^\top \Gamma_{J_k}^{-1}\zeta_{1,k} + \frac{1}{N}\sum_{k=1}^N R_{J_k,J_N}^\top \Gamma_{J_k}^{-1}\mathbb{E}[(\nabla_z F_0(Z_0))\,\Psi_{J_k}(Z_0)X^\top]\Delta\check{\theta}_{k-1} \right\|_2 = o\left(N^{-\frac{1}{2}}\right), \qquad \text{a.s}$$

Using the same strategy, for $l=2$, let $S_N = \sum_{k=1}^N R_{J_k,J_N}^\top \Gamma_{J_k}^{-1}\zeta_{2,k}$, we have that

$$\mathbb{E}_{N-1}\|S_N\|_{\Gamma_{J_N}}^2 \le \|S_{N-1}\|_{\Gamma_{J_{N-1}}}^2 + CJ_N^3\|\Delta\check{\theta}_{N-1}\|_2^2 \le \|S_{N-1}\|_{\Gamma_{J_{N-1}}}^2 + CN^{3\alpha_\dagger - 2\alpha_\theta}\log^{2c_\theta}(N), \qquad \text{a.s}$$

So $\|S_N\|_{\Gamma_{J_N}}^2 = O_{\text{a.s.}}\left(N^{(1+3\alpha_\dagger-2\alpha_\theta)_+}\text{PolyLog}\right)$. Then

$$\left\| \frac{1}{N}\sum_{k=1}^N R_{J_k,J_N}^\top \Gamma_{J_k}^{-1}\zeta_{2,k} \right\|_2 = \frac{1}{N}\|S_N\|_2 \le \frac{C}{N}\|S_N\|_{\Gamma_{J_N}} = O\left(N^{\frac{(1+3\alpha_\dagger-2\alpha_\theta)_+}{2}-1}\text{PolyLog}\right), \qquad \text{a.s}$$

When $\alpha_\theta > \frac{3}{2}\alpha_\dagger$, the above term is of order $o_{\text{a.s.}}(N^{-\frac{1}{2}})$.

For $l=3$, using Lemma C.1, we have that

$$\begin{aligned}\left\| \frac{1}{N}\sum_{k=1}^N R_{J_k,J_N}^\top \Gamma_{J_k}^{-1}\zeta_{3,k} \right\|_2 &\le \frac{C}{N}\sum_{k=1}^N \|\zeta_{3,k}\|_2 \le \frac{C}{N}\sum_{k=1}^N J_k^{\frac{3}{2}-s}\sum_{i=1}^B \|X_{i,k}\|_2\|\Delta\check{\theta}_{k-1}\|_2 \\ &= O\left(N^{\left(1+\left(\frac{3}{2}-s\right)\alpha_\dagger-\alpha_\theta\right)_+-1}\text{PolyLog}\right), \qquad \text{a.s.}\end{aligned}$$

so when $\alpha_\theta + (s-\frac{3}{2})\alpha_\dagger > \frac{1}{2}$, we have that the above term is of order $o_{\text{a.s.}}(N^{-\frac{1}{2}})$.

For $l=4$, we look at the second term $\frac{1}{B}\sum_{i=1}^B \left(F_0(Z_{0,i,k}) - \mathbb{P}_{J_k}(F_0)(Z_{0,i,k})\right)\Psi_{J_k}\left(Z_{0,i,k}\right)$ first. Noting that

$$\mathbb{E}_{k-1}\sum_{i=1}^B \left(F_0(Z_{0,i,k}) - \mathbb{P}_{J_k}(F_0)(Z_{0,i,k})\right)\Psi_{J_k}\left(Z_{0,i,k}\right) = 0$$

and

$$\mathbb{E}_{k-1}\left\|\sum_{i=1}^B \left(F_0(Z_{0,i,k}) - \mathbb{P}_{J_k}(F_0)(Z_{0,i,k})\right)\Psi_{J_k}\left(Z_{0,i,k}\right)\right\|_2^2 \le J_k^{1-2s},$$

then by Lemma C.1 we have that

$$\begin{aligned}&\left\| \frac{1}{N}\sum_{k=1}^N R_{J_k,J_N}^\top \Gamma_{J_k}^{-1}\left(\frac{1}{B}\sum_{i=1}^B \left(F_0(Z_{0,i,k}) - \mathbb{P}_{J_k}(F_0)(Z_{0,i,k})\right)\Psi_{J_k}\left(Z_{0,i,k}\right)\right)\right\|_2 \\ &= O\left(N^{\frac{(1-(2s-1)\alpha_\dagger)_+}{2}-1}\text{PolyLog}\right), \qquad \text{a.s}\end{aligned}$$

When $2s > 1$, the above term is of order $o_{\text{a.s.}}(N^{-\frac{1}{2}})$. Then

$$\left\| \frac{1}{N} \sum_{k=1}^{N} R_{J_k,J_N}^{\top} \Gamma_{J_k}^{-1} \zeta_{4,k} - \frac{1}{N} \sum_{k=1}^{N} R_{J_k,J_N}^{\top} \Gamma_{J_k}^{-1} \frac{1}{B} \sum_{i=1}^{B} \varepsilon_{i,k} \Psi_{J_k} \left(Z_{0,i,k} \right) \right\|_2 = o\left(N^{-\frac{1}{2}} \right), \qquad \text{a.s.}$$

We next look at $l = 5$. Note that

$$\zeta_{5,k} = \left(\Gamma_{J_k} - \widehat{\Gamma}_{J_k,k} \right) R_{J_{k-1},J_k}^{\top} \Delta\widehat{\beta}_{k-1} + \left(\widehat{\Gamma}_{J_k,k} - \check{\Gamma}_{J_k,k} \right) R_{J_{k-1},J_k}^{\top} \Delta\widehat{\beta}_{k-1},$$

so we let

$$S_{1,N} = \sum_{k=1}^{N} R_{J_k,J_N}^{\top} \Gamma_{J_k}^{-1} \left(\Gamma_{J_k} - \widehat{\Gamma}_{J_k,k} \right) R_{J_{k-1},J_k}^{\top} \Delta\widehat{\beta}_{k-1},$$

and

$$S_{2,N} = \sum_{k=1}^{N} R_{J_k,J_N}^{\top} \Gamma_{J_k}^{-1} \left(\widehat{\Gamma}_{J_k,k} - \check{\Gamma}_{J_k,k} \right) R_{J_{k-1},J_k}^{\top} \Delta\widehat{\beta}_{k-1}.$$

Since

$$\mathbb{E}_{N-1} \|S_{1,N}\|_{\Gamma_{J_N}}^2 \leq \|S_{1,N-1}\|_{\Gamma_{J_{N-1}}}^2 + C J_N^2 \|\Delta\widehat{\beta}_{N-1}\|_2^2 \leq \|S_{1,N-1}\|_{\Gamma_{J_{N-1}}}^2 + C N^{3\alpha_{\dagger} - \alpha_{\eta}} \mathfrak{L}_N^2, \qquad \text{a.s.}$$

using the proof method before, we can show that

$$\frac{1}{N} \|S_{1,N}\|_2 = O\left(N^{\frac{(1+3\alpha_{\dagger}-\alpha_{\eta})_+}{2} - 1} \text{PolyLog} \right), \qquad \text{a.s.}$$

when $\alpha_\eta > 3\alpha_\dagger$, we have that $\frac{1}{N}\|S_{1,N}\|_2$ is $o_{\text{a.s.}}(N^{-\frac{1}{2}})$. For $S_{2,N}$, the mean-value theorem and the basis derivative bound give

$$\|\widehat{\Gamma}_{J_k,k} - \check{\Gamma}_{J_k,k}\|_{\text{op}} \leq C J_k^2 T_k \|\Delta\check{\theta}_{k-1}\|_2, \qquad T_k = \frac{1}{B} \sum_{i=1}^{B} (1 + \|X_{i,k}\|_2),$$

where $\mathbb{E}_{k-1} T_k^2 \leq C$. Therefore

$$\|S_{2,N}\|_2 \leq C \sum_{k=1}^{N} k^{2\alpha_{\dagger} - \alpha_{\theta} - \frac{\alpha_{\eta} - \alpha_{\dagger}}{2}} T_k \text{PolyLog}, \qquad \text{a.s.}$$

The second part of Lemma C.1 yields

$$\frac{1}{N} \|S_{2,N}\|_2 = O\left(N^{2\alpha_{\dagger} - \alpha_{\theta} - \frac{\alpha_{\eta} - \alpha_{\dagger}}{2}} \text{PolyLog} \right), \qquad \text{a.s.}$$

The strict condition $2\alpha_\theta + \alpha_\eta > 1 + 5\alpha_\dagger$ leads to that the above is $o_{\mathrm{a.s.}}(N^{-1/2})$; the same conclusion holds for $N^{-1}\sum_{k\le N} R_{J_k,J_N}^\top \Gamma_{J_k}^{-1}\zeta_{5,k}$.

Together, we have shown that

$$\begin{aligned}\Delta\overline{\beta}_N &= \frac{1}{N}\sum_{k=1}^N R_{J_k,J_N}^\top \Gamma_{J_k}^{-1}\frac{1}{B}\sum_{i=1}^B \varepsilon_{i,k}\Psi_{J_k}(Z_{0,i,k}) - \frac{1}{N}\sum_{k=1}^N R_{J_k,J_N}^\top \Gamma_{J_k}^{-1}\mathbb{E}[(\nabla_z F_0(Z_0))\Psi_{J_k}(Z_0)X^\top]\Delta\check{\theta}_{k-1} \\ &+ \frac{1}{N}\sum_{k=1}^N \big(R_{J_k,J_N}^\top \beta_{J_k,0} - \beta_{J_N,0}\big) + r_N, \qquad \|r_N\|_2 = o_{\mathrm{a.s.}}(N^{-1/2}).\end{aligned}$$

where the second last term describes the impacts of changing pseudo true coefficients. This proves the first result.

To prove the second result, recall that we define $\overline{F}_N(z) = \Psi_{J_N}^\top(z)\overline{\beta}_N$. We immediately have that

$$\begin{aligned}\overline{F}_N(z) - F_0(z) &= (\Psi_{J_N}(z)^\top\overline{\beta}_N - \Psi_{J_N}(z)^\top\beta_{J_N,0}) + (\mathbb{P}_{J_N}(F_0)(z) - F_0(z)) \\ &= (\mathbb{P}_{J_N}(F_0)(z) - F_0(z)) + \frac{1}{N}\sum_{k=1}^N ((\mathbb{P}_{J_k}(F_0)(z) - \mathbb{P}_{J_N}(F_0)(z))) \\ &+ \frac{1}{N}\sum_{k=1}^N \Psi_{J_k}(z)^\top\Gamma_{J_k}^{-1}\frac{1}{B}\sum_{i=1}^B \varepsilon_{i,k}\Psi_{J_k}(Z_{0,i,k}) \\ &- \frac{1}{N}\sum_{k=1}^N \Psi_{J_k}(z)^\top\Gamma_{J_k}^{-1}\mathbb{E}[(\nabla_z F_0(Z_0))\Psi_{J_k}(Z_0)X^\top]\Delta\check{\theta}_{k-1} + r_N^F(z),\end{aligned}$$

where $r_N^F(z) := \Psi_{J_N}(z)^\top r_N$. Hence, by $\sup_z \|\Psi_{J_N}(z)\|_2 \lesssim J_N^{1/2}$,

$$\|r_N^F(\cdot)\|_\infty = o\big(N^{-(1-\alpha_\dagger)/2}\big), \qquad \text{a.s.}$$

The two bias terms combine exactly as

$$\mathbb{P}_{J_N}F_0 - F_0 + \frac{1}{N}\sum_{k=1}^N(\mathbb{P}_{J_k}F_0 - \mathbb{P}_{J_N}F_0) = \frac{1}{N}\sum_{k=1}^N(\mathbb{P}_{J_k}F_0 - F_0).$$

Consequently their uniform norm is at most $CN^{-1}\sum_{k=1}^N J_k^{-s}$. Recall that $\mu_0(\theta_0, z) = \mathbb{E}(X|Z_0 = z)$, and put $h(z) = (\nabla_z F_0(z))\mu_0(\theta_0, z) \in \mathbb{R}^p$ and $E_J(z) = \mathbb{P}_J h(z) - h(z)$. Then

$$\Psi_J(z)^\top\Gamma_J^{-1}\mathbb{E}\left(\Psi_J(Z_0)\nabla_z F_0(Z_0)X^\top\right) = \{(\mathbb{P}_J h)(z)\}^\top,$$

Since $\|E_J\|_\infty \le CJ^{-s_\mu}$ under Condition 6, the generated-regressor contribution in the preceding expansion, including its sign, is exactly

$$-h(z)^\top \frac{1}{N}\sum_{k=1}^N \Delta\check{\theta}_{k-1} - \frac{1}{N}\sum_{k=1}^N E_{J_k}(z)^\top \Delta\check{\theta}_{k-1}.$$

Uniformly in z, its norm is bounded by

$$C\left\|\frac{1}{N}\sum_{k=1}^N \Delta\check{\theta}_{k-1}\right\|_2 + \frac{C}{N}\sum_{k=1}^N J_k^{-s_\mu}\|\Delta\check{\theta}_{k-1}\|_2 = C\left(\left\|\frac{1}{N}\sum_{k=1}^N \Delta\check{\theta}_{k-1}\right\|_2 + A_{\mu,N}\right).$$

Finally, to study the behavior of $\frac{1}{N}\sum_{k=1}^N \Psi_{J_k}(z)^\top \Gamma_{J_k}^{-1}\frac{1}{B}\sum_{i=1}^B \varepsilon_{i,k}\Psi_{J_k}(Z_{0,i,k})$, we define $\varepsilon_{1,N,i,k}$ and $\varepsilon_{2,N,i,k}$ as in (C.4) with M_N to be chosen later. For any $\varrho > 0$ and M_N, we have that

$$\Pr\left(\sup_{z\in\mathcal{Z}_0}\left\|\frac{1}{N}\sum_{k=1}^N \Psi_{J_k}(z)^\top \Gamma_{J_k}^{-1}\frac{1}{B}\sum_{i=1}^B \varepsilon_{2,N,i,k}\Psi_{J_k}(Z_{0,i,k})\right\|_2 > \varrho\right) \le C\varrho^{-1}J_N M_N^{1-\kappa}.$$

For any positive integer L, Condition 7(iii) provides $z_1,\ldots,z_L \in \mathcal{Z}_0$ such that

$$\sup_{z\in\mathcal{Z}_0}\min_{1\le l\le L}\|\Psi_{J_N}(z) - \Psi_{J_N}(z_l)\|_2 \le CJ_N^{3/2}L^{-1}.$$

By nestedness, $\Psi_{J_k}(z) = R_{J_k,J_N}\Psi_{J_N}(z)$ for every $J_k \le J_N$, and $\sup_{J_1\le J_2}\|R_{J_1,J_2}\|_{\mathrm{op}} < \infty$ under the normalized-basis construction. Hence this single terminal net controls all lower-dimensional rows appearing in the sum, up to a fixed multiplicative constant. As a result,

$$\begin{aligned}
&\sup_{z\in\mathcal{Z}_0}\left\|\frac{1}{N}\sum_{k=1}^N \Psi_{J_k}(z)^\top \Gamma_{J_k}^{-1}\frac{1}{B}\sum_{i=1}^B \varepsilon_{1,N,i,k}\Psi_{J_k}(Z_{0,i,k})\right\|_2 \\
&\le \max_{1\le l\le L}\left\|\frac{1}{N}\sum_{k=1}^N \Psi_{J_k}(z_l)^\top \Gamma_{J_k}^{-1}\frac{1}{B}\sum_{i=1}^B \varepsilon_{1,N,i,k}\Psi_{J_k}(Z_{0,i,k})\right\|_2 + \frac{CJ_N^2}{NL}\sum_{k=1}^N \frac{1}{B}\sum_{i=1}^B |\varepsilon_{1,N,i,k}|
\end{aligned}$$

Note that $\frac{CJ_N^2}{NL}\sum_{k=1}^N \frac{1}{B}\sum_{i=1}^B |\varepsilon_{1,N,i,k}| \leq CL^{-1}N^{2\alpha_\dagger}M_N$. And

$$\begin{aligned}
&\Pr\left(\max_{1\leq l\leq L}\left\|\frac{1}{N}\sum_{k=1}^N \Psi_{J_k}(z_l)^\top \Gamma_{J_k}^{-1}\frac{1}{B}\sum_{i=1}^B \varepsilon_{1,N,i,k}\Psi_{J_k}(Z_{0,i,k})\right\|_2 > \varrho\right)\\
&\leq \sum_{1\leq l\leq L}\Pr\left(\left\|\frac{1}{N}\sum_{k=1}^N \Psi_{J_k}(z_l)^\top \Gamma_{J_k}^{-1}\frac{1}{B}\sum_{i=1}^B \varepsilon_{1,N,i,k}\Psi_{J_k}(Z_{0,i,k})\right\|_2 > \varrho\right)\\
&\leq C\exp\left(\log(L) - C_1N^{1-\alpha_\dagger}\varrho^2\right) + C\exp\left(\log(L) - C_2N^{1-\alpha_\dagger}M_N^{-1}\varrho\right)
\end{aligned}$$

If we choose

$$\varrho = \sqrt{C_1^{-1}N^{-(1-\alpha_\dagger)}(\log(L) + 2\log(N))},$$

$$M_N = \sqrt{\frac{C_2^2N^{1-\alpha_\dagger}}{C_1(\log(L)+2\log(N))}},$$

and

$$L = \left\lceil \frac{N^{1+\alpha_\dagger}}{\log N}\right\rceil,$$

we have that

$$\sup_{z\in\mathcal{Z}_0}\left\|\frac{1}{N}\sum_{k=1}^N \Psi_{J_k}(z)^\top \Gamma_{J_k}^{-1}\frac{1}{B}\sum_{i=1}^B \varepsilon_{1,N,i,k}\Psi_{J_k}(Z_{0,i,k})\right\|_2 = O\left(N^{-\frac{1-\alpha_\dagger}{2}}\log^{\frac{1}{2}}(N)\right), a.s.,$$

and if $\kappa > 2 + \frac{2(1+\alpha_\dagger)}{1-\alpha_\dagger}$, we have that

$$\sum_{N=1}^\infty P\left(\sup_{z\in\mathcal{Z}_0}\left\|\frac{1}{N}\sum_{k=1}^N \Psi_{J_k}(z)^\top \Gamma_{J_k}^{-1}\frac{1}{B}\sum_{i=1}^B \varepsilon_{2,N,i,k}\Psi_{J_k}(Z_{0,i,k})\right\|_2 > \varrho\right) < \infty$$

So together we have that

$$\sup_{z\in\mathcal{Z}_0}\left|\frac{1}{N}\sum_{k=1}^N \Psi_{J_k}(z)^\top \Gamma_{J_k}^{-1}\frac{1}{B}\sum_{i=1}^B \varepsilon_{i,k}\Psi_{J_k}(Z_{0,i,k})\right| = O\left(N^{-\frac{1-\alpha_\dagger}{2}}\log^{\frac{1}{2}}(N)\right), a.s.$$

Combining this display with the exact bias and generated-regressor decompositions gives

$$\|\overline{F}_N - F_0\|_\infty = O_{\text{a.s.}}\left(\sqrt{\frac{J_N\log N}{N}} + \frac{1}{N}\sum_{k=1}^N J_k^{-s} + \left\|\frac{1}{N}\sum_{k=1}^N \Delta\check{\theta}_{k-1}\right\|_2 + A_{\mu,N}\right),$$

which is (C.3) and finishes the proof.

Algorithm 3 Joint Refinement Learner with Plug-in AME Average (Phase II)

Require: streaming batches $\mathcal{W}_k$, $k = 1, \dots, N$; sieve dimensions $\{J_k\}$; learning rates $\{\xi_k\}$; initial estimate $\widetilde{\omega}_0 = (\widetilde{\theta}_0^\top, \widetilde{\beta}_0^\top)^\top$; predictable conditioning matrices $\{\widehat{G}_k\}$; predictable guardrail centers $\{\widehat{\omega}_k\}$; constant $C_\Omega > 0$.

1: $\overline{\overline{\theta}}_0 \leftarrow \widetilde{\theta}_0$, $\overline{\overline{\beta}}_0 \leftarrow \widetilde{\beta}_0$

2: **for** $k = 1, 2, \dots, N$ **do**

3: Construct $\widehat{\Omega}_k = \left\{\omega \in \mathbb{R}^{p+J_k} : \|\omega - \widehat{\omega}_k\|_2 \le \frac{C_\Omega}{2} J_k^{-2}\right\}$

4: Re-embed the pre-update state: $\omega_{k-1}^{(k)} \leftarrow \mathcal{R}_{J_{k-1},J_k}^\top \widetilde{\omega}_{k-1}$

5: **Projected joint gradient update**:

$$\widetilde{\omega}_k \leftarrow \Pi_{\widehat{\Omega}_k}\left[\omega_{k-1}^{(k)} - \frac{\xi_k \widehat{G}_k}{B} \sum_{i=1}^{B} \nabla_\omega \mathcal{L}_{J_k}\left(\omega_{k-1}^{(k)}, W_{i,k}\right)\right].$$

6: **Update PR averages:**

$$\overline{\overline{\theta}}_k \leftarrow \frac{1}{k}\widetilde{\theta}_k + \frac{k-1}{k}\overline{\overline{\theta}}_{k-1}, \qquad \overline{\overline{\beta}}_k \leftarrow \frac{1}{k}\widetilde{\beta}_k + \frac{k-1}{k} R_{J_{k-1},J_k}^\top \overline{\overline{\beta}}_{k-1}$$

7: **return** $\overline{\overline{\theta}}_N$ and $\overline{\overline{F}}_N(\cdot) = \Psi_{J_N}(\cdot)^\top \overline{\overline{\beta}}_N$

D. Phase II Joint Local Refinement

D.1. Pseudo Code

The pseudo code for the second phase local refinement is given in Algorithm 3.

D.2. Sufficient Conditions and Discussions

Condition 10. *(i)* $\|\nabla_{zz}\Psi_J\|_\infty \le CJ^{5/2}$ *and* $\|\nabla_{zzz}\Psi_J\|_\infty \le CJ^{7/2}$*; (ii)* $\mathbb{E}[(1 + |\varepsilon_{i,k}|^\kappa)(1 + \|X_{i,k}\|_2^\kappa)] < \infty$ *for some* $\kappa > 2 + \max\{\frac{2}{\alpha_\xi}, \frac{2(1+\alpha_\dagger)}{1-\alpha_\dagger}\}$*; (iii)* $\|\nabla_{zz}[F_0(\cdot) - \mathbb{P}_J(F_0)(\cdot)]\|_\infty \le C_{F_0} J^{2-s}$, $\|\nabla_{zzz}[F_0(\cdot) - \mathbb{P}_J(F_0)(\cdot)]\|_\infty \le C_{F_0} J^{3-s}$, *and* $\|\nabla_{zzz} F_0\|_\infty < \infty$*, where* C_{F_0} *is the constant defined in Condition 5(ii).*

Condition 11. *There holds*

$$0 < \underline{\lambda}_{\mathcal{L}} \le \inf_J \lambda_{\min}(\nabla_{\omega\omega}\mathcal{L}_J(\omega_{J,0})) \le \sup_J \lambda_{\max}(\nabla_{\omega\omega}\mathcal{L}_J(\omega_{J,0})) \le \overline{\lambda}_{\mathcal{L}} < \infty.$$

Condition 12. *There is a fixed $\delta_\Omega \in (0,1)$ such that, eventually almost surely,*

$$\|\widehat{\omega}_k - \omega_{J_k,0}\|_2 \le (1-\delta_\Omega)\frac{C_\Omega}{2} J_k^{-2}.$$

Moreover, $\sup_{\omega\in\widehat{\Omega}_k} \|\nabla_{\omega\omega}\mathcal{L}_{J_k}(\omega) - \nabla_{\omega\omega}\mathcal{L}_{J_k}(\omega_{J_k,0})\|_{\mathrm{op}} \le \underline{\lambda}_{\mathcal{L}}/2$, and, for every fixed $C_\omega > 0$, $\sup_{\|\omega-\omega_{J,0}\|_2 \le C_\omega J^{-2}} \|\nabla_{\omega\omega}\mathcal{L}_J(\omega)\|_{\mathrm{op}} \le C'_\omega$ with C'_ω independent of J.

For later reference, let the deterministic Phase II basin be

$$\Omega_J := \left\{\omega \in \mathbb{R}^{p+J} : \|\omega - \omega_{J,0}\|_2 \le C_\Omega J^{-2}\right\}. \tag{D.1}$$

Condition 13 (Estimated conditioning matrix)**.** *(i) For every k, $\widehat{G}_k \in \mathbb{R}^{(p+J_k)\times(p+J_k)}$ is $\mathcal{F}_{k-1}$-measurable, symmetric, and positive definite; (ii) There are deterministic symmetric positive-definite matrices $G_{J,0}$ and constants $0 < \underline{g} \le \overline{g} < \infty$ such that*

$$\underline{g} \le \inf_J \lambda_{\min}(G_{J,0}) \le \sup_J \lambda_{\max}(G_{J,0}) \le \overline{g}.$$

Suppose also that, for some $c_G > 0$,

$$\inf_J \inf_{\omega\in\Omega_J} \lambda_{\min}\left(\frac{G_{J,0}\nabla_{\omega\omega}\mathcal{L}_J(\omega) + \nabla_{\omega\omega}\mathcal{L}_J(\omega)G_{J,0}}{2}\right) \ge c_G; \tag{D.2}$$

(iii) Writing $d_{G,k} := \|\widehat{G}_k - G_{J_k,0}\|_{\mathrm{op}}$, suppose that, for some $\alpha_G > 0$,

$$d_{G,k} \le \|\widehat{G}_k - G_{J_k,0}\|_F = O_{\mathrm{a.s.}}(k^{-\alpha_G}), \qquad \alpha_G > \frac{1-\alpha_\xi+\alpha_\dagger}{2}. \tag{D.3}$$

Thus $\widehat{G}_k$ also has eigenvalues uniformly bounded away from zero and infinity eventually almost surely.

Remark D.1. *The consistency $d_{G,k} = o_{\mathrm{a.s.}}(1)$, the uniform eigenvalue bounds, and (D.2) are already sufficient for localization, eventual inactivity of the projection, and the raw-iterate rates. The compatibility condition is automatic for the inverse-Hessian target $G_{J,0} = \{\nabla_{\omega\omega}\mathcal{L}_J(\omega_{J,0})\}^{-1}$ after C_Ω is chosen so that the local Hessian-variation bound underlying Condition 12 holds throughout Ω_J: the matrix in (D.2) is then a uniformly small perturbation of the identity. For a general positive-definite target, bounded eigenvalues alone do not ensure Euclidean one-step contraction; alternatively one may use projection in the $G_{J,0}^{-1}$-metric, in which case (D.2) is not needed.*

On the other side, a rate is needed only for the strong $o_{\mathrm{a.s.}}(N^{-1/2})$ remainder in the PR *representation. The polynomial condition (D.3) is a transparent sufficient condition. The proof below uses the two primitive requirements*

$$\sum_{k=1}^{\infty} \frac{J_k d_{G,k}^2}{k} < \infty, \qquad \frac{1}{\sqrt{N}} \sum_{k=1}^{N} d_{G,k} k^{-(\alpha_\xi - \alpha_\dagger)/2} \text{PolyLog} \to 0, \quad a.s.,$$

together with $d_{G,k} = o_{\text{a.s.}}(1)$ and the analogous, smaller secondary state term. The first condition makes the predictably gain-weighted score a common-Hilbert-space martingale with summable normalized quadratic variation; the second controls the linear state component. Condition (D.3) implies both because $\alpha_\xi < 1 - 3\alpha_\dagger$ and $\alpha_G > (1 - \alpha_\xi + \alpha_\dagger)/2$. These primitive conditions may be imposed directly instead of the polynomial rate.

Condition 14. *Let $s > \frac{7}{2}$ be the approximation error rate of F_0 defined in Condition 5 and s_μ be defined in Condition 6(ii), and define*

$$q_\star := s + \min(s_\mu,\, s - 1). \tag{D.4}$$

The learning rate α_ξ and sieve growth exponent $\alpha_\dagger$ satisfy $\max\left\{\frac{1}{2} + 3\alpha_\dagger,\ 1 + (4 - 2s)\alpha_\dagger\right\} < \alpha_\xi < \min\left\{1 - 3\alpha_\dagger,\ \frac{1}{2} + (s - \frac{1}{2})\alpha_\dagger\right\}$, and moreover, $q_\star \alpha_\dagger > \frac{1}{2}$.

Remark D.2. *Condition 10 further regulates the norm of second- and third-order derivatives of spline basis functions. Such regulation is necessary when we consider NLS criterion. Note that for normalized spline basis functions, such regulation is automatically valid. Condition 11 regulates the curvature of the NLS loss function at the pseudo true parameter $\omega_{J,0}$. In the following proofs, we show that $\|\mathbb{M}_J - \nabla_{\omega\omega}\mathcal{L}_J(\omega_{J,0})\|_F \le CJ^{\frac{3}{2}-s}$, where*

$$\mathbb{M}_J = \mathbb{E}\begin{pmatrix} (\nabla_z F_0(Z_0))^2 X X^\top & \nabla_z F_0(Z_0) X \Psi_J(Z_0)^\top \\ \nabla_z F_0(Z_0) \Psi_J(Z_0) X^\top & \Psi_J(Z_0)\Psi_J(Z_0)^\top \end{pmatrix}.$$

When $s > \frac{3}{2}$, Condition 11 is equivalent to assuming uniformly (with respect to J) bounded eigenvalues from both below and above for $\mathbb{M}_J$. Condition 12 imposes a strict interior condition that is used to show eventual inactivity of projection. The Phase I rates in Theorem B.1 and Theorem C.1, with tuning chosen strictly inside Condition 9, are sufficient to furnish this margin after an almost surely finite time.

D.3. Discussion on Condition 12

This subsection shows that under mild conditions, the local basin with radius of order J^{-2} preserves the curvature. To show this, note that the Hessian matrix of the NLS loss function is given by

$$\nabla_{\omega\omega}\mathcal{L}_J(\omega, W) = \begin{pmatrix} \nabla_{\theta\theta}\mathcal{L}_J(\omega, W) & \nabla_{\theta\beta}\mathcal{L}_J(\omega, W) \\ \nabla_{\beta\theta}\mathcal{L}_J(\omega, W) & \nabla_{\beta\beta}\mathcal{L}_J(\omega, W) \end{pmatrix},$$

where by simple calculation

$$\nabla_{\theta\theta}\mathcal{L}_J(\omega, W) = ((\nabla_z\Psi_J(x_0+X^\top\theta))^\top\beta)^2 XX^\top - (Y - \Psi_J(x_0+X^\top\theta)^\top\beta)(\nabla_{zz}\Psi_J(x_0+X^\top\theta))^\top\beta XX^\top,$$

$$\nabla_{\theta\beta}\mathcal{L}_J(\omega, W) = (\nabla_z\Psi_J(x_0+X^\top\theta))^\top\beta X\Psi_J(x_0+X^\top\theta)^\top - (Y - \Psi_J(x_0+X^\top\theta)^\top\beta)X(\nabla_z\Psi_J(x_0+X^\top\theta))^\top,$$

and

$$\nabla_{\beta\beta}\mathcal{L}_J(\omega, W) = \Psi_J(x_0 + X^\top\theta)\Psi_J(x_0 + X^\top\theta)^\top.$$

We will calculate the bound of the derivative for each of the Hessian block with the local basin. In particular, we prove the following result.

Lemma D.1. *Suppose that, for $m = 0, 1, 2, 3$,*

$$J^{-(2m+1)/2}\|\nabla_z^m\Psi_J\|_\infty \le C, \qquad \|\nabla_z^m F_0\|_\infty \le C, \qquad \|\nabla_z^m(F_0 - \mathbb{P}_J F_0)\|_\infty \le CJ^{m-s},$$

with $s > 3$, and suppose $\mathbb{E}\|X\|_2^4 < \infty$. For every fixed $C_1 < \infty$ there is $C_2(C_1) < \infty$, independent of J, such that, whenever $\|\omega_i - \omega_{J,0}\|_2 \le C_1 J^{-2}$, $i = 1, 2$,

$$\|\nabla_{\omega\omega}\mathcal{L}_J(\omega_1) - \nabla_{\omega\omega}\mathcal{L}_J(\omega_2)\|_F \le C_2 J^2\|\omega_1 - \omega_2\|_2.$$

Proof. Write $t_\theta = x_0 + X^\top\theta$, $f_\beta(t) = \Psi_J(t)^\top\beta$, and $r_\omega = F_0(Z_0) - f_\beta(t_\theta)$. On the stated ball, the inverse inequalities and the approximation assumptions imply, uniformly in t,

$$|f_\beta(t)| + |f'_\beta(t)| \le C, \qquad |f''_\beta(t)| \le CJ^{1/2}, \qquad |f_\beta^{(3)}(t)| \le CJ^{3/2}, \tag{D.5}$$

and

$$|r_\omega| \le C\{J^{-3/2} + J^{-2}\|X\|_2\}. \tag{D.6}$$

Indeed, the coefficient displacement contributes $J^{(2m+1)/2}J^{-2}$ to the mth derivative, while $f_{\beta_{J,0}} = \mathbb{P}_J F_0$; the index displacement contributes $C\|X\|_2 J^{-2}$ to the residual.

It is enough to bound the directional derivative of the population Hessian. Let $h = (u^\top, v^\top)^\top$ be a unit vector and put $\dot{t} = X^\top u$. Along $\omega + qh$,

$$\dot{f} = f'_\beta \dot{t} + \Psi_J^\top v, \quad \dot{f}' = f''_\beta \dot{t} + (\nabla_z \Psi_J)^\top v, \quad \dot{f}'' = f^{(3)}_\beta \dot{t} + (\nabla_{zz} \Psi_J)^\top v, \quad \dot{r} = -\dot{f}.$$

Taking expectation of the sample Hessian replaces Y by $F_0(Z_0)$, so the three population Hessian blocks are

$$\begin{aligned}
H_{\beta\beta}(\omega) &= \mathbb{E}[\Psi_J(t_\theta)\Psi_J(t_\theta)^\top], \\
H_{\theta\theta}(\omega) &= \mathbb{E}[\{(f'_\beta)^2 - r_\omega f''_\beta\} X X^\top], \\
H_{\theta\beta}(\omega) &= \mathbb{E}[X\{f'_\beta \Psi_J(t_\theta)^\top - r_\omega \nabla_z \Psi_J(t_\theta)^\top\}].
\end{aligned}$$

For the first block,

$$\|DH_{\beta\beta}(\omega)[h]\|_F \le 2\,\mathbb{E}\big[\|\Psi_J(t_\theta)\|_2 \|\nabla_z \Psi_J(t_\theta)\|_2 |\dot{t}|\big] \le CJ^2.$$

For the scalar multiplying $XX^\top$ in $H_{\theta\theta}$,

$$D\{(f'_\beta)^2 - r_\omega f''_\beta\}[h] = 2 f'_\beta \dot{f}' + \dot{f}\, f''_\beta - r_\omega \dot{f}''.$$

Using (D.5)–(D.6), its absolute value is at most $C\{J^{3/2} + J^{1/2}\|X\|_2 + J^{-1/2}\|X\|_2^2\}$. Consequently, by $\mathbb{E}\|X\|_2^4 < \infty$,

$$\|DH_{\theta\theta}(\omega)[h]\|_F \le CJ^{3/2}.$$

Finally,

$$\begin{aligned}
D\{f'_\beta \Psi_J^\top - r_\omega (\nabla_z \Psi_J)^\top\}[h] = {} & \dot{f}' \Psi_J^\top + f'_\beta (\nabla_z \Psi_J)^\top \dot{t} \\
& + \dot{f} (\nabla_z \Psi_J)^\top - r_\omega (\nabla_{zz} \Psi_J)^\top \dot{t}.
\end{aligned}$$

The same envelopes, followed by multiplication by X and Hölder's inequality, give $\|DH_{\theta\beta}(\omega)[h]\|_F \le CJ^2$. The transpose block has the same bound. Integrating these directional bounds along the line segment joining ω_1 and ω_2, which lies in the convex local ball, proves the claim. □

D.4. Proofs of Lemma 5.3

Similar to the proof of Theorem C.1, the proof is divided into three steps. To keep the already lengthy score decomposition readable, the three baseline steps are first written for an auxiliary identity-gain path; throughout those steps, $\widetilde{\omega}_k$, D_k, and T_k refer to that auxiliary path. Subsection D.4.4 then repeats the needed stability bounds for the actual conditioned path, replaces the identity by the population matrices $G_{J,0}$, and finally controls the predictable estimates $\widehat{G}_k$. That subsection verifies both the gain-dependent raw recursion and the cancellation that leaves the PR influence function unchanged. Before starting, we record two auxiliary lemmas: a *centered* bound for the sieve score at the pseudo-true parameter, and a *telescoping* bound for the sieve-target jumps generated by nestedness of the sieve spaces. Throughout, recall $q_\star = s + \min(s_\mu, s-1)$ from (D.4), $\mathcal{J}(m) = \min\{k : J_k \geq m\}$, and $h(z) := (\nabla_z F_0(z))\, \mu_0(\theta_0, z)$.

Lemma D.2 (Centered sieve score)**.** *Let Condition 5–7 and Condition 10 hold with $s > 2$. Define*

$$b_J := \mathbb{E}\left[\nabla_\omega \mathcal{L}_J(\omega_{J,0}, W)\right] = \left(b_{\theta,J}^\top, b_{\beta,J}^\top\right)^\top.$$

Then (i) $b_{\beta,J} = 0$; (ii) $\|b_{\theta,J}\|_2 \leq CJ^{-q_\star}$; (iii) $\mathbb{E}\|\nabla_\omega \mathcal{L}_J(\omega_{J,0}, W) - b_J\|_2^2 \leq CJ$.

Proof. Recall $\nabla_\beta \mathcal{L}_J(\omega_{J,0}, W) = -(Y - \Psi_J(Z_0)^\top \beta_{J,0})\Psi_J(Z_0)$. Since $\mathbb{E}(\varepsilon | x_0, X) = 0$,

$$b_{\beta,J} = -\mathbb{E}\left[\left(F_0(Z_0) - \mathbb{P}_J(F_0)(Z_0)\right)\Psi_J(Z_0)\right] = 0$$

by the definition of the L_2-projection (C.1). This proves (i). For (ii), $\nabla_\theta \mathcal{L}_J(\omega_{J,0}, W) = -(Y - \Psi_J(Z_0)^\top \beta_{J,0})(\nabla_z \Psi_J(Z_0))^\top \beta_{J,0} X$, so conditioning on Z_0,

$$\begin{aligned} b_{\theta,J} &= -\mathbb{E}\left[\left(F_0 - \mathbb{P}_J(F_0)\right)(Z_0)\, \nabla_z \mathbb{P}_J(F_0)(Z_0)\, \mu_0(\theta_0, Z_0)\right] \\ &= -\mathbb{E}\left[\left(F_0 - \mathbb{P}_J(F_0)\right)(Z_0)\, h(Z_0)\right] - \mathbb{E}\left[\left(F_0 - \mathbb{P}_J(F_0)\right)(Z_0)\left(\nabla_z \mathbb{P}_J(F_0) - \nabla_z F_0\right)(Z_0)\, \mu_0(\theta_0, Z_0)\right]. \end{aligned}$$

By the L_2-projection property, $\mathbb{E}[(F_0 - \mathbb{P}_J(F_0))(Z_0) g(Z_0)] = 0$ for every $g \in \mathcal{S}_J$ componentwise; applying this with $g = \mathbb{P}_J(h)$ and using Condition 5 and Condition 6(ii),

$$\left\|\mathbb{E}\left[\left(F_0 - \mathbb{P}_J(F_0)\right)(Z_0) h(Z_0)\right]\right\|_2 = \left\|\mathbb{E}\left[\left(F_0 - \mathbb{P}_J(F_0)\right)(Z_0)\left(h - \mathbb{P}_J(h)\right)(Z_0)\right]\right\|_2 \leq CJ^{-s} \cdot C_\mu J^{-s_\mu}.$$

For the second term, Condition 5(ii) gives $\|F_0 - \mathbb{P}_J(F_0)\|_\infty \leq CJ^{-s}$ and $\|\nabla_z[\mathbb{P}_J(F_0) - F_0]\|_\infty \leq CJ^{1-s}$, while $\mathbb{E}\|\mu_0(\theta_0, Z_0)\|_2 \leq (\mathbb{E}\|X\|_2^2)^{1/2} < \infty$; hence this term is $O(J^{-s} \cdot J^{1-s}) = O(J^{-(2s-1)})$. Since $q_\star = s + \min(s_\mu, s-1) \leq \min\{s + s_\mu, 2s - 1\}$, (ii) follows. For (iii), $\nabla_z \mathbb{P}_J(F_0)$ is uniformly bounded for $s > 1$, so $\mathbb{E}\|\nabla_\theta \mathcal{L}_J(\omega_{J,0}, W)\|_2^2 \leq C\mathbb{E}[(1 + \varepsilon^2)\|X\|_2^2] \leq C$ by Condition 6; and

$$\mathbb{E}\|\nabla_\beta \mathcal{L}_J(\omega_{J,0}, W)\|_2^2 \leq C\, \mathbb{E}\|\Psi_J(Z_0)\|_2^2 = C\, \mathrm{tr}(\Gamma_J) \leq CJ$$

by Condition 7(ii). Subtracting the mean b_J does not increase the order. $\square$

Lemma D.3 (Telescoping bound for sieve-target jumps)**.** *Let Condition 5 and Condition 7 hold, and recall from Section 3.2 that the sieve spaces are nested with the exact refinement relation (3.3). For $m \geq J_0+1$ define the target jump*

$$\delta_m := \left\|\beta_{m,0} - R_{m-1,m}^\top \beta_{m-1,0}\right\|_2 .$$

Then there is a constant C_δ independent of m such that:

(i) $\delta_m \leq C_\delta m^{-s}$;

(ii) $\sum_{m>M} \delta_m^2 \leq C_\delta^2 M^{-2s}$ *for every* $M \geq J_0$;

(iii) for every $\nu \geq 0$ there is $C(\nu) < \infty$ such that for all $M \geq 2J_0$,

$$\sum_{m=J_0+1}^{M} m^\nu \delta_m \leq C(\nu)\left(M^{\nu+\frac{1}{2}-s} + \log M\right), \qquad \sum_{m=J_0+1}^{M} m^\nu \delta_m^2 \leq C(\nu)\left(M^{\nu-2s} + \log M\right).$$

Proof. By the exact refinement relation $\Psi_{m-1}(z) = R_{m-1,m}\Psi_m(z)$, the vector $\beta_{m,0} - R_{m-1,m}^\top \beta_{m-1,0}$ collects the coefficients, in the basis Ψ_m, of the *detail* function

$$\Delta_m(z) := \mathbb{P}_m(F_0)(z) - \mathbb{P}_{m-1}(F_0)(z).$$

Since $\mathbb{P}_m$ is the orthogonal projection of $L_2(Z_0)$ onto $\mathcal{S}_m$ and $\mathcal{S}_{m-1} \subseteq \mathcal{S}_m$, we have $\Delta_m \perp \mathcal{S}_{m-1}$ in $L_2(Z_0)$, while $\Delta_{m'} \in \mathcal{S}_{m'} \subseteq \mathcal{S}_{m-1}$ for every $m' < m$. Hence

$$\begin{aligned}
&\sum_{m=M+1}^{M'} \|\Delta_m\|_{L_2(Z_0)}^2 = \|\mathbb{P}_{M'}(F_0) - \mathbb{P}_M(F_0)\|_{L_2(Z_0)}^2 \\
&= \|\mathbb{P}_{M'}(F_0 - \mathbb{P}_M(F_0))\|_{L_2(Z_0)}^2 \leq \|F_0 - \mathbb{P}_M(F_0)\|_\infty^2 \leq C^2 M^{-2s},
\end{aligned}$$

where the last step uses Condition 5. Letting $M' \to \infty$ and using $\delta_m^2 \leq \lambda_{\min}(\Gamma_m)^{-1}\|\Delta_m\|_{L_2(Z_0)}^2 \leq C\|\Delta_m\|_{L_2(Z_0)}^2$ by Condition 7(ii) proves (ii); taking $M = m-1$ and keeping the single summand δ_m^2 proves (i). For (iii), decompose $\{J_0+1, \ldots, M\}$ into dyadic blocks $B_\ell = \{m : 2^\ell \leq m < 2^{\ell+1}\}$. By the Cauchy–Schwarz inequality and (ii),

$$\sum_{m \in B_\ell} m^\nu \delta_m \leq 2^{(\ell+1)\nu}\left(2^\ell\right)^{\frac{1}{2}}\left(\sum_{m \geq 2^\ell} \delta_m^2\right)^{\frac{1}{2}} \leq C 2^{\ell\left(\nu+\frac{1}{2}-s\right)}, \qquad \sum_{m \in B_\ell} m^\nu \delta_m^2 \leq C 2^{\ell(\nu-2s)}.$$

Summing the geometric series over $\ell \le \log_2 M$ gives the displayed bounds, where the term $\log M$ covers the cases in which the exponent is nonpositive. $\square$

Remark D.3. *Since $J_k = [J_0 k^{\alpha_\dagger}]$, we have $J_k - J_{k-1} \in \{0, 1\}$ for all k sufficiently large, so eventually each sieve expansion inserts exactly one knot. For the finitely many early steps with $J_k - J_{k-1} > 1$, telescoping through intermediate dimensions and $\sup_{J_1 \le J_2} \|R_{J_1,J_2}\|_{\mathrm{op}} < \infty$ give $\|R^\top_{J_{k-1},J_k}\beta_{J_{k-1},0} - \beta_{J_k,0}\|_2 \le C\sum_{m=J_{k-1}+1}^{J_k}\delta_m$, so these steps only affect constants. In particular, Lemma D.3(i) recovers the bound $\|R^\top_{J_{k-1},J_k}\beta_{J_{k-1},0} - \beta_{J_k,0}\|_2 \le CJ_k^{-s}$ used in the proof of Theorem C.1, while Lemma D.3(ii)–(iii) are strictly sharper* on average *across expansions: the multiresolution (nested-projection) structure makes the jump sizes square-summable at rate M^{-2s}, which is the property exploited in Section D.4.3 below.*

Lemma D.4 (Common-space martingale averaging)**.** *Let $U_k = (U^\top_{\theta,k}, U^\top_{\beta,k})^\top \in \mathbb{R}^{p+J_k}$ be an $\mathcal{F}_k$-adapted square-integrable martingale difference array. Suppose*

$$\sum_{k=1}^{\infty} \frac{1}{k}\,\mathbb{E}_{k-1}\|U_k\|_2^2 < \infty \qquad a.s. \tag{D.7}$$

Then

$$\frac{1}{\sqrt{N}}\left\|\sum_{k=1}^{N} \mathcal{R}^\top_{J_k,J_N} U_k\right\|_2 \longrightarrow 0 \qquad a.s. \tag{D.8}$$

Moreover, if $C_k \in \mathbb{R}^{q\times(p+J_k)}$ is $\mathcal{F}_{k-1}$-measurable for a fixed q and $\sup_k \|C_k\|_{\mathrm{op}} < \infty$ almost surely, then

$$\max_{1\le n\le N} \frac{1}{\sqrt{N}}\left\|\sum_{k=1}^{n} C_k U_k\right\|_2 \longrightarrow 0 \qquad a.s. \tag{D.9}$$

The conclusions remain true after multiplying U_k by any predictable scalar a_k with $|a_k| \le 1$.

Proof. Put $\mathcal{H} = \mathbb{R}^p \oplus L_2(P_Z)$ and define the deterministic common-coordinate embedding

$$\iota_J(v_\theta, v_\beta) = (v_\theta, \Psi_J^\top v_\beta).$$

By Condition 7,

$$c\|v\|_2^2 \le \|\iota_J v\|_{\mathcal{H}}^2 = \|v_\theta\|_2^2 + v_\beta^\top \Gamma_J v_\beta \le C\|v\|_2^2$$

uniformly in J. Exact re-embedding gives, for $J_1 \le J_2$,

$$\iota_{J_2}(\mathcal{R}^\top_{J_1,J_2} v) = \iota_{J_1}(v), \tag{D.10}$$

because $\Psi_{J_1} = R_{J_1,J_2}\Psi_{J_2}$. After localization at the stopping times at which the series in

(D.7) first exceeds an integer, the Hilbert-space martingale $\sum_k k^{-1/2}\iota_{J_k}(U_k)$ has summable conditional quadratic variation and therefore converges almost surely. Kronecker's lemma in $\mathcal{H}$ yields $N^{-1/2}\sum_{k\le N}\iota_{J_k}(U_k)\to 0$. Equation (D.10) and norm equivalence prove (D.8).

For the row statement, C_kU_k is an $\mathbb{R}^q$-valued martingale difference and its normalized conditional quadratic variation is summable by (D.7); the same argument gives $n^{-1/2}\sum_{k\le n}C_kU_k\to 0$. Given $\epsilon>0$, this endpoint convergence bounds all sufficiently late partial sums by $\epsilon\sqrt{n}$, while the finitely many early sums divided by $\sqrt{N}$ vanish. This proves (D.9). A predictable bounded scalar multiplier preserves all arguments. □

D.4.1. Obtaining Initial Convergence Rate

Summary: This part establishes the following results for $\widetilde{\omega}_k$: for any $r>1$,

$$\|\Delta\widetilde{\omega}_k\|_2 = O\left(k^{\frac{1+\alpha_\dagger-2\alpha_\xi}{2}}\log^{\frac{r}{2}}(k) + k^{\frac{1-\alpha_\xi}{2}-q_\star\alpha_\dagger}\log^{\frac{r}{2}}(k) + k^{-\left(s-\frac{1}{2}\right)\alpha_\dagger}\log^{\frac{r}{2}}(k)\right), \qquad \text{a.s.}$$

Under Condition 14, each of the three exponents is strictly smaller than $-2\alpha_\dagger$, so the projection is inactive almost surely for k large.

Define $\Delta\widetilde{\omega}_k = \widetilde{\omega}_k - \omega_{J_k,0}$, where recall that $\omega_{J_k,0} = (\theta_0^\top, \beta_{J_k,0}^\top)^\top$. Note that

$$
\begin{aligned}
&\mathcal{R}_{J_{k-1},J_k}^\top \widetilde{\omega}_{k-1} - \frac{\xi_k}{B}\sum_{i=1}^{B} \nabla_\omega \mathcal{L}_{J_k}\left(\mathcal{R}_{J_{k-1},J_k}^\top \widetilde{\omega}_{k-1}, W_{i,k}\right) - \omega_{J_k,0} \\
&= \mathcal{R}_{J_{k-1},J_k}^\top \Delta\widetilde{\omega}_{k-1} - \frac{\xi_k}{B}\sum_{i=1}^{B}\Big(\nabla_\omega \mathcal{L}_{J_k}(\mathcal{R}_{J_{k-1},J_k}^\top \widetilde{\omega}_{k-1}, W_{i,k}) - \nabla_\omega \mathcal{L}_{J_k}(\mathcal{R}_{J_{k-1},J_k}^\top \omega_{J_{k-1},0}, W_{i,k})\Big) \\
&- \frac{\xi_k}{B}\sum_{i=1}^{B}\Big(\nabla_\omega \mathcal{L}_{J_k}(\mathcal{R}_{J_{k-1},J_k}^\top \omega_{J_{k-1},0}, W_{i,k}) - \nabla_\omega \mathcal{L}_{J_k}(\omega_{J_k,0}, W_{i,k})\Big) \\
&- \frac{\xi_k}{B}\sum_{i=1}^{B} \nabla_\omega \mathcal{L}_{J_k}(\omega_{J_k,0}, W_{i,k}) + \mathcal{R}_{J_{k-1},J_k}^\top \omega_{J_{k-1},0} - \omega_{J_k,0} \\
&= \int_0^1 \Big(\mathbb{I}_{p+J_k} - \xi_k \nabla_{\omega\omega}\mathcal{L}_{J_k}\left(\mathcal{R}_{J_{k-1},J_k}^\top\left(\omega_{J_{k-1},0} + \tau\Delta\widetilde{\omega}_{k-1}\right)\right)\Big) d\tau \mathcal{R}_{J_{k-1},J_k}^\top \Delta\widetilde{\omega}_{k-1} \\
&- \frac{\xi_k}{B}\sum_{i=1}^{B}\int_0^1 \Big(\nabla_{\omega\omega}\mathcal{L}_{J_k}\left(\mathcal{R}_{J_{k-1},J_k}^\top\left(\omega_{J_{k-1},0} + \tau\Delta\widetilde{\omega}_{k-1}\right), W_{i,k}\right) \\
&\qquad\qquad - \nabla_{\omega\omega}\mathcal{L}_{J_k}\left(\mathcal{R}_{J_{k-1},J_k}^\top\left(\omega_{J_{k-1},0} + \tau\Delta\widetilde{\omega}_{k-1}\right)\right)\Big) d\tau \mathcal{R}_{J_{k-1},J_k}^\top \Delta\widetilde{\omega}_{k-1} \\
&- \frac{\xi_k}{B}\sum_{i=1}^{B}\Big(\nabla_\omega \mathcal{L}_{J_k}(\mathcal{R}_{J_{k-1},J_k}^\top \omega_{J_{k-1},0}, W_{i,k}) - \nabla_\omega \mathcal{L}_{J_k}(\omega_{J_k,0}, W_{i,k})\Big) \\
&- \frac{\xi_k}{B}\sum_{i=1}^{B}\left(\nabla_\omega \mathcal{L}_{J_k}(\omega_{J_k,0}, W_{i,k}) + \varphi_{J_k}(W_{i,k})\right) + \frac{\xi_k}{B}\sum_{i=1}^{B}\varphi_{J_k}(W_{i,k}) + \mathcal{R}_{J_{k-1},J_k}^\top \omega_{J_{k-1},0} - \omega_{J_k,0}.
\end{aligned}
$$

where

$$
\varphi_J(W) = \begin{pmatrix} (Y - F_0(Z_0))\,\nabla_z F_0(Z_0) X \\ (Y - F_0(Z_0))\,\Psi_J(Z_0) \end{pmatrix}.
$$

Now we analyze the above terms one by one. When $J_k = J_{k-1}$, $\mathcal{R}_{J_{k-1},J_k}^\top \widetilde{\omega}_{k-1} = \widetilde{\omega}_{k-1} \in \widehat{\Omega}_{k-1}$. Since $\widehat{\Omega}_{k-1}$ is convex by construction, Condition 11 and Condition 12 together lead to that the smallest eigenvalue of $\nabla_{\omega\omega}\mathcal{L}_{J_k}(\omega_{J_{k-1},0} + \tau\Delta\widetilde{\omega}_{k-1}) = \nabla_{\omega\omega}\mathcal{L}_{J_{k-1}}(\omega_{J_{k-1},0} + \tau\Delta\widetilde{\omega}_{k-1})$ is uniformly lower bounded from 0 for all $\tau \in [0,1]$. When $J_k > J_{k-1}$, note that

$$
\|\mathcal{R}_{J_{k-1},J_k}^\top\left(\omega_{J_{k-1},0} + \tau\Delta\widetilde{\omega}_{k-1}\right) - \omega_{J_k,0}\|_2 \le \|\mathcal{R}_{J_{k-1},J_k}^\top \omega_{J_{k-1},0} - \omega_{J_k,0}\|_2 + C\|\Delta\widetilde{\omega}_{k-1}\|_2 \le CJ_k^{-2},
$$

where the last inequality comes from the fact that $\widehat{\Omega}_k$ has radius of order J_k^{-2}, Condition 12, $s > 2$, and

$$
\|\mathcal{R}_{J_{k-1},J_k}^\top \omega_{J_{k-1},0} - \omega_{J_k,0}\|_2 = \|R_{J_{k-1},J_k}^\top \beta_{J_{k-1},0} - \beta_{J_k,0}\|_2 \le CJ_k^{-s},
$$

according to the proof of Theorem C.1. So $\nabla_{\omega\omega}\mathcal{L}_{J_k}(\mathcal{R}^\top_{J_{k-1},J_k}(\omega_{J_{k-1},0}+\tau\Delta\widetilde{\omega}_{k-1}))$ has eigenvalues bounded (in absolute value) by some constant independent of k and τ following Condition 12. Together there holds

$$
\begin{aligned}
&\mathbb{E}_{k-1}\left\|\int_0^1\left(\mathbb{I}_{p+J_k}-\xi_k\nabla_{\omega\omega}\mathcal{L}_{J_k}\left(\mathcal{R}^\top_{J_{k-1},J_k}\left(\omega_{J_{k-1},0}+\tau\Delta\widetilde{\omega}_{k-1}\right)\right)\right)d\tau\mathcal{R}^\top_{J_{k-1},J_k}\Delta\widetilde{\omega}_{k-1}\right\|_2^2\\
&\le(1-C\xi_k+C\mathbf{1}_k)\left\|\Delta\widetilde{\omega}_{k-1}\right\|_2^2,
\end{aligned}
$$

regardless of whether $J_k=J_{k-1}$ or $J_k>J_{k-1}$, where recall that $\mathbf{1}_k=\mathbf{1}(J_k>J_{k-1})$.

Note that the Hessian matrix of the NLS loss function is given by

$$
\nabla_{\omega\omega}\mathcal{L}_J(\omega,W)=\begin{pmatrix}\nabla_{\theta\theta}\mathcal{L}_J(\omega,W) & \nabla_{\theta\beta}\mathcal{L}_J(\omega,W)\\ \nabla_{\beta\theta}\mathcal{L}_J(\omega,W) & \nabla_{\beta\beta}\mathcal{L}_J(\omega,W)\end{pmatrix},
$$

where by simple calculation

$$
\nabla_{\theta\theta}\mathcal{L}_J(\omega,W)=((\nabla_z\Psi_J(x_0+X^\top\theta))^\top\beta)^2XX^\top-(Y-\Psi_J(x_0+X^\top\theta)^\top\beta)(\nabla_{zz}\Psi_J(x_0+X^\top\theta))^\top\beta XX^\top,
$$

$$
\nabla_{\theta\beta}\mathcal{L}_J(\omega,W)=(\nabla_z\Psi_J(x_0+X^\top\theta))^\top\beta X\Psi_J(x_0+X^\top\theta)^\top-(Y-\Psi_J(x_0+X^\top\theta)^\top\beta)X(\nabla_z\Psi_J(x_0+X^\top\theta))^\top,
$$

and

$$
\nabla_{\beta\beta}\mathcal{L}_J(\omega,W)=\Psi_J(x_0+X^\top\theta)\Psi_J(x_0+X^\top\theta)^\top.
$$

When $\omega\in\Omega_J$, inverse inequalities and the $O(J^{-2})$ coefficient radius give

$$
\begin{aligned}
\left\|\Psi_J^\top\beta-F_0\right\|_\infty&=O(J^{-3/2}+J^{-s}),\\
\left\|(\nabla_z\Psi_J)^\top\beta-\nabla_zF_0\right\|_\infty&=O(J^{-1/2}+J^{1-s}),\\
\left\|(\nabla_{zz}\Psi_J)^\top\beta-\nabla_{zz}F_0\right\|_\infty&=O(J^{1/2}+J^{2-s}).
\end{aligned}
$$

The last quantity is not uniformly bounded in J. Using the bounded true derivatives gives the envelope

$$
\|\nabla_{\omega\omega}\mathcal{L}_J(\omega,W)\|_F\le CJ^{1/2}(1+|Y|)\|X\|_2^2+CJ^{3/2}(1+|Y|)\|X\|_2+CJ.
$$

Condition 10(ii), which in particular has $\kappa>4$, makes the squared expectation of this envelope $O(J^3)$. Therefore

$$\mathbb{E}_{k-1}\left\|\frac{\xi_k}{B}\sum_{i=1}^{B}\int_0^1\left(\nabla_{\omega\omega}\mathcal{L}_{J_k}\left(\mathcal{R}_{J_{k-1},J_k}^{\top}\left(\omega_{J_{k-1},0}+\tau\Delta\widetilde{\omega}_{k-1}\right),W_{i,k}\right)\right.\right.$$
$$\left.\left.-\nabla_{\omega\omega}\mathcal{L}_{J_k}\left(\mathcal{R}_{J_{k-1},J_k}^{\top}\left(\omega_{J_{k-1},0}+\tau\Delta\widetilde{\omega}_{k-1}\right)\right)\right)d\tau\mathcal{R}_{J_{k-1},J_k}^{\top}\Delta\widetilde{\omega}_{k-1}\right\|_2^2\leq C\xi_k^2J_{k-1}^3\|\Delta\widetilde{\omega}_{k-1}\|_2^2.$$

This also leads to

$$\mathbb{E}_{k-1}\left\|\frac{\xi_k}{B}\sum_{i=1}^{B}\left(\nabla_\omega\mathcal{L}_{J_k}(\mathcal{R}_{J_{k-1},J_k}^{\top}\omega_{J_{k-1},0},W_{i,k})-\nabla_\omega\mathcal{L}_{J_k}(\omega_{J_k,0},W_{i,k})\right)\right\|_2^2$$
$$\leq C\xi_k^2J_k^3\|\mathcal{R}_{J_{k-1},J_k}^{\top}\omega_{J_{k-1},0}-\omega_{J_k,0}\|_2^2\leq C\xi_k^2J_k^{3-2s},$$

Moreover, note that

$$\nabla_\theta\mathcal{L}_J(\omega,W)=-(Y-\Psi_J(x_0+X^\top\theta)^\top\beta)(\nabla_z\Psi_J(x_0+X^\top\theta))^\top\beta X$$

and

$$\nabla_\beta\mathcal{L}_J(\omega,W)=-(Y-\Psi_J(x_0+X^\top\theta)^\top\beta)\Psi_J(x_0+X^\top\theta).$$

Note that due to J^{-s} and J^{1-s} global approximation error of F_0 and ∇_zF_0, we have that

$$\|\nabla_\omega\mathcal{L}_J(\omega_{J,0},W)+\varphi_J(W)\|_2\leq C(1+|Y|)(1+\|X\|_2)J^{\frac{3}{2}-s},$$

which leads to

$$\mathbb{E}_{k-1}\left\|\frac{\xi_k}{B}\sum_{i=1}^{B}\left(\nabla_\omega\mathcal{L}_{J_k}(\omega_{J_k,0},W_{i,k})+\varphi_{J_k}(W_{i,k})\right)\right\|_2^2\leq C\xi_k^2J_k^{3-2s}.$$

Also, we have that

$$\mathbb{E}_{k-1}\|\varphi_{J_k}(W_{i,k})\|_2^2\leq CJ_k.$$

Given the above results, we now expand the squared norm of the pre-projection update. Write

$$g_k:=\frac{1}{B}\sum_{i=1}^{B}\nabla_\omega\mathcal{L}_{J_k}(\omega_{J_k,0},W_{i,k}),\qquad\mathbb{E}_{k-1}\,g_k=b_{J_k},$$

where b_{J_k} is the mean sieve score of Lemma D.2, so that $\|b_{J_k}\|_2\leq CJ_k^{-q_\star}$ and $\mathbb{E}_{k-1}\|g_k-b_{J_k}\|_2^2\leq CJ_k$. All remaining stochastic terms in the decomposition displayed at the beginning of this subsection are handled exactly as bounded above. Collecting terms, the conditional mean of the pre-projection increment consists of (a) the contraction of $\mathcal{R}_{J_{k-1},J_k}^\top \Delta\widetilde{\omega}_{k-1}$ through the integrated population Hessian, (b) the deterministic mean score $-\xi_k b_{J_k}$, and, at expansion times only, (c) the deterministic terms $\mathcal{R}_{J_{k-1},J_k}^\top \omega_{J_{k-1},0} - \omega_{J_k,0}$ and

$$-\mathbb{E}\frac{\xi_k}{B}\sum_{i=1}^{B}[\nabla_\omega \mathcal{L}_{J_k}(\mathcal{R}_{J_{k-1},J_k}^\top \omega_{J_{k-1},0}, W_{i,k}) - \nabla_\omega \mathcal{L}_{J_k}(\omega_{J_k,0}, W_{i,k})].$$

All of these norms are at most CJ_k^{-s} by Lemma D.3(i), Remark D.3, and Condition 12. Using $|ab| \le \frac{ca^2}{2} + \frac{b^2}{2c}$ (any $c > 0$) on the cross term between the contracted error and (b),

$$2\xi_k \left|\left\langle \left(\mathbb{I} - \xi_k\overline{\mathbb{H}}_k\right) \mathcal{R}_{J_{k-1},J_k}^\top \Delta\widetilde{\omega}_{k-1},\ b_{J_k} \right\rangle\right| \le c\,\xi_k\|\Delta\widetilde{\omega}_{k-1}\|_2^2 + C\xi_k\|b_{J_k}\|_2^2 \le c\,\xi_k\|\Delta\widetilde{\omega}_{k-1}\|_2^2 + C\xi_k J_k^{-2q_\star},$$

where $\overline{\mathbb{H}}_k$ denotes the integrated population Hessian appearing in the contraction term, and c is chosen small relative to the contraction constant. The variance of the centered score contributes $\xi_k^2\,\mathbb{E}_{k-1}\|g_k - b_{J_k}\|_2^2 \le C\xi_k^2 J_k$, the variance of the $\frac{\xi_k}{B}\sum_{i=1}^{B}\Big[\nabla_\omega \mathcal{L}_{J_k}(\omega_{J_k,0}, W_{i,k}) + \varphi_{J_k}(W_{i,k})\Big]$-type remainder contributes $C\xi_k^2 J_k^{3-2s} \le C\xi_k^2 J_k$ since $s > 1$, and the cross terms at expansion times are bounded by $C\mathbf{1}_k(\|\Delta\widetilde{\omega}_{k-1}\|_2^2 + J_k^{-2s})$. So when $J_k = J_{k-1}$, there holds

$$\begin{aligned}
&\mathbb{E}_{k-1}\|\mathcal{R}_{J_{k-1},J_k}^\top \widetilde{\omega}_{k-1} - \frac{\xi_k}{B}\sum_{i=1}^{B}\nabla_\omega \mathcal{L}_{J_k}\left(\mathcal{R}_{J_{k-1},J_k}^\top \widetilde{\omega}_{k-1}, W_{i,k}\right) - \omega_{J_k,0}\|_2^2 \\
&\le \left(1 - C\xi_k + C\xi_k^2 J_k^3\right)\|\Delta\widetilde{\omega}_{k-1}\|_2^2 + C\xi_k^2 J_k + C\xi_k J_k^{-2q_\star},
\end{aligned}$$

and when $J_k > J_{k-1}$, we have that

$$\begin{aligned}
&\mathbb{E}_{k-1}\|\mathcal{R}_{J_{k-1},J_k}^\top \widetilde{\omega}_{k-1} - \frac{\xi_k}{B}\sum_{i=1}^{B}\nabla_\omega \mathcal{L}_{J_k}\left(\mathcal{R}_{J_{k-1},J_k}^\top \widetilde{\omega}_{k-1}, W_{i,k}\right) - \omega_{J_k,0}\|_2^2 \\
&\le C\left(1 - C\xi_k + C\xi_k^2 J_k^3 + C\xi_k J_k^{3-2s}\right)\|\Delta\widetilde{\omega}_{k-1}\|_2^2 + C\xi_k^2 J_k + C\xi_k J_k^{-2q_\star} + CJ_k^{-2s}.
\end{aligned}$$

Then use the fact that $\omega_{J_k,0} \in \widehat{\Omega}_k$ according to Condition 12, we have that

$$\mathbb{E}_{k-1}\|\Delta\widetilde{\omega}_k\|_2^2 \le \left(1 - C\xi_k + C\mathbf{1}_k\right)\|\Delta\widetilde{\omega}_{k-1}\|_2^2 + C\xi_k^2 J_k + C\xi_k J_k^{-2q_\star} + C\mathbf{1}_k J_k^{-2s}, \tag{D.11}$$

when $\alpha_\xi > 3\alpha_\dagger$ and $s > \frac{5}{2}$.

Condition 12 gives $\omega_{J_k,0} \in \widehat{\Omega}_k$ eventually almost surely. On that probability-one tail, nonexpansiveness of projection gives

$$\|\Delta\widetilde{\omega}_k\|_2^2 \le \left\| \mathcal{R}_{J_{k-1},J_k}^\top \widetilde{\omega}_{k-1} - \frac{\xi_k}{B}\sum_{i=1}^B \nabla_\omega \mathcal{L}_{J_k}\left(\mathcal{R}_{J_{k-1},J_k}^\top \widetilde{\omega}_{k-1}, W_{i,k}\right) - \omega_{J_k,0} \right\|_2^2$$

This leads to

$$\mathbb{E}_{k-1}\|\Delta\widetilde{\omega}_k\|_2^2 \le (1 - C\xi_k + C\mathbf{1}_k)\|\Delta\widetilde{\omega}_{k-1}\|_2^2 + C\xi_k^2 J_k + C\xi_k J_k^{-2q_\star} + C\mathbf{1}_k J_k^{-2s}.$$

The same blockwise Robbins–Siegmund argument as in Section C.4.2 yields

$$\|\Delta\widetilde{\omega}_k\|_2 = O\left(k^{\frac{1+\alpha_\dagger-2\alpha_\xi}{2}}\log^{\frac{r}{2}}(k) + k^{\frac{1-\alpha_\xi}{2}-q_\star\alpha_\dagger}\log^{\frac{r}{2}}(k) + k^{-\left(s-\frac{1}{2}\right)\alpha_\dagger}\log^{\frac{r}{2}}(k)\right), \qquad \text{a.s.}$$

We finally verify that under Condition 14 each exponent exceeds $2\alpha_\dagger$, so that the *projected* error satisfies $\|\Delta\widetilde{\omega}_k\|_2 = o(J_k^{-2})$ almost surely. First, $\frac{2\alpha_\xi-\alpha_\dagger-1}{2} > 2\alpha_\dagger$ iff $\alpha_\xi > \frac{1}{2} + \frac{5}{2}\alpha_\dagger$, which is implied by $\alpha_\xi > \frac{1}{2} + 3\alpha_\dagger$. Second, using $\alpha_\xi > \frac{1}{2} + 3\alpha_\dagger$ we get $\frac{1-\alpha_\xi}{2} < \frac{1}{4} - \frac{3}{2}\alpha_\dagger$, so $q_\star\alpha_\dagger - \frac{1-\alpha_\xi}{2} > \frac{1}{2} - \frac{1}{4} + \frac{3}{2}\alpha_\dagger > 2\alpha_\dagger$ whenever $q_\star\alpha_\dagger > \frac{1}{2}$. Third, $(s-\frac{1}{2})\alpha_\dagger > 2\alpha_\dagger$ iff $s > \frac{5}{2}$.

Given such result, we easily have that

$$\begin{aligned}
\|\widetilde{\omega}_k - \widehat{\omega}_k\|_2 &\le \|\widetilde{\omega}_k - \omega_{J_k,0}\|_2 + \|\widehat{\omega}_k - \omega_{J_k,0}\|_2 \\
&\le o(J_k^{-2}) + \frac{(1-\delta_\Omega)C_\Omega J_k^{-2}}{2} < \frac{C_\Omega J_k^{-2}}{2}, \qquad \text{a.s.}
\end{aligned}$$

This finally validates the projection inactivity.

D.4.2. Obtain the Refined Rate

Summary: This part decomposes the raw iterate error into a regular *component D_k and a* deterministic *sieve-expansion transient T_k,*

$$\Delta\widetilde{\omega}_k = D_k + T_k,$$

and builds the refined rates. In particular, we have that

$$\|D_k\|_2 = O\left(k^{-\frac{\alpha_\xi-\alpha_\dagger}{2}}\mathrm{PolyLog} + k^{-(2s-2)\alpha_\dagger}\right), \qquad \|T_k\|_2 \le Ck^{-s\alpha_\dagger}, \qquad \text{a.s.},$$

so that

$$\|\Delta\widetilde{\omega}_k\|_2 = O(k^{-\frac{\alpha_\xi-\alpha_\dagger}{2}} PolyLog + k^{-s\alpha_\dagger}), \qquad a.s.$$

In Section D.4.1, we have shown that the projection is inactive almost surely for k large. In this case, almost surely the following holds for k sufficiently large

$$\widetilde{\omega}_k = \mathcal{R}_{J_{k-1},J_k}^\top \widetilde{\omega}_{k-1} - \frac{\xi_k}{B}\sum_{i=1}^{B} \nabla_\omega \mathcal{L}_{J_k}\left(\mathcal{R}_{J_{k-1},J_k}^\top \widetilde{\omega}_{k-1}, W_{i,k}\right). \tag{D.12}$$

To simplify our following exposition, without loss of generality we assume that (D.12) holds true for all k; otherwise we can start from a random integer k_0, whose impacts will be exponentially decaying so are negligible. Define

$$\mathbb{M}_J = \mathbb{E}\begin{pmatrix} (\nabla_z F_0(Z_0))^2 XX^\top & \nabla_z F_0(Z_0) X\Psi_J(Z_0)^\top \\ \nabla_z F_0(Z_0)\Psi_J(Z_0)X^\top & \Psi_J(Z_0)\Psi_J(Z_0)^\top \end{pmatrix}.$$

At the sieve reference point,

$$\begin{aligned}
\nabla_{\theta\theta}\mathcal{L}_J(\omega_{J,0}) &= \mathbb{E}\big[\big\{(\nabla_z \mathbb{P}_J F_0(Z_0))^2 - (F_0 - \mathbb{P}_J F_0)(Z_0)\nabla_{zz}\mathbb{P}_J F_0(Z_0)\big\} XX^\top\big], \\
\nabla_{\theta\beta}\mathcal{L}_J(\omega_{J,0}) &= \mathbb{E}\big[\nabla_z \mathbb{P}_J F_0(Z_0) X\Psi_J(Z_0)^\top - (F_0 - \mathbb{P}_J F_0)(Z_0) X\nabla_z \Psi_J(Z_0)^\top\big].
\end{aligned}$$

By Condition 6 and Condition 10 and the fact that $s > 2$, we have that

$$\|\mathbb{M}_J - \nabla_{\omega\omega}\mathcal{L}_J(\omega_{J,0})\|_F \le CJ^{\frac{3}{2}-s},$$

where C is a constant independent of J. From the proof of the first step, let

$$e_k^- := \mathcal{R}_{J_{k-1},J_k}^\top \Delta\widetilde{\omega}_{k-1}, \qquad A_k := \big(\mathbb{I}_{p+J_k} - \xi_k \mathbb{M}_{J_k}\big)\mathcal{R}_{J_{k-1},J_k}^\top.$$

Then the unprojected recursion has the exact decomposition

$$\Delta\widetilde{\omega}_k = A_k \Delta\widetilde{\omega}_{k-1} - \xi_k \sum_{\ell=1}^{6} \zeta_{\ell,k} + \zeta_{7,k}, \tag{D.13}$$

where

$$
\begin{aligned}
\zeta_{1,k} &:= \{\nabla_{\omega\omega}\mathcal{L}_{J_k}(\omega_{J_k,0}) - \mathbb{M}_{J_k}\}e_k^-,\\
\zeta_{2,k} &:= \int_0^1 \Big[\nabla_{\omega\omega}\mathcal{L}_{J_k}\{\mathcal{R}_{J_{k-1},J_k}^\top(\omega_{J_{k-1},0} + \tau\Delta\widetilde{\omega}_{k-1})\} - \nabla_{\omega\omega}\mathcal{L}_{J_k}(\omega_{J_k,0})\Big] d\tau\, e_k^-,\\
\zeta_{3,k} &:= \frac{1}{B}\sum_{i=1}^B \int_0^1 \Big[\nabla_{\omega\omega}\mathcal{L}_{J_k}\{\mathcal{R}_{J_{k-1},J_k}^\top(\omega_{J_{k-1},0} + \tau\Delta\widetilde{\omega}_{k-1}), W_{i,k}\} - \nabla_{\omega\omega}\mathcal{L}_{J_k}\{\mathcal{R}_{J_{k-1},J_k}^\top(\omega_{J_{k-1},0} + \tau\Delta\widetilde{\omega}_{k-1})\}\Big] d\tau\, e_k^-\\
\zeta_{4,k} &:= \frac{1}{B}\sum_{i=1}^B \Big[\nabla_\omega\mathcal{L}_{J_k}(\mathcal{R}_{J_{k-1},J_k}^\top\omega_{J_{k-1},0}, W_{i,k}) - \nabla_\omega\mathcal{L}_{J_k}(\omega_{J_k,0}, W_{i,k})\Big],\\
\zeta_{5,k} &:= \frac{1}{B}\sum_{i=1}^B \Big[\nabla_\omega\mathcal{L}_{J_k}(\omega_{J_k,0}, W_{i,k}) + \varphi_{J_k}(W_{i,k})\Big],\\
\zeta_{6,k} &:= -\frac{1}{B}\sum_{i=1}^B \varphi_{J_k}(W_{i,k}),\\
\zeta_{7,k} &:= \mathcal{R}_{J_{k-1},J_k}^\top\omega_{J_{k-1},0} - \omega_{J_k,0}.
\end{aligned}
$$

Thus (D.13) yields

$$
\Delta\widetilde{\omega}_k = \left(\prod_{j=1}^k A_j\right)\Delta\widetilde{\omega}_0 - \sum_{j=1}^k \xi_j \left(\prod_{m=j+1}^k A_m\right)\sum_{l=1}^6 \zeta_{l,j} + \sum_{j=1}^k \left(\prod_{m=j+1}^k A_m\right)\zeta_{7,j},
$$

where following the previous convention, $\prod_{j=1}^k A_j = A_k A_{k-1}\cdots A_1$, $\prod_{m=j+1}^k A_m = A_k A_{k-1}\cdots A_{j+1}$ if $j \le k-1$ and $\prod_{m=j+1}^k A_m = \mathbb{I}_{p+J_k}$ if $j = k$.

Before analyzing these terms, we now formally introduce the transient/regular decomposition announced in the summary. Decompose the score-difference term at expansion times into its (deterministic) mean and its martingale part: $\zeta_{4,k} = \mathbb{E}\zeta_{4,k} + \zeta_{4,k}^\circ$, noting that $\zeta_{4,k}$ is evaluated at the deterministic points $\mathcal{R}_{J_{k-1},J_k}^\top\omega_{J_{k-1},0}$ and $\omega_{J_k,0}$, so $\mathbb{E}_{k-1}\zeta_{4,k} = \mathbb{E}\zeta_{4,k}$ is nonrandom. Define the deterministic *sieve-expansion transient*

$$
T_k := \sum_{j=1}^k \left(\prod_{m=j+1}^k A_m\right)\tau_j, \qquad \tau_j := \zeta_{7,j} - \xi_j\,\mathbb{E}\zeta_{4,j}, \qquad D_k := \Delta\widetilde{\omega}_k - T_k, \tag{D.14}
$$

so that $T_k = A_k T_{k-1} + \tau_k$ with $T_0 = 0$, and

$$
D_k = A_k D_{k-1} - \xi_k\left(\zeta_{1,k} + \zeta_{2,k} + \zeta_{3,k} + \zeta_{4,k}^\circ + \zeta_{5,k} + \zeta_{6,k}\right), \qquad D_0 = \Delta\widetilde{\omega}_0. \tag{D.15}
$$

Note that $\tau_j \ne 0$ only at expansion times, and by Condition 12 (boundedness of the population Hessian on the local basin, which contains the segment between $\mathcal{R}_{J_{j-1},J_j}^\top \omega_{J_{j-1},0}$ and $\omega_{J_j,0}$ since $s > 2$), a mean-value expansion gives $\|\mathbb{E}\zeta_{4,j}\|_2 \le C\|\mathcal{R}_{J_{j-1},J_j}^\top \omega_{J_{j-1},0} - \omega_{J_j,0}\|_2$. Hence, by Lemma D.3 and Remark D.3,

$$\|\tau_j\|_2 \le C\mathbf{1}_j \sum_{m=J_{j-1}+1}^{J_j} \delta_m \le C\mathbf{1}_j J_j^{-s}. \tag{D.16}$$

The next lemma collects the properties of T_k that will be used below and in Section D.4.3.

Lemma D.5 (Checkpoint transient)**.** *Let the conditions of Lemma 5.3 hold. Then there are constants $c, C > 0$ such that:*

(i) $\left\|\prod_{m=j+1}^{k} A_m\right\|_{\mathrm{op}} \le C\exp\left(-c\left(k^{1-\alpha_\xi} - j^{1-\alpha_\xi}\right)\right)$ *for all $j \le k$;*

(ii) $\|T_k\|_2 \le Ck^{-s\alpha_\dagger}$ *for all k, and* $\|T_k\|_2^2 \le C\sum_{m\le J_k}\exp\left(-2c\left(k^{1-\alpha_\xi} - \mathcal{J}(m)^{1-\alpha_\xi}\right)\right)\delta_m^2$*;*

(iii) for every $\nu \ge 0$ there is $C(\nu)$ such that for all N,

$$\sum_{k=1}^{N} k^\nu\|T_k\|_2 \le C(\nu)\left(N^{\nu+\alpha_\xi+\left(\frac{1}{2}-s\right)\alpha_\dagger} + \log N\right), \quad \sum_{k=1}^{N} k^\nu\|T_k\|_2^2 \le C(\nu)\left(N^{\nu+\alpha_\xi+(1-2s)\alpha_\dagger} + \log N\right);$$

(iv) for every fixed $r > 0$ and all m sufficiently large, $\|T_{\mathcal{J}(m)-1}\|_2 \le \mathcal{J}(m)^{-r}$; that is, the transient is superpolynomially small at the end of each constant-dimension block.

Proof. (i) is obviously from our previous proofs. For (ii), by (D.16) and (i), and merging Remark D.3 into constants,

$$\|T_k\|_2 \le C\sum_{m\le J_k}\exp\left(-c\left(k^{1-\alpha_\xi} - \mathcal{J}(m)^{1-\alpha_\xi}\right)\right)\delta_m =: C\sum_{m\le J_k} e_m(k)\,\delta_m.$$

For consecutive m, the gap between expansion times satisfies $\mathcal{J}(m+1) - \mathcal{J}(m) \asymp \mathcal{J}(m)^{1-\alpha_\dagger}$, so

$$\frac{e_m(k)}{e_{m+1}(k)} = \exp\left(-c\left(\mathcal{J}(m+1)^{1-\alpha_\xi} - \mathcal{J}(m)^{1-\alpha_\xi}\right)\right) \le \exp\left(-c'\mathcal{J}(m)^{1-\alpha_\xi-\alpha_\dagger}\right) \le \frac{1}{2}$$

for all $m \ge m_0$, since $1 - \alpha_\xi - \alpha_\dagger > 0$. Hence the weights $e_m(k)$ decay super-geometrically as m decreases below J_k, $\sum_{m\le J_k} e_m(k) \le C$, and $\sum_{m\le J_k} e_m(k)\delta_m \le C\max\{\delta_m : J_k/2 \le m \le J_k\} + C\exp(-c''k^{1-\alpha_\xi}) \le CJ_k^{-s} \le Ck^{-s\alpha_\dagger}$ by Lemma D.3(i). The second claim in (ii) follows directly from Cauchy–Schwarz: $\|T_k\|_2^2 \le (\sum_m e_m(k))(\sum_m e_m(k)\delta_m^2) \le C\sum_m e_m(k)\delta_m^2$.

Relabelling the positive decay constant absorbs the harmless distinction between $e_m(k)$ and $e_m(k)^2$. For (iii), interchanging the order of summation and using, for $\varrho := \alpha_\xi < 1$,

$$\sum_{k\geq j} k^\nu \exp\left(-c\left(k^{1-\varrho} - j^{1-\varrho}\right)\right) \leq C j^{\nu+\varrho},$$

which follows by splitting $k \leq 2j$ (where $k^{1-\varrho} - j^{1-\varrho} \geq (1-\varrho)(k-j)(2j)^{-\varrho}$, giving a geometric sum of size Cj^ϱ with $k^\nu \leq (2j)^\nu$) from $k > 2j$ (where $k^{1-\varrho} - j^{1-\varrho} \geq (1-2^{-(1-\varrho)})k^{1-\varrho}$, giving an $O(1)$ tail), we obtain

$$\sum_{k=1}^{N} k^\nu \|T_k\|_2 \leq C \sum_{m\leq J_N} \delta_m \,\mathcal{J}(m)^{\nu+\alpha_\xi} \leq C \sum_{m\leq J_N} m^{\frac{\nu+\alpha_\xi}{\alpha_\dagger}} \delta_m,$$

and Lemma D.3(iii) with $J_N \leq CN^{\alpha_\dagger}$ gives the first display; the second is identical with δ_m^2 in place of δ_m. (iv) The most recent expansion before time $\mathcal{J}(m) - 1$ occurs at $\mathcal{J}(m-1)$, and by (i)–(ii),

$$\|T_{\mathcal{J}(m)-1}\|_2 \leq C \exp\left(-c\left((\mathcal{J}(m)-1)^{1-\alpha_\xi} - \mathcal{J}(m-1)^{1-\alpha_\xi}\right)\right) \leq C \exp\left(-c' \mathcal{J}(m)^{1-\alpha_\xi-\alpha_\dagger}\right),$$

which is superpolynomially small since $1 - \alpha_\xi - \alpha_\dagger > 0$. $\square$

Lemma D.5 provides the bound for $\|T_k\|_2$. It remains to bound $\|D_k\|_2$. In view of (D.15), the regular component satisfies $D_k = (\prod_{j=1}^k A_j)\Delta\widetilde{\omega}_0 - \sum_{l\in\{1,2,3,5,6\}} Q_{l,k} - Q^\circ_{4,k}$, where $Q_{l,k} = A_k Q_{l,k-1} + \xi_k \zeta_{l,k}$. and $Q^\circ_{4,k}$ is defined with $\zeta^\circ_{4,k}$ in place of $\zeta_{4,k}$. For $(\prod_{j=1}^k A_j)\Delta\widetilde{\omega}_0$, similar to the proof in Section C.4.3, when $J_{k-1} = J_k$, $\mathcal{R}_{J_{k-1},J_k} = \mathbb{I}_{p+J_k}$, so $A_k = \mathbb{I}_{p+J_k} - \xi_k \mathbb{M}_{J_k}$ and $\|A_k\|_{\text{op}} \leq (1 - C\xi_k)$; on the other side, when $J_k > J_{k-1}$, we have that $\|A_k\|_{\text{op}} \leq C$. Together, we have that

$$\begin{aligned}\left\|\left(\prod_{j=1}^{k} A_j\right)\Delta\widetilde{\omega}_0\right\|_2 &\leq C^{J_k} \prod_{j=1}^{k} (1 - C\xi_j)\,\|\Delta\widetilde{\omega}_0\|_2 \leq C \exp\left(-Ck^{1-\alpha_\xi} + Ck^{\alpha_\dagger}\right)\\ &\leq C \exp\left(-Ck^{1-\alpha_\xi}\right)\end{aligned}$$

when $1 - \alpha_\xi > \alpha_\dagger$. It then remains to look at $Q_{l,k}$ one by one.

We first show that $Q_{1,k}$ does not determine the convergence rate of $\Delta\widetilde{\omega}_k$ under the given conditions. In particular, let $\|\Delta\widetilde{\omega}_k\|_2 = O(k^{-\alpha_\omega})$ a.s. holds, where α_ω is some preliminary positive rate. Such rate exists according to our first-step proof. Then

$$
\begin{aligned}
\|Q_{1,k}\|_2 &\le \|A_k\|_{\mathrm{op}}\|Q_{1,k-1}\|_2 + C\xi_k J_k^{\frac{3}{2}-s}\|\Delta\widetilde{\omega}_{k-1}\|_2 \\
&\le \left(1 - Ck^{-\alpha_\xi} + C\mathbf{1}_k\right)\|Q_{1,k-1}\|_2 + Ck^{-\alpha_\xi-\alpha_\omega+(\frac{3}{2}-s)\alpha_\dagger}, \qquad \text{a.s.}
\end{aligned}
$$

where the exponent $(\frac{3}{2}-s)\alpha_\dagger$ comes from the fact that $\|\mathbb{M}_J - \nabla_{\omega\omega}\mathcal{L}_J(\omega_{J,0})\|_F \le CJ^{\frac{3}{2}-s}$. Without loss of generality, we can assume the above holds for all k (otherwise we can consider sufficiently large k only). Iterations gives

$$
\begin{aligned}
\|Q_{1,k}\|_2 \le{}& C\exp\left(-C\sum_{j=1}^{k} j^{-\alpha_\xi} + Ck^{\alpha_\dagger}\right)\|Q_{1,0}\|_2 \\
&+ C\sum_{j=1}^{k}\exp\left(-C\sum_{m=j+1}^{k} m^{-\alpha_\xi} + C(k^{\alpha_\dagger} - j^{\alpha_\dagger} + 1)\right) j^{-\alpha_\xi-\alpha_\omega+(\frac{3}{2}-s)\alpha_\dagger} \\
\le{}& C\exp\left(-Ck^{1-\alpha_\xi} + Ck^{\alpha_\dagger}\right)\|Q_{1,0}\|_2 \\
&+ C\exp\left(-C^*k^{1-\alpha_\xi} + Ck^{\alpha_\dagger}\right)\sum_{j=1}^{k}\exp\left(C^*j^{1-\alpha_\xi} - Cj^{\alpha_\dagger}\right) j^{-\alpha_\xi-\alpha_\omega+(\frac{3}{2}-s)\alpha_\dagger}.
\end{aligned}
$$

So when $1-\alpha_\xi > \alpha_\dagger$, by Lemma F.5, there holds $\|Q_{1,k}\|_2 = O_{\text{a.s.}}(k^{-\alpha_\omega+(\frac{3}{2}-s)\alpha_\dagger})$. When $s > \frac{3}{2}$, we have that $\|Q_{1,k}\|_2 = o(k^{-\alpha_\omega})$ a.s., so it does not affect the bound of the convergence rate of $\Delta\widetilde{\omega}_k$.

For $Q_{2,k}$, note that the first Hessian argument in $\zeta_{2,k}$ differs from $\omega_{J_k,0}$ by $\zeta_{7,k} + \tau e_k^-$. Hence Lemma D.1, bounded re-embedding, and Young's inequality give

$$
\begin{aligned}
\|\zeta_{2,k}\|_2 &\le CJ_k^2\left\{\|\Delta\widetilde{\omega}_{k-1}\|_2^2 + \mathbf{1}_k J_k^{-s}\|\Delta\widetilde{\omega}_{k-1}\|_2\right\} \\
&\le CJ_k^2\|\Delta\widetilde{\omega}_{k-1}\|_2^2 + C\mathbf{1}_k J_k^{2-2s}.
\end{aligned} \tag{D.17}
$$

Thus

$$
\|Q_{2,k}\|_2 \le (1 - Ck^{-\alpha_\xi} + C\mathbf{1}_k)\|Q_{2,k-1}\|_2 + Ck^{-\alpha_\xi-2\alpha_\omega+2\alpha_\dagger} + C\mathbf{1}_k k^{-\alpha_\xi-(2s-2)\alpha_\dagger}, \qquad \text{a.s.}
$$

Lemma D.5(i) therefore yields

$$
\|Q_{2,k}\|_2 = O\left(k^{-2\alpha_\omega+2\alpha_\dagger} + k^{-(2s-2)\alpha_\dagger}\right), \qquad \text{a.s.}
$$

When $\alpha_\omega > 2\alpha_\dagger$, which holds given the proof of first part, the first term does not affect the convergence rate.

For $Q_{3,k}$, we obviously have that

$$\mathbb{E}_{k-1}\|Q_{3,k}\|_2^2 \leq (1 - C\xi_k + C\mathbf{1}_k)\|Q_{3,k-1}\|_2^2 + C\xi_k^2 J_k^3 \|\Delta\widetilde{\omega}_{k-1}\|_2^2.$$

This leads to that

$$\|Q_{3,k}\|_2 = O\left(k^{\frac{1+3\alpha_\dagger - 2\alpha_\xi}{2} - \alpha_\omega} \log^{\frac{r}{2}}(k)\right), \qquad \text{a.s.}$$

When $\alpha_\xi > \frac{1+3\alpha_\dagger}{2}$, we have that $Q_{3,k}$ does not affect the convergence rate of $\Delta\widetilde{\omega}_{k-1}$.

For $Q_{4,k}^\circ$, recall that the deterministic mean $\mathbb{E}\zeta_{4,k}$ has been moved into the transient T_k via (D.14), so only the martingale part $\zeta_{4,k}^\circ$ remains, and $\zeta_{4,k}^\circ = 0$ for any k with $J_k = J_{k-1}$. By the Hessian bound used for $\zeta_{4,k}$ in Section D.4.1 and Lemma D.3(i),

$$\mathbb{E}_{k-1}\|\zeta_{4,k}^\circ\|_2^2 \leq \mathbb{E}\|\zeta_{4,k}\|_2^2 \leq CJ_k^3 \|\mathcal{R}_{J_{k-1},J_k}^\top \omega_{J_{k-1},0} - \omega_{J_k,0}\|_2^2 \leq C\mathbf{1}_k J_k^{3-2s},$$

so that

$$\mathbb{E}_{k-1}\|Q_{4,k}^\circ\|_2^2 \leq \left(1 - Ck^{-\alpha_\xi} + C\mathbf{1}_k\right)\|Q_{4,k-1}^\circ\|_2^2 + C\mathbf{1}_k k^{-2\alpha_\xi + (3-2s)\alpha_\dagger}.$$

Multiplying by the weight $k^{\alpha_\xi - \alpha_\dagger}\log^{-r}(k)$, the forcing term is summable:

$$\sum_k \mathbf{1}_k k^{\alpha_\xi - \alpha_\dagger - 2\alpha_\xi + (3-2s)\alpha_\dagger} \log^{-r}(k) \asymp \sum_m \mathcal{J}(m)^{-\alpha_\xi + (2-2s)\alpha_\dagger} \log^{-r}(\mathcal{J}(m)) < \infty,$$

since $-\alpha_\xi + (2-2s)\alpha_\dagger < -\alpha_\dagger$ holds trivially for $s > \frac{3}{2}$. Following the same argument used for $Q_{4,k}$-type terms previously (supermartingale convergence along constant-dimension blocks), we obtain

$$\|Q_{4,k}^\circ\|_2 = O\left(k^{-\frac{\alpha_\xi - \alpha_\dagger}{2}} \log^{\frac{r}{2}}(k)\right), \qquad \text{a.s.}$$

For $Q_{5,k}$, we first note that since $\mathbb{E}[\varphi_J(W)] = 0$, the mean of $\zeta_{5,k}$ is precisely the mean sieve score of Lemma D.2:

$$\mathbb{E}\zeta_{5,k} = \mathbb{E}\left[\nabla_\omega \mathcal{L}_{J_k}(\omega_{J_k,0}, W)\right] = b_{J_k}, \qquad \|b_{J_k}\|_2 \leq CJ_k^{-q_\star} = CJ_k^{-s-\min(s_\mu, s-1)}.$$

Moreover, $\mathbb{E}\|\zeta_{5,k}\|_2^2 \leq CJ^{3-2s}\mathbb{E}[(1+|Y|^2)(1+\|X\|_2^2)] \leq CJ^{3-2s}$.

Given the above two bounds, we have that

$$Q_{5,k} = \xi_k \mathbb{E}_{k-1}\zeta_{5,k} + \xi_k(\zeta_{5,k} - \mathbb{E}_{k-1}\zeta_{5,k}) + A_k Q_{5,k-1}$$

So we only need to bound two sequences $Q_{5,k,1}$ and $Q_{5,k,2}$ defined by $Q_{5,k,1} = \xi_k \mathbb{E}_{k-1}\zeta_{5,k} + A_k Q_{5,k-1,1}$ and $Q_{5,k,2} = \xi_k(\zeta_{5,k} - \mathbb{E}_{k-1}\zeta_{5,k}) + A_k Q_{5,k-1,2}$. Obviously the first term is of order $O_{\text{a.s.}}(k^{-(s+\min(s_\mu,\,s-1))\alpha_\dagger})$ by Lemma F.5, while the second term, by the previous proof method for $Q_{4,k}$, is of order $O_{\text{a.s.}}(k^{-\frac{2\alpha_\xi+(2s-3)\alpha_\dagger-1}{2}}\text{PolyLog})$. This shows that

$$\|Q_{5,k}\|_2 = O\left(k^{-(s+\min(s_\mu,\,s-1))\alpha_\dagger} + k^{-\frac{2\alpha_\xi+(2s-3)\alpha_\dagger-1}{2}}\text{PolyLog}\right), \qquad \text{a.s.}$$

For $Q_{6,k}$, applying the proof of Theorem C.1 to εX together shows that

$$\|Q_{6,k}\|_2 = O\left(k^{-\frac{\alpha_\xi-\alpha_\dagger}{2}}\log^{\frac{1}{2}}(k)\right).$$

Finally, the terms driven by $\zeta_{7,j}$ and by $\mathbb{E}\zeta_{4,j}$ constitute exactly the transient T_k of (D.14), and Lemma D.5(ii) leads to $\|T_k\|_2 \le Ck^{-s\alpha_\dagger}$.

Consequently, we have that

$$\|D_k\|_2 = O\left(k^{-\frac{\alpha_\xi-\alpha_\dagger}{2}}\text{PolyLog} + k^{-(2s-2)\alpha_\dagger}\right), \quad \|\Delta\widetilde{\omega}_k\|_2 = O\left(k^{-\frac{\alpha_\xi-\alpha_\dagger}{2}}\text{PolyLog} + k^{-s\alpha_\dagger}\right), \quad \text{a.s.} \tag{D.18}$$

Here the component $k^{-(2s-2)\alpha_\dagger}$ of $\|D_k\|_2$ collects the contributions of $Q_{1,k}$ and $Q_{2,k}$ driven by the transient part of $\Delta\widetilde{\omega}$ (namely $O(k^{-(2s-\frac{3}{2})\alpha_\dagger})$ and $O(k^{-(2s-2)\alpha_\dagger})$ respectively), while the $Q_{5,k}$ mean component $O(k^{-q_\star\alpha_\dagger})$ is absorbed into the leading term because $q_\star\alpha_\dagger > \frac{1}{2} > \frac{\alpha_\xi-\alpha_\dagger}{2}$, and the $Q_{5,k}$ variance component is absorbed because $\frac{2\alpha_\xi+(2s-3)\alpha_\dagger-1}{2} \ge \frac{\alpha_\xi-\alpha_\dagger}{2}$ iff $\alpha_\xi + (2s-2)\alpha_\dagger \ge 1$, which follows from $\alpha_\xi > 1+(4-2s)\alpha_\dagger$. The second display in (D.18) follows from the first, Lemma D.5(ii), and $(2s-2)\alpha_\dagger \ge s\alpha_\dagger$ for $s \ge 2$. This gives the refined convergence rate of $\|\Delta\widetilde{\omega}_k\|_2$

D.4.3. PR Averaging and the Asymptotic Linear Representation

Summary: This part will prove the main results without conditioning.

From the proof of Section D.4.2, we have the decomposition $\Delta\widetilde{\omega}_k = D_k + T_k$ from (D.14)–(D.15). We first dispose of the transient component of the PR average directly — this is the checkpoint-aware step. By Lemma D.5(iii) with $\nu = 0$, and $\|\mathcal{R}_{J_k,J_N}\|_{\text{op}} \le C$,

$$\left\|\frac{1}{N}\sum_{k=1}^{N}\mathcal{R}_{J_k,J_N}^\top T_k\right\|_2 \le \frac{C}{N}\sum_{k=1}^{N}\|T_k\|_2 \le \frac{C}{N}\left(N^{\alpha_\xi+(\frac{1}{2}-s)\alpha_\dagger} + \log N\right) = o\left(N^{-\frac{1}{2}}\right), \tag{D.19}$$

where the final step holds because $\alpha_\xi < \frac{1}{2} + (s - \frac{1}{2})\alpha_\dagger$ under Condition 14.

It remains to analyze the regular component. For $\mathcal{J}(m) + 1 \leq k \leq \mathcal{J}(m+1) - 1$, sieve dimension holds constant, so $A_k = \mathbb{I}_{p+m} - \xi_k \mathbb{M}_m$, $\tau_k = 0$, and $\zeta^\circ_{4,k} = 0$, whence by (D.15),

$$D_k = \left(\mathbb{I}_{p+m} - \xi_k \mathbb{M}_m\right) D_{k-1} - \xi_k \sum_{l \in \{1,2,3,5,6\}} \zeta_{l,k}, \quad \mathcal{J}(m) + 1 \leq k \leq \mathcal{J}(m+1) - 1.$$

This implies that

$$D_{k-1} = \xi_k^{-1} \mathbb{M}_m^{-1} \left(D_{k-1} - D_k\right) - \mathbb{M}_m^{-1} \sum_{l \in \{1,2,3,5,6\}} \zeta_{l,k}, \ \mathcal{J}(m) + 1 \leq k \leq \mathcal{J}(m+1) - 1$$

and

$$\begin{aligned}
&\sum_{k=\mathcal{J}(m)}^{\mathcal{J}(m+1)-2} D_k = \sum_{k=\mathcal{J}(m)+1}^{\mathcal{J}(m+1)-1} D_{k-1} \\
&= \mathbb{M}_m^{-1} \left[\sum_{k=\mathcal{J}(m)+1}^{\mathcal{J}(m+1)-2} \left(\xi_{k+1}^{-1} - \xi_k^{-1}\right) D_k + \xi_{\mathcal{J}(m)+1}^{-1} D_{\mathcal{J}(m)} - \xi_{\mathcal{J}(m+1)-1}^{-1} D_{\mathcal{J}(m+1)-1} \right] \\
&- \mathbb{M}_m^{-1} \sum_{k=\mathcal{J}(m)+1}^{\mathcal{J}(m+1)-1} \sum_{l \in \{1,2,3,5,6\}} \zeta_{l,k}.
\end{aligned}$$

Now we look at $\overline{\overline{\omega}}_N = \frac{1}{N} \sum_{k=1}^N \mathcal{R}_{J_k,J_N}^\top \widetilde{\omega}_k$. Using $\Delta \widetilde{\omega}_k = D_k + T_k$, (D.19), and the same block decomposition as before applied to D_k,

$$\begin{aligned}
\Delta\overline{\overline{\omega}}_N &= \frac{1}{N}\sum_{m=J_0}^{J_N-1}\mathcal{R}_{m,J_N}^{\top}\sum_{k=\mathcal{J}(m)}^{\mathcal{J}(m+1)-2}D_k+\frac{1}{N}\sum_{k=\mathcal{J}(J_N)}^{N}D_k+\frac{1}{N}\sum_{m=J_0}^{J_N-1}\mathcal{R}_{m,J_N}^{\top}D_{\mathcal{J}(m+1)-1}\\
&\quad+\frac{1}{N}\sum_{m=J_0}^{J_N-1}\sum_{k=\mathcal{J}(m)}^{\mathcal{J}(m+1)-1}\left(\mathcal{R}_{m,J_N}^{\top}\omega_{m,0}-\omega_{J_N,0}\right)+o\left(N^{-\frac{1}{2}}\right)\\
&=\frac{1}{N}\sum_{m=J_0}^{J_N-1}\mathcal{R}_{m,J_N}^{\top}\mathbb{M}_m^{-1}\sum_{k=\mathcal{J}(m)+1}^{\mathcal{J}(m+1)-2}\left(\xi_{k+1}^{-1}-\xi_k^{-1}\right)D_k+\frac{1}{N}\mathbb{M}_{J_N}^{-1}\sum_{k=\mathcal{J}(J_N)+1}^{N-1}\left(\xi_{k+1}^{-1}-\xi_k^{-1}\right)D_k\\
&\quad+\frac{1}{N}\sum_{m=J_0}^{J_N-1}\mathcal{R}_{m,J_N}^{\top}\mathbb{M}_m^{-1}\left(\xi_{\mathcal{J}(m)+1}^{-1}D_{\mathcal{J}(m)}-\xi_{\mathcal{J}(m+1)-1}^{-1}D_{\mathcal{J}(m+1)-1}\right)\\
&\quad+\frac{1}{N}\mathbb{M}_{J_N}^{-1}\left(\xi_{\mathcal{J}(J_N)+1}^{-1}D_{\mathcal{J}(J_N)}-\xi_N^{-1}D_N\right)\\
&\quad-\frac{1}{N}\sum_{m=J_0}^{J_N-1}\mathcal{R}_{m,J_N}^{\top}\mathbb{M}_m^{-1}\sum_{k=\mathcal{J}(m)+1}^{\mathcal{J}(m+1)-1}\sum_{l\in\{1,2,3,5,6\}}\zeta_{l,k}-\frac{1}{N}\mathbb{M}_{J_N}^{-1}\sum_{k=\mathcal{J}(J_N)+1}^{N}\sum_{l\in\{1,2,3,5,6\}}\zeta_{l,k}\\
&\quad+\frac{1}{N}\sum_{m=J_0}^{J_N-1}\mathcal{R}_{m,J_N}^{\top}D_{\mathcal{J}(m+1)-1}+\frac{1}{N}\sum_{m=J_0}^{J_N-1}\sum_{k=\mathcal{J}(m)}^{\mathcal{J}(m+1)-1}\left(\mathcal{R}_{m,J_N}^{\top}\omega_{m,0}-\omega_{J_N,0}\right)+o\left(N^{-\frac{1}{2}}\right).
\end{aligned}$$

We analyze the above equation term by term, using the refined rate (D.18) for D_k; note that each bound below is stated for the two components of (D.18) separately. Due to the boundedness of eigenvalues of $\mathbb{M}_m$ and $\|\mathcal{R}_{m,J_N}\|_{\mathrm{op}}$, we have that

$$\begin{aligned}
&\left\|\frac{1}{N}\sum_{m=J_0}^{J_N-1}\mathcal{R}_{m,J_N}^{\top}\mathbb{M}_m^{-1}\sum_{k=\mathcal{J}(m)+1}^{\mathcal{J}(m+1)-2}\left(\xi_{k+1}^{-1}-\xi_k^{-1}\right)D_k+\frac{1}{N}\mathbb{M}_{J_N}^{-1}\sum_{k=\mathcal{J}(J_N)+1}^{N-1}\left(\xi_{k+1}^{-1}-\xi_k^{-1}\right)D_k\right\|_2\\
&\leq\frac{1}{N}\sum_{k=1}^{N}\left(\xi_{k+1}^{-1}-\xi_k^{-1}\right)\|D_k\|_2\leq\frac{1}{N}\sum_{k=1}^{N}\left(k^{\alpha_\xi-1-\frac{\alpha_\xi-\alpha_\dagger}{2}}\mathrm{PolyLog}+k^{\alpha_\xi-1-(2s-2)\alpha_\dagger}\right)\\
&\leq CN^{\frac{\alpha_\xi+\alpha_\dagger}{2}-1}\mathrm{PolyLog}+CN^{\alpha_\xi-(2s-2)\alpha_\dagger-1}+CN^{-1}\log(N),\qquad\text{a.s.}
\end{aligned}$$

$$\begin{aligned}
&\left\|\frac{1}{N}\sum_{m=J_0}^{J_N-1}\mathcal{R}_{m,J_N}^{\top}\mathbb{M}_m^{-1}\left(\xi_{\mathcal{J}(m)+1}^{-1}D_{\mathcal{J}(m)}-\xi_{\mathcal{J}(m+1)-1}^{-1}D_{\mathcal{J}(m+1)-1}\right)+\frac{1}{N}\mathbb{M}_{J_N}^{-1}\left(\xi_{\mathcal{J}(J_N)+1}^{-1}D_{\mathcal{J}(J_N)}-\xi_N^{-1}D_N\right)\right\|_2\\
&\leq\frac{C}{N}\left(\sum_{m=J_0}^{J_N}\left(m^{\frac{\alpha_\xi}{\alpha_\dagger}-\frac{\alpha_\xi-\alpha_\dagger}{2\alpha_\dagger}}\mathrm{PolyLog}+m^{\frac{\alpha_\xi-(2s-2)\alpha_\dagger}{\alpha_\dagger}}\right)+N^{\alpha_\xi-\frac{\alpha_\xi-\alpha_\dagger}{2}}\mathrm{PolyLog}+N^{\alpha_\xi-(2s-2)\alpha_\dagger}\right)\\
&\leq CN^{\frac{\alpha_\xi+3\alpha_\dagger}{2}-1}\mathrm{PolyLog}+CN^{\alpha_\xi+\alpha_\dagger-(2s-2)\alpha_\dagger-1}+CN^{-1}\log(N),\qquad\text{a.s.}
\end{aligned}$$

$$\left\| \frac{1}{N} \sum_{m=J_0}^{J_N-1} \mathcal{R}_{m,J_N}^\top D_{\mathcal{J}(m+1)-1} \right\|_2 \le \frac{C}{N} \sum_{m=J_0}^{J_N} \left(m^{-\frac{\alpha_\xi - \alpha_\dagger}{2\alpha_\dagger}} \text{PolyLog} + m^{-\frac{(2s-2)\alpha_\dagger}{\alpha_\dagger}} \right) \le \frac{C}{N} \left(N^{\frac{3\alpha_\dagger - \alpha_\xi}{2}} \text{PolyLog} + \log N \right),$$

Note that the above terms are all $o(N^{-\frac{1}{2}})$ a.s.: the leading components require $\alpha_\xi + 3\alpha_\dagger < 1$, exactly as before, while the components generated by the secondary rate $k^{-(2s-2)\alpha_\dagger}$ in (D.18) are $o(N^{-\frac{1}{2}})$ whenever $\alpha_\xi < \frac{1}{2} + (2s-3)\alpha_\dagger$, which is implied by $\alpha_\xi < \frac{1}{2} + (s - \frac{1}{2})\alpha_\dagger$ since $2s - 3 \ge s - \frac{1}{2}$ for $s \ge \frac{5}{2}$.

It only remains to look at

$$\frac{1}{N} \sum_{m=J_0}^{J_N-1} \mathcal{R}_{m,J_N}^\top \mathbb{M}_m^{-1} \sum_{k=\mathcal{J}(m)+1}^{\mathcal{J}(m+1)-1} \sum_{l \in \{1,2,3,5,6\}} \zeta_{l,k} + \frac{1}{N} \mathbb{M}_{J_N}^{-1} \sum_{k=\mathcal{J}(J_N)+1}^{N} \sum_{l \in \{1,2,3,5,6\}} \zeta_{l,k}.$$

Similar to the proof in Section D.4.2, we look at each l separately. Note that $\zeta_{1,k}$, $\zeta_{2,k}$ and $\zeta_{3,k}$ depend on the *full* iterate error $\Delta\widetilde{\omega}_{k-1} = D_{k-1} + T_{k-1}$, so we bound their sums using both components: the refined rate (D.18) for D, and the checkpoint-aware sums of Lemma D.5(iii) for T.

For $l = 1$, obviously we have that $\|\zeta_{1,k}\|_2 \le J_k^{\frac{3}{2}-s} \|\Delta\widetilde{\omega}_{k-1}\|_2 \le J_k^{\frac{3}{2}-s} (\|D_{k-1}\|_2 + \|T_{k-1}\|_2)$, so

$$\begin{aligned}
&\left\| \frac{1}{N} \sum_{m=J_0}^{J_N-1} \mathcal{R}_{m,J_N}^\top \mathbb{M}_m^{-1} \sum_{k=\mathcal{J}(m)+1}^{\mathcal{J}(m+1)-1} \zeta_{1,k} + \frac{1}{N} \mathbb{M}_{J_N}^{-1} \sum_{k=\mathcal{J}(J_N)+1}^{N} \zeta_{1,k} \right\|_2 \\
&\le \frac{C}{N} \sum_{k=1}^{N} \|\zeta_{1,k}\|_2 \le \frac{C}{N} \sum_{k=1}^{N} k^{-\frac{\alpha_\xi - \alpha_\dagger}{2} + (\frac{3}{2}-s)\alpha_\dagger} \text{PolyLog} + \frac{C}{N} \sum_{k=1}^{N} k^{(\frac{3}{2}-s)\alpha_\dagger} \left(k^{-(2s-2)\alpha_\dagger} + \|T_{k-1}\|_2 \right) \\
&\le C N^{-\frac{\alpha_\xi - \alpha_\dagger}{2} + (\frac{3}{2}-s)\alpha_\dagger} \text{PolyLog} + C N^{(\frac{7}{2}-3s)\alpha_\dagger} + \frac{C}{N} \left(N^{\alpha_\xi + (2-2s)\alpha_\dagger} + \log N \right), \qquad \text{a.s.,}
\end{aligned}$$

where the sum involving $\|T_{k-1}\|_2$ is bounded by Lemma D.5(iii) (the factor $k^{(\frac{3}{2}-s)\alpha_\dagger}$ is decreasing, so for the checkpoint at $\mathcal{J}(m)$ it may be replaced by $\mathcal{J}(m)^{(\frac{3}{2}-s)\alpha_\dagger}$ inside the checkpoint sums, giving $\sum_m m^{\frac{\alpha_\xi + (\frac{3}{2}-s)\alpha_\dagger}{\alpha_\dagger}} \delta_m$-type sums, which Lemma D.3(iii) bounds by $C(N^{\alpha_\xi + (\frac{3}{2}-s)\alpha_\dagger + (\frac{1}{2}-s)\alpha_\dagger} + \log N) = C(N^{\alpha_\xi + (2-2s)\alpha_\dagger} + \log N)$). When $\alpha_\xi + (2s-4)\alpha_\dagger > 1$, the first term is $o(N^{-\frac{1}{2}})$; the second is $o(N^{-\frac{1}{2}})$ since $(3s - \frac{7}{2})\alpha_\dagger \ge (2s-1)\alpha_\dagger \ge q_\star \alpha_\dagger > \frac{1}{2}$ for $s \ge \frac{5}{2}$; and the third is $o(N^{-\frac{1}{2}})$ since $\alpha_\xi + (2-2s)\alpha_\dagger < \frac{1}{2}$ is implied by $\alpha_\xi < \frac{1}{2} + (s - \frac{1}{2})\alpha_\dagger$, using $2s - 2 \ge s - \frac{1}{2}$ for $s \ge \frac{3}{2}$. So the $l = 1$ term is $o_{\text{a.s.}}(N^{-\frac{1}{2}})$.

The block normal equations contain only no-jump indices. Hence (D.17) gives

$$\left\| \frac{1}{N} \sum_{m=J_0}^{J_N-1} \mathcal{R}_{m,J_N}^{\top} \mathbb{M}_m^{-1} \sum_{k=\mathcal{J}(m)+1}^{\mathcal{J}(m+1)-1} \zeta_{2,k} + \frac{1}{N} \mathbb{M}_{J_N}^{-1} \sum_{k=\mathcal{J}(J_N)+1}^{N} \zeta_{2,k} \right\|_2$$
$$\leq \frac{C}{N} \sum_{k=1}^{N} J_k^2 \|\Delta\widetilde{\omega}_{k-1}\|_2^2 \leq \frac{C}{N} \sum_{k=1}^{N} J_k^2 \|D_k\|_2^2 + \frac{C}{N} \sum_{k=1}^{N} J_k^2 \|T_k\|_2^2.$$

The index shifts are harmless because J_k/J_{k-1} is uniformly bounded. For the D-part, (D.18) gives

$$\frac{C}{N} \sum_{k=1}^{N} J_k^2 \|D_k\|_2^2 \leq C N^{-\alpha_\xi + 3\alpha_\dagger} \mathrm{PolyLog} + C N^{(6-4s)\alpha_\dagger} \mathrm{PolyLog}, \qquad \text{a.s.},$$

which is $o(N^{-\frac{1}{2}})$ when $\alpha_\xi > \frac{1}{2} + 3\alpha_\dagger$, using also $(4s-6)\alpha_\dagger \geq (2s-1)\alpha_\dagger \geq q_\star \alpha_\dagger > \frac{1}{2}$ for $s \geq \frac{5}{2}$. For the T-part, Lemma D.5(iii) (with the decreasing-weight device described under $l = 1$) gives

$$\frac{C}{N} \sum_{k=1}^{N} J_k^2 \|T_k\|_2^2 \leq \frac{C}{N} \left(N^{\alpha_\xi + (3-2s)\alpha_\dagger} + \log N \right) = o\left(N^{-\frac{1}{2}} \right),$$

since $\alpha_\xi + (3-2s)\alpha_\dagger < \frac{1}{2}$ is implied by $\alpha_\xi < \frac{1}{2} + (s - \frac{1}{2})\alpha_\dagger$, using $2s - 3 \geq s - \frac{1}{2}$ for $s \geq \frac{5}{2}$. So the $l = 2$ term is $o_{\text{a.s.}}(N^{-\frac{1}{2}})$.

For later gain and AME arguments, we also record an all-index checkpoint bound. Let $t_m = \mathcal{J}(m)$ and $R = (\alpha_\xi - \alpha_\dagger)/2$. Equations (D.18) and Lemma D.5(iv) imply

$$\|\Delta\widetilde{\omega}_{\mathcal{J}(m)-1}\|_2 = O\left(m^{-\frac{\alpha_\xi - \alpha_\dagger}{2\alpha_\dagger}} \mathrm{PolyLog}(m) + m^{-s} \right), \qquad \text{a.s.}$$

Because Condition 14 gives $\alpha_\dagger < 1/12$ and $\alpha_\xi > 9\alpha_\dagger$, we have $R/\alpha_\dagger > 4$. Therefore (D.17) yields

$$\sum_{\mathcal{J}(m) \leq N} \|\zeta_{2,\mathcal{J}(m)}\|_2 \leq C \sum_{m \leq J_N} m^2 \Bigg[\{ m^{-\frac{\alpha_\xi - \alpha_\dagger}{2\alpha_\dagger}} \mathrm{PolyLog} + m^{-s} \}^2$$
$$+ m^{-s} \{ m^{-\frac{\alpha_\xi - \alpha_\dagger}{2\alpha_\dagger}} \mathrm{PolyLog} + m^{-s} \} \Bigg] = O_{\text{a.s.}}(1). \quad \text{(D.20)}$$

Together with the no-jump calculation above, this gives

$$\sum_{k \leq N} \|\zeta_{2,k}\|_2 = o(\sqrt{N}), \qquad \text{a.s.} \quad \text{(D.21)}$$

For $l = 3$, let $\chi_k = \mathbf{1}\{J_k = J_{k-1}\}$ (and set $\chi_1 = 0$). Up to finitely many initialization indices, the block expression in (D.15) is exactly equal to $\frac{1}{N}\sum_{k=1}^{N} \mathcal{R}_{J_k,J_N}^{\top} \mathbb{M}_{J_k}^{-1} \chi_k \zeta_{3,k}$. Define $U_{3,k} = \chi_k \mathbb{M}_{J_k}^{-1}\zeta_{3,k}$, $U_{3,k}$ is a martingale difference. The empirical-Hessian bound and the uniform inverse bound give $\mathbb{E}_{k-1}\|U_{3,k}\|_2^2 \le CJ_k^3\|\Delta\widetilde{\omega}_{k-1}\|_2^2$. Put $a = \alpha_\dagger$ and $\rho = \alpha_\xi$. By (D.18), Lemma D.5(ii), and $\|D + T\|^2 \le 2\|D\|^2 + 2\|T\|^2$, we have that

$$\sum_{k=2}^{\infty} \frac{1}{k}\mathbb{E}_{k-1}\|U_{3,k}\|_2^2 \le C\sum_k k^{-1-\alpha_\xi+4\alpha_\dagger}\mathrm{PolyLog} + C\sum_k k^{-1+(7-4s)\alpha_\dagger}\mathrm{PolyLog}$$
$$+ C\sum_k k^{-1+(3-2s)\alpha_\dagger} < \infty \qquad \text{a.s.}$$

Here $\alpha_\xi > 4\alpha_\dagger$ follows from $\alpha_\xi > \frac{1}{2} + 3\alpha_\dagger$ and $\alpha_\dagger < \frac{1}{2}$, while $s > 7/2$ makes the other two exponents strictly below -1. Lemma D.4 therefore implies that the $l = 3$ block sum is $o_{\text{a.s.}}(N^{-1/2})$.

For $l = 4$, recall that the block sums involve $\zeta_{4,k}^\circ$ only through indices $\mathcal{J}(m) + 1 \le k \le \mathcal{J}(m+1) - 1$ at which the sieve dimension does not change, so $\zeta_{4,k}^\circ = 0$ there and this term vanishes identically (the deterministic mean $\mathbb{E}\zeta_{4,k}$ at expansion times is part of the transient T_k, already handled in (D.19)).

For $l = 5$, decompose $\zeta_{5,k} = b_{J_k} + \zeta_{5,k}^\circ$, where $\mathbb{E}_{k-1}\zeta_{5,k}^\circ = 0$, $\|b_{J_k}\|_2 \le CJ_k^{-q_\star}$, and $\mathbb{E}_{k-1}\|\zeta_{5,k}^\circ\|_2^2 \le CJ_k^{3-2s}$. The deterministic part satisfies the exact power-sum bound $\frac{C}{N}\sum_{k=1}^{N} J_k^{-q_\star} = o(N^{-1/2})$ because $q_\star\alpha_\dagger > 1/2$. For the centered part set $U_{5,k} = \chi_k\mathbb{M}_{J_k}^{-1}\zeta_{5,k}^\circ$. It is a martingale difference and

$$\sum_{k=2}^{\infty} \frac{1}{k}\mathbb{E}_{k-1}\|U_{5,k}\|_2^2 \le C\sum_{k=2}^{\infty} k^{-1+(3-2s)\alpha_\dagger} < \infty$$

because $s > 3/2$. Lemma D.4 gives an $o_{\text{a.s.}}(N^{-1/2})$ averaged centered term. Hence the entire $l = 5$ contribution is $o_{\text{a.s.}}(N^{-1/2})$.

The above implies that the following a.s. holds

$$\Delta\overline{\overline{\omega}}_N = \frac{1}{N}\sum_{k=1}^{N} \mathcal{R}_{J_k,J_N}^{\top}\mathbb{M}_{J_k}^{-1}\frac{1}{B}\sum_{i=1}^{B}\varepsilon_{i,k}\begin{pmatrix}\nabla_z F_0(Z_{i,k}(\theta_0))X_{i,k}\\ \Psi_{J_k}(Z_{i,k}(\theta_0))\end{pmatrix}$$
$$+ \frac{1}{N}\sum_{k=1}^{N}\left(\mathcal{R}_{J_k,J_N}^{\top}\omega_{J_k,0} - \omega_{J_N,0}\right) + o\left(N^{-\frac{1}{2}}\right).$$

Note that the first summation on the right-hand side omits the expansion indices, but define

$\chi_k = \mathbf{1}\{J_k > J_{k-1}\}$ and $U_{6,k} = \chi_k \mathbb{M}_{J_k}^{-1} \zeta_{6,k}$. Then $\mathbb{E}_{k-1}\|U_{6,k}\|_2^2 \le C\chi_k J_k$. We have that

$$\sum_{k=1}^{\infty} \frac{1}{k} \mathbb{E}_{k-1}\|U_{6,k}\|_2^2 \le C \sum_m \frac{m}{\mathcal{J}(m)} \asymp \sum_m m^{1-1/\alpha_\dagger} < \infty,$$

since $\alpha_\dagger < 1/2$. Lemma D.4 shows that the restored checkpoint score is $o_{\text{a.s.}}(\sqrt{N})$ before division by N.

To further simplify notations, define

$$\mathbb{M}_{J,11} = \mathbb{E}\left[(\nabla_z F_0(Z_0))^2 X X^\top\right], \quad \mathbb{M}_{J,12} = \mathbb{E}\left[\nabla_z F_0(Z_0) X \Psi_J(Z_0)^\top\right], \quad \mathbb{M}_{J,22} = \Gamma_J.$$

We have

$$\begin{aligned}
&\mathbb{M}_J^{-1} \begin{pmatrix} \varepsilon \nabla_z F_0(Z_0) X \\ \varepsilon \Psi_J(Z_0) \end{pmatrix} \\
&= \varepsilon \begin{pmatrix} \left(\mathbb{M}_{J,11} - \mathbb{M}_{J,12}\mathbb{M}_{J,22}^{-1}\mathbb{M}_{J,12}^\top\right)^{-1} \left(\nabla_z F_0(Z_0) X - \mathbb{M}_{J,12}\mathbb{M}_{J,22}^{-1}\Psi_J(Z_0)\right) \\ \left(\mathbb{M}_{J,22} - \mathbb{M}_{J,12}^\top\mathbb{M}_{J,11}^{-1}\mathbb{M}_{J,12}\right)^{-1} \left(\Psi_J(Z_0) - \mathbb{M}_{J,12}^\top\mathbb{M}_{J,11}^{-1}\nabla_z F_0(Z_0) X\right) \end{pmatrix}
\end{aligned}$$

We note that by Condition 6,

$$\begin{aligned}
&\left\|\Psi_J(\cdot)^\top \Gamma_J^{-1} \mathbb{E}\left(\nabla_z F_0(Z_0)\Psi_J(Z_0) X^\top\right) - \nabla_z F_0(\cdot)\mu_0(\theta_0, \cdot)^\top\right\|_\infty \\
&= \left\|\Psi_J(\cdot)^\top \Gamma_J^{-1} \mathbb{E}\left(\nabla_z F_0(Z_0)\Psi_J(Z_0)\mu_0(\theta_0, Z_0)^\top\right) - \nabla_z F_0(\cdot)\mu_0(\theta_0, \cdot)^\top\right\|_\infty \le CJ^{-s_\mu}.
\end{aligned}$$

So

$$\begin{aligned}
&\left\|\mathbb{M}_{J,12}\mathbb{M}_{J,22}^{-1}\mathbb{M}_{J,12}^\top - \mathbb{E}\left((\nabla_z F_0(Z_0))^2 \mu_0(\theta_0, Z_0)\mu_0(\theta_0, Z_0)^\top\right)\right\|_F \\
&\le \mathbb{E}[\|\nabla_z F_0(Z_0) X\|_2] \left\|\Psi_J(\cdot)^\top \Gamma_J^{-1} \mathbb{E}\left(\nabla_z F_0(Z_0)\Psi_J(Z_0) X^\top\right) - \nabla_z F_0(\cdot)\mu_0(\theta_0, \cdot)^\top\right\|_\infty \le CJ^{-s_\mu}.
\end{aligned}$$

This, due to uniformly lower bounded eigenvalues of $\mathbb{M}_\theta$ according to Condition 11 and $\|\mathbb{M}_J - \nabla_{\omega\omega}\mathcal{L}_J(\omega_{J,0})\|_F \le CJ^{\frac{3}{2}-s}$, implies that

$$\left\|\left(\mathbb{M}_{J,11} - \mathbb{M}_{J,12}\mathbb{M}_{J,22}^{-1}\mathbb{M}_{J,12}^\top\right)^{-1} - \mathbb{M}_\theta^{-1}\right\|_F \le CJ^{-s_\mu} \to 0, J \to \infty,$$

where recall that $\mathbb{M}_\theta = \mathbb{E}[(\nabla_z F_0(Z_0))^2 (X - \mu_0(\theta_0, Z_0))(X - \mu_0(\theta_0, Z_0))^\top]$. Similarly, we have that

$$\|\mathbb{M}_{J,12}\mathbb{M}_{J,22}^{-1}\Psi_J(Z_0) - \nabla_z F_0(Z_0)\mu_0(\theta_0, Z_0)\|_2 \le CJ^{-s_\mu}$$

uniformly for all Z_0. Together, we have that

$$\begin{aligned}
&\varepsilon \left(\mathbb{M}_{J,11} - \mathbb{M}_{J,12}\mathbb{M}_{J,22}^{-1}\mathbb{M}_{J,12}^{\top}\right)^{-1} \left(\nabla_z F_0(Z_0)X - \mathbb{M}_{J,12}\mathbb{M}_{J,22}^{-1}\Psi_J(Z_0)\right)\\
&= \varepsilon \mathbb{M}_\theta^{-1}\nabla_z F_0(Z_0)\left(X - \mu_0(\theta_0, Z_0)\right) + \varepsilon\delta_{J,\theta}(W),\\
&\qquad \|\delta_{J,\theta}(W)\|_2 \le CJ^{-s_\mu}(1 + \|X\|_2 + \|\mu_0(\theta_0, Z_0)\|_2) \text{ for all } W.
\end{aligned}$$

The above implies that

$$\begin{aligned}
\Delta\overline{\overline{\theta}}_N &= \frac{1}{N}\sum_{k=1}^{N}\frac{1}{B}\sum_{i=1}^{B}\varepsilon_{i,k}\mathbb{M}_\theta^{-1}\nabla_z F_0(Z_{0,i,k})\left(X_{i,k} - \mu_0(\theta_0, Z_{0,i,k})\right)\\
&\quad + \frac{1}{N}\sum_{k=1}^{N}\frac{1}{B}\sum_{i=1}^{B}\varepsilon_{i,k}\delta_{J_k,\theta}(W_{i,k}) + o\left(N^{-\frac{1}{2}}\right), \qquad \text{a.s.}
\end{aligned} \tag{D.22}$$

To finish the first result, set

$$U_{\theta,k} = \frac{1}{B}\sum_{i=1}^{B}\varepsilon_{i,k}\delta_{J_k,\theta}(W_{i,k}).$$

This is a fixed-dimensional martingale difference and the displayed envelope for $\delta_{J,\theta}$, together with Condition 10, gives $\mathbb{E}_{k-1}\|U_{\theta,k}\|_2^2 \le CJ_k^{-2s_\mu}$. Consequently $\sum_k k^{-1}\mathbb{E}_{k-1}\|U_{\theta,k}\|_2^2 < \infty$. The row part of Lemma D.4 yields $N^{-1}\sum_{k\le N}U_{\theta,k} = o_{\text{a.s.}}(N^{-1/2})$.

Also, for $\Delta\overline{\overline{\beta}}_N$, using the same reasoning, we have that

$$\begin{aligned}
\Delta\overline{\overline{\beta}}_N &= \frac{1}{N}\sum_{k=1}^{N}R_{J_k,J_N}^{\top}\mathbb{M}_{\beta,J_k}^{-1}\frac{1}{B}\sum_{i=1}^{B}\varepsilon_{i,k}\left(\Psi_{J_k}\left(Z_{0,i,k}\right) - P_{J_k}\nabla_z F_0(Z_{0,i,k})X_{i,k}\right)\\
&\quad + \frac{1}{N}\sum_{k=1}^{N}\left(R_{J_k,J_N}^{\top}\beta_{J_k,0} - \beta_{J_N,0}\right) + o\left(N^{-\frac{1}{2}}\right), \qquad \text{a.s.}
\end{aligned}$$

where recall that $P_J = \left(\mathbb{E}\left[\Psi_J(Z_0)\nabla_z F_0(Z_0)X^{\top}\right]\right)\left(\mathbb{E}\left[\left(\nabla_z F_0(Z_0)\right)^2 XX^{\top}\right]\right)^{-1}$.

D.4.4. Extension to estimated conditioning matrices

We now pass from the auxiliary identity-gain calculation to the update adjusted by conditioning matrix $\widehat{G}_k$. For simplicity, define

$$G_k^0 = G_{J_k,0}, \qquad E_k = \widehat{G}_k - G_k^0, \qquad B_k = (G_k^0)^{-1}E_k, \qquad d_k = \|E_k\|_{\text{op}},$$

and

$$\bar{g}_k(\omega) = \frac{1}{B}\sum_{i=1}^{B} \nabla_\omega \mathcal{L}_{J_k}(\omega, W_{i,k}).$$

All the analysis are based on the probability-one event on which the gain, localization, and rate conditions hold.

Stability, localization, and raw rates. On a no-jump step let H_k be the integrated population Hessian on the local segment. Condition 13 gives $d_k = o(1)$, and the local Hessian bound imply, eventually,

$$\lambda_{\min}\left\{\frac{\widehat{G}_k H_k + H_k \widehat{G}_k}{2}\right\} \geq c_G/2, \qquad \|\widehat{G}_k H_k\|_{\text{op}} \leq C.$$

Hence

$$\|(I - \xi_k \widehat{G}_k H_k)v\|_2^2 \leq (1 - c\xi_k)\|v\|_2^2 \tag{D.23}$$

for all sufficiently large k. At a sieve expansion, bounded re-embedding contributes the same $C\mathbf{1}_k$ factor as before. Premultiplication by the bounded predictable $\widehat{G}_k$ changes only constants in the score and empirical Hessian bounds. Repeating the centered-score expansion leading to (D.11) therefore gives the explicit recurrence

$$\mathbb{E}_{k-1}\|\Delta\widetilde{\omega}_k^G\|_2^2 \leq (1 - c\xi_k + C\mathbf{1}_k)\|\Delta\widetilde{\omega}_{k-1}^G\|_2^2 + C\xi_k^2 J_k + C\xi_k J_k^{-2q_\star} + C\mathbf{1}_k J_k^{-2s}. \tag{D.24}$$

Thus the initial rate and eventual inactivity of the projection follow by the same blockwise Robbins–Siegmund argument.

The refined decomposition also survives. The extra linear term generated by E_k is absorbed into the contraction in (D.23); its centered score has conditional covariance bounded by $Cd_k^2 J_k \leq CJ_k$; and all nonlinear terms are unchanged up to a constant. At a sieve increase, the local-Hessian remainder also contains the second term in (D.17), with the actual gain-path error in place of $\Delta\widetilde{\omega}_{k-1}$. Bounded predictable premultiplication changes only its constant, and Young's inequality produces the same $k^{-(2s-2)\alpha_\dagger}$ deterministic floor. Thus it does not alter either the raw-rate bootstrap or the componentwise theta rate. Let D_k^G and T_k^G denote the counterparts of D_k and T_k with preconditioning, the above implies that

$$\|D_k^G\|_2 = O_{\text{a.s.}}\left(k^{-\frac{\alpha_\xi - \alpha_\dagger}{2}}\text{PolyLog} + k^{-(2s-2)\alpha_\dagger}\right), \qquad \|T_k^G\|_2 \leq Ck^{-s\alpha_\dagger}, \tag{D.25}$$

with the weighted transient bounds in Lemma D.5.

Exact population-gain cancellation. Using the same score remainders as in (D.15), evaluated on the conditioned path, the no-projection recursion is

$$\Delta\widetilde{\omega}_k^G = A_k^G \Delta\widetilde{\omega}_{k-1}^G - \xi_k G_k^0 \left(\sum_{\ell=1}^{6} \zeta_{\ell,k} + \zeta_{G,k} \right) + \zeta_{7,k}, \tag{D.26}$$

where

$$A_k^G = (I - \xi_k G_k^0 \mathbb{M}_{J_k}) \mathcal{R}_{J_{k-1},J_k}^\top, \qquad \zeta_{G,k} = B_k \,\overline{g}_k(\mathcal{R}_{J_{k-1},J_k}^\top \widetilde{\omega}_{k-1}^G).$$

Define the checkpoint transient with $\tau_k^G = \zeta_{7,k} - \xi_k G_k^0 \mathbb{E}_{k-1}\zeta_{4,k}$. On a constant-$m$ block,

$$D_k^G = (I - \xi_k G_{m,0}\mathbb{M}_m) D_{k-1}^G - \xi_k G_{m,0} Z_k^G,$$

where

$$Z_k^G = \zeta_{1,k} + \zeta_{2,k} + \zeta_{3,k} + \zeta_{4,k}^\circ + \zeta_{5,k} + \zeta_{6,k} + \zeta_{G,k}.$$

The transition is uniformly stable because

$$I - \xi G_{m,0}\mathbb{M}_m = G_{m,0}^{1/2}\{I - \xi G_{m,0}^{1/2}\mathbb{M}_m G_{m,0}^{1/2}\} G_{m,0}^{-1/2}.$$

Most importantly, the exact normal equation is

$$D_{k-1}^G = \xi_k^{-1}\mathbb{M}_m^{-1} G_{m,0}^{-1}(D_{k-1}^G - D_k^G) - \mathbb{M}_m^{-1} Z_k^G. \tag{D.27}$$

After blockwise Abel summation, $G_{m,0}^{-1}$ appears only in boundary and step-size-difference terms. Those rows are uniformly bounded, so the boundary estimates in Subsection D.4.3 are unchanged. The gain cancels exactly from the score term in (D.27); consequently the leading PR score $-\mathbb{M}_m^{-1}\zeta_{6,k}$ remains.

Lemma D.6 (Estimated-gain averaging remainder)**.** *Under the primitive conditions in Remark D.1, for every predictable scalar a_k with $|a_k| \leq 1$,*

$$\frac{1}{\sqrt{N}} \left\| \sum_{k=1}^{N} a_k \mathcal{R}_{J_k,J_N}^\top \mathbb{M}_{J_k}^{-1} \zeta_{G,k} \right\|_2 \longrightarrow 0 \qquad a.s. \tag{D.28}$$

If $C_k \in \mathbb{R}^{q\times(p+J_k)}$ is predictable, q is fixed, and $\sup_k \|C_k\|_{\mathrm{op}} < \infty$, then also

$$\max_{1\leq n\leq N} \frac{1}{\sqrt{N}} \left\| \sum_{k=1}^{n} a_k C_k \mathbb{M}_{J_k}^{-1} \zeta_{G,k} \right\|_2 \longrightarrow 0 \qquad a.s. \tag{D.29}$$

Proof. The exact score decomposition is

$$\overline{g}_k(\mathcal{R}_{J_{k-1},J_k}^{\top}\widetilde{\omega}_{k-1}^{G}) = \mathbb{M}_{J_k}\mathcal{R}_{J_{k-1},J_k}^{\top}(D_{k-1}^{G} + T_{k-1}^{G}) + \sum_{\ell=1}^{6}\zeta_{\ell,k}. \tag{D.30}$$

For the regular state part, uniform bounds on $\mathbb{M}_J^{\pm 1}$ and $(G_J^0)^{-1}$ give

$$\frac{1}{\sqrt{N}}\sum_{k\le N} d_k\|D_{k-1}^{G}\|_2 \longrightarrow 0$$

by the second primitive condition in Remark D.1 and (D.25); the secondary state exponent $(2s-2)\alpha_\dagger$ is larger than R. Since $d_k \le 1$ eventually, Lemma D.5(iii) gives the same conclusion for T_{k-1}^{G}. These absolute bounds are uniform over partial sums and remain valid after any bounded row premultiplication.

Now we analyze the ζ terms in the PR average. The ζ_1 sum and the no-jump part of the ζ_2 sum are absolutely $o_{\text{a.s.}}(\sqrt{N})$ by the $l = 1, 2$ calculations in Section D.4.3. The checkpoint calculation (D.20) also applies to the actual gain path by (D.25), so the expansion-round ζ_2 sum is $O_{\text{a.s.}}(1)$. Multiplication by $\mathbb{M}_{J_k}^{-1}B_k$, whose norm is $O(d_k)$ and eventually bounded, does not enlarge them. For ζ_3, define $U_{G_3,k} = \mathbb{M}_{J_k}^{-1}B_k\zeta_{3,k}$. It is a martingale difference and $\mathbb{E}_{k-1}\|U_{G_3,k}\|_2^2 \le Cd_k^2J_k^3\|\Delta\widetilde{\omega}_{k-1}^{G}\|_2^2$. The normalized quadratic variation is summable by the calculation for $U_{3,k}$ above, since $d_k \le 1$ eventually. Lemma D.4 handles both (D.28) and the row-uniform version. For ζ_4 term, again decompose $\zeta_{4,k} = \mathbb{E}_{k-1}\zeta_{4,k} + \zeta_{4,k}^{\circ}$. For the centered part, $\mathbb{E}_{k-1}\|\mathbb{M}_{J_k}^{-1}B_k\zeta_{4,k}^{\circ}\|_2^2 \le Cd_k^2\mathbf{1}_kJ_k^{3-2s}$, whose normalized sum is finite. The mean is supported on checkpoints and

$$\sum_{k=1}^{N}\|\mathbb{M}_{J_k}^{-1}B_k\mathbb{E}_{k-1}\zeta_{4,k}\|_2 \le C\sum_{m\le J_N}\delta_m = O(\log N) = o(\sqrt{N})$$

by Lemma D.3(iii). Next write $\zeta_{5,k} = b_{J_k} + \zeta_{5,k}^{\circ}$. Its mean contributes $O(\sum_{k\le N}J_k^{-q_\star}) = o(\sqrt{N})$. The centered part is a martingale difference with conditional second moment at most $Cd_k^2J_k^{3-2s}$, so Lemma D.4 applies. Finally, define $U_{G_6,k} = \mathbb{M}_{J_k}^{-1}B_k\zeta_{6,k}$. $U_{G_6,k}$ is a martingale difference with $\mathbb{E}_{k-1}\|U_{G_6,k}\|_2^2 \le Cd_k^2J_k$. The first primitive condition in Remark D.1 is exactly the normalized quadratic-variation condition for this term. Summing the seven pieces in (D.30) proves both conclusions. We also point that the predictability of $\widehat{G}_k$ is used at every martingale step. □

Apply Lemma D.6 with the no-jump indicator in the block normal equations. Together with

(D.27), the identity-gain remainder calculations, and the transient average, it yields the same joint PR representation. This finishes the entire of Lemma 5.2 in the main text.

D.5. Proof of Theorem 5.1

Let

$$\psi_{\theta,k} = \frac{1}{B}\sum_{i=1}^{B} \varepsilon_{i,k}\mathbb{M}_\theta^{-1}F_0'(Z_{0,i,k})\{X_{i,k} - \mu_0(\theta_0, Z_{0,i,k})\}.$$

These are independent, mean-zero p-vectors with covariance Ω_θ and the moment required by the classical multivariate LIL. Its cluster-set form gives that the limit points of

$$(2N \log\log N)^{-1/2}\Omega_\theta^{-1/2}\sum_{k=1}^{N}\psi_{\theta,k}$$

form the closed unit ball. The almost-sure $o(N^{-1/2})$ representation in Lemma 5.3 is negligible on the LIL scale, proving the first part of the theorem including the stated directional equivalent.

To prove the second part, use the second expansion in Lemma 5.3 and retain its deterministic approximation part exactly as $\frac{1}{N}\sum_{k=1}^{N}\{\mathbb{P}_{J_k}F_0(z) - F_0(z)\}$. The basis envelope $\sup_{z\in\mathcal{Z}_0}\|\Psi_{J_N}(z)\|_2 \le CJ_N^{1/2}$ and the common-space martingale maximal bound give the stochastic term $O_{\text{a.s.}}(J_N^{1/2}N^{-1/2}(\log N)^{1/2})$. Uniform sieve approximation gives

$$\sup_{z\in\mathcal{Z}_0}\left|\frac{1}{N}\sum_{k=1}^{N}\{\mathbb{P}_{J_k}F_0(z) - F_0(z)\}\right| \le \frac{C}{N}\sum_{k=1}^{N}J_k^{-s}.$$

This proves the second part of the theorem.

D.6. Proof of Theorem 5.2

The batch influences $\psi_{\theta,k}$ defined in the preceding proof are iid, mean zero, have covariance Ω_θ, and have a finite $2+\delta$ moment under Condition 10. The multivariate invariance principle therefore gives $N^{-1/2}\sum_{k\le[Nr]}\psi_{\theta,k} \Rightarrow \Omega_\theta^{1/2}\mathbb{W}_p(r)$ for $r \in [0,1]$. It remains only to prove the uniform equivalence

$$\max_{0\le r\le 1}\left\|\frac{1}{\sqrt{N}}\sum_{k=1}^{[Nr]}\frac{1}{B}\sum_{i=1}^{B}\varepsilon_{i,k}\mathbb{M}_\theta^{-1}\nabla_z F_0(Z_{0,i,k})\left(X_{i,k}-\mu_0(\theta_0,Z_{0,i,k})\right)-\frac{[Nr]}{\sqrt{N}}\Delta\overline{\overline{\theta}}_{[Nr]}\right\|_2 \to_{a.s.} 0$$

For simplicity, define $e_N=\Delta\overline{\overline{\theta}}_N-\frac{1}{N}\sum_{k=1}^{N}\frac{1}{B}\sum_{i=1}^{B}\varepsilon_{i,k}\mathbb{M}_\theta^{-1}\nabla_z F_0(Z_{0,i,k})\left(X_{i,k}-\mu_0(\theta_0,Z_{0,i,k})\right)$, then it remains to show that

$$\max_{0\le r\le 1}\left\|\frac{[Nr]}{\sqrt{N}}e_{[Nr]}\right\|_2 \to_{a.s.} 0.$$

According to Lemma 5.3, we have that $\|e_N\|_2=o\left(N^{-\frac{1}{2}}\right)$. So for any $\epsilon>0$, we can find path-dependent N_1 such that for any $N>N_1$, there holds $\|e_N\|_2\le\frac{\epsilon}{\sqrt{N}}$. Also, for such ϵ, we can also find path-dependent N_2 such that $\max_{1\le k\le N_1}\|\frac{ke_k}{\sqrt{N}}\|_2\le\epsilon$. So together we have that for $N\ge N_2$, we have

$$\max_{0\le r\le 1}\left\|\frac{[Nr]}{\sqrt{N}}e_{[Nr]}\right\|_2\le\max\left\{\max_{1\le k\le N_1}\left\|\frac{ke_k}{\sqrt{N}}\right\|_2,\max_{N_1<k\le N}\left\|\frac{ke_k}{\sqrt{N}}\right\|_2\right\}\le\max\left\{\epsilon,\max_{N_1<k\le N}\frac{\epsilon\sqrt{k}}{\sqrt{N}}\right\}\le\epsilon.$$

Since ϵ can be arbitrarily small, this shows the result.

E. Average Marginal Effects

E.1. Additional Conditions for Theorem 5.3

Condition 15. *(i) $\mathbb{E}\|\nabla_z\mu_0(\theta_0,Z_0)\|_2^2<\infty$; (ii) Let $f_Z(z)$ denote the density of Z_0, which is continuously differentiable. There holds $\lim_{z\to\pm\infty}\mu_0(\theta_0,z)f_Z(z)=0$; define $\ell(z)=\nabla_z f_Z(z)/f_Z(z)$. Then $\mathbb{E}\ell^2(Z_0)<\infty$, and for all J there holds*

$$\mathbb{E}\left|\ell(Z_0)-\mathbb{P}_J(\ell)(Z_0)\right|^2\le C_f J^{-2s_f}$$

for some $s_f>0$. Moreover, $(s+s_f)\alpha_\dagger>\frac{1}{2}$; (iii) Define $\Upsilon_k=B^{-1}\sum_{i=1}^{B}\Upsilon_{i,k}$. There holds

$$\mathbb{E}\Upsilon_k\Upsilon_k^\top\longrightarrow V_\Upsilon$$

as $k\to\infty$ for some positive definite matrix V_Υ.

E.2. Proof of Theorem 5.3

Note that
$$\widetilde{\tau}_k = \frac{1}{B}\sum_{i=1}^{B}\left(\nabla_z \Psi_{J_{k-1}}\left(Z_{i,k}(\widetilde{\theta}_{k-1})\right)\right)^\top \widetilde{\beta}_{k-1}\widetilde{\theta}_{k-1}$$
can be decomposed as follows:
$$\begin{aligned}
\widetilde{\tau}_k &= \frac{1}{B}\sum_{i=1}^{B}\left(\nabla_z \Psi_{J_{k-1}}\left(Z_{i,k}(\widetilde{\theta}_{k-1})\right)\right)^\top \widetilde{\beta}_{k-1}\widetilde{\theta}_{k-1} \\
&= \underbrace{\frac{1}{B}\sum_{i=1}^{B}\left(\nabla_z \Psi_{J_{k-1}}(Z_{0,i,k})\right)^\top \widetilde{\beta}_{k-1}\widetilde{\theta}_{k-1}}_{(i)} \\
&\quad + \underbrace{\frac{1}{B}\sum_{i=1}^{B}\widetilde{\beta}_{k-1}^\top \left(\nabla_{zz}\Psi_{J_{k-1}}(Z_{0,i,k})\right) X_{i,k}^\top \Delta\widetilde{\theta}_{k-1}\widetilde{\theta}_{k-1}}_{(ii)} \\
&\quad + \underbrace{\frac{1}{B}\sum_{i=1}^{B}\int_0^1 (1-\tau)\,\widetilde{\beta}_{k-1}^\top \nabla_{zzz}\Psi_{J_{k-1}}\left(Z_{0,i,k}+\tau X_{i,k}^\top \Delta\widetilde{\theta}_{k-1}\right)\left(X_{i,k}^\top \Delta\widetilde{\theta}_{k-1}\right)^2 \widetilde{\theta}_{k-1}\,d\tau.}_{(iii)}
\end{aligned}$$

We first look at (iii). By Condition 6, Condition 10, the decomposition $\Delta\widetilde{\omega}_{k-1} = D_{k-1}+T_{k-1}$, and the proof of Section D.4.2,
$$\|(iii)\|_2 \le \frac{C}{B}\sum_{i=1}^{B}\|X_{i,k}\|_2^2\left\{\|D_{k-1}\|_2^2 + J_{k-1}^{\frac{7}{2}}\|D_{k-1}\|_2^3 + \|T_{k-1}\|_2^2 + J_{k-1}^{\frac{7}{2}}\|T_{k-1}\|_2^3\right\}, \qquad \text{a.s.}$$

Since $s > \frac{7}{2}$, $J_{k-1}^{7/2}\|T_{k-1}\|_2 \le C$ eventually. Therefore, by (D.18), Lemma D.5(ii)–(iii), and the same weighted strong-law argument used above,
$$\begin{aligned}
\left\|\frac{1}{N}\sum_{k=1}^{N}(iii)\right\|_2 &= O\left(N^{\left(\frac{7\alpha_\dagger}{2}-\frac{\alpha_\xi-\alpha_\dagger}{2}\right)_+ -\alpha_\xi+\alpha_\dagger}\mathrm{PolyLog} + N^{-(4s-4)\alpha_\dagger}\mathrm{PolyLog}\right. \\
&\qquad \left. + N^{(\frac{19}{2}-6s)\alpha_\dagger}\mathrm{PolyLog} + N^{-1}\mathrm{PolyLog}\right) \\
&\quad + O\left(\frac{1}{N}\left\{N^{\alpha_\xi+(1-2s)\alpha_\dagger} + \log N\right\}\right) \\
&= o\left(N^{-\frac{1}{2}}\right), \qquad \text{a.s.}
\end{aligned}$$

The leading term is $o(N^{-1/2})$ under Condition 14. When the positive part is active, it is enough that $3\alpha_\xi > 1+10\alpha_\dagger$, which follows from $\alpha_\xi > \frac{1}{2}+3\alpha_\dagger$ and $\alpha_\dagger < \frac{1}{2}$; when it is inactive, $\alpha_\xi > \frac{1}{2}+\alpha_\dagger$ suffices. The remaining terms are smaller under the same conditions. So we will only look at (i) and (ii).

We first look at the second term. For (ii), we can see that

$$\begin{aligned}
(ii) &= \frac{1}{B}\sum_{i=1}^{B} \beta_{J_{k-1},0}^\top \left(\nabla_{zz}\Psi_{J_{k-1}}(Z_{0,i,k})\right) \widetilde{\theta}_{k-1} X_{i,k}^\top \Delta\widetilde{\theta}_{k-1} \qquad (iv) \\
&\quad + \frac{1}{B}\sum_{i=1}^{B} \Delta\widetilde{\beta}_{k-1}^\top \left(\nabla_{zz}\Psi_{J_{k-1}}(Z_{0,i,k})\right) \widetilde{\theta}_{k-1} X_{i,k}^\top \Delta\widetilde{\theta}_{k-1} \qquad (v).
\end{aligned}$$

We first look at (iv). The exact decomposition is

$$\begin{aligned}
&(iv) - \frac{1}{B}\sum_{i=1}^{B} \nabla_{zz}F_0(Z_{0,i,k})\theta_0 X_{i,k}^\top \Delta\widetilde{\theta}_{k-1} \\
&= \frac{1}{B}\sum_{i=1}^{B} \nabla_{zz}\left\{\mathbb{P}_{J_{k-1}}(F_0) - F_0\right\}(Z_{0,i,k})\theta_0 X_{i,k}^\top \Delta\widetilde{\theta}_{k-1} \\
&\quad + \frac{1}{B}\sum_{i=1}^{B} \nabla_{zz}\mathbb{P}_{J_{k-1}}(F_0)(Z_{0,i,k})\Delta\widetilde{\theta}_{k-1} X_{i,k}^\top \Delta\widetilde{\theta}_{k-1}.
\end{aligned}$$

For the first term, integration by parts gives

$$\begin{aligned}
&\left\|\mathbb{E}\left[\nabla_{zz}\left\{\mathbb{P}_J(F_0) - F_0\right\}(Z_0)X^\top\right]\right\|_2 \\
&= \left\| -\int \nabla_z\left\{\mathbb{P}_J(F_0) - F_0\right\}(z)\left\{\nabla_z\mu_0(\theta_0,z)^\top f_Z(z) + \mu_0(\theta_0,z)^\top \nabla_z f_Z(z)\right\}dz \right\|_2 \\
&\le CJ^{1-s}\left[\left\{\mathbb{E}\|\nabla_z\mu_0(\theta_0,Z_0)\|_2^2\right\}^{1/2} + \left\{\mathbb{E}\|\mu_0(\theta_0,Z_0)\|_2^2\right\}^{1/2}\left\{\mathbb{E}\ell^2(Z_0)\right\}^{1/2}\right] \le CJ^{1-s}.
\end{aligned}$$

Hence, using $\Delta\widetilde{\omega}_{k-1} = D_{k-1} + T_{k-1}$,

$$\begin{aligned}
\frac{1}{N}\sum_{k=1}^{N} J_{k-1}^{1-s}\|\Delta\widetilde{\theta}_{k-1}\|_2 &\le CN^{(1-s)\alpha_\dagger - \frac{\alpha_\xi-\alpha_\dagger}{2}}\mathrm{PolyLog} + CN^{(3-3s)\alpha_\dagger}\mathrm{PolyLog} + CN^{-1}\mathrm{PolyLog} \\
&\quad + \frac{C}{N}\left\{N^{\alpha_\xi + (\frac{1}{2}-s)\alpha_\dagger} + \log N\right\} = o(N^{-\frac{1}{2}}), \qquad \text{a.s.}
\end{aligned}$$

After subtracting its conditional mean, the same martingale-difference argument gives an $o_{\mathrm{a.s.}}(N^{-1/2})$ running average. Indeed, using $\Delta\widetilde{\omega}_{k-1} = D_{k-1}+T_{k-1}$, (D.18), and Lemma D.5(ii), its conditional second moments divided by k are summable.

For the second term in the exact decomposition of (iv), boundedness of $\nabla_{zz}\mathbb{P}_{J_{k-1}}(F_0)$ gives

$$\begin{aligned}
&\frac{1}{N}\sum_{k=1}^{N}\left\|\frac{1}{B}\sum_{i=1}^{B}\nabla_{zz}\mathbb{P}_{J_{k-1}}(F_0)(Z_{0,i,k})\Delta\widetilde{\theta}_{k-1}X_{i,k}^{\top}\Delta\widetilde{\theta}_{k-1}\right\|_2 \\
&\le CN^{-\alpha_\xi+\alpha_\dagger}\mathrm{PolyLog}+CN^{-(4s-4)\alpha_\dagger}\mathrm{PolyLog}+CN^{-1}\mathrm{PolyLog} \\
&\quad+\frac{C}{N}\left\{N^{\alpha_\xi+(1-2s)\alpha_\dagger}+\log N\right\}=o(N^{-\frac{1}{2}}), \qquad \text{a.s.}
\end{aligned}$$

Consequently,

$$\frac{1}{N}\sum_{k=1}^{N}(iv)=\frac{1}{N}\sum_{k=1}^{N}\frac{1}{B}\sum_{i=1}^{B}\nabla_{zz}F_0(Z_{0,i,k})\theta_0X_{i,k}^{\top}\Delta\widetilde{\theta}_{k-1}+o(N^{-\frac{1}{2}}), \qquad \text{a.s.}$$

Now we look at $\left\|N^{-1}\sum_{k=1}^{N}(v)\right\|_2$. Similar to the previous strategy, we decompose each summand into its conditional mean and martingale-difference part. Note that

$$\begin{aligned}
\mathbb{E}\left[\nabla_{zz}\Psi_J(Z_0)X^{\top}\right]&=\int\nabla_{zz}\Psi_J(z)\mu_0(\theta_0,z)^{\top}f_Z(z)\,dz \\
&=-\mathbb{E}\left[\nabla_z\Psi_J(Z_0)\left\{\mu_0(\theta_0,Z_0)^{\top}\ell(Z_0)+\nabla_z\mu_0(\theta_0,Z_0)^{\top}\right\}\right].
\end{aligned}$$

Therefore, by Condition 15(i)–(ii),

$$\left\|\mathbb{E}\left[\nabla_{zz}\Psi_J(Z_0)X^{\top}\right]\right\|_F\le CJ^{\frac{3}{2}}.$$

It follows that

$$\begin{aligned}
\left\|\frac{1}{N}\sum_{k=1}^{N}\mathbb{E}_{k-1}(v)\right\|_2&\le\frac{C}{N}\sum_{k=1}^{N}J_k^{\frac{3}{2}}\left(\|D_{k-1}\|_2^2+\|T_{k-1}\|_2^2\right) \\
&\le CN^{\frac{5\alpha_\dagger}{2}-\alpha_\xi}\mathrm{PolyLog}+CN^{(\frac{11}{2}-4s)\alpha_\dagger}\mathrm{PolyLog}+CN^{-1}\mathrm{PolyLog} \\
&\quad+\frac{C}{N}\left\{N^{\alpha_\xi+(\frac{5}{2}-2s)\alpha_\dagger}+\log N\right\}=o(N^{-\frac{1}{2}}), \qquad \text{a.s.}
\end{aligned}$$

On the other side,

$$\sum_{k=2}^{\infty}\frac{\mathbb{E}_{k-1}\|(v)-\mathbb{E}_{k-1}(v)\|_2^2}{k} \le C\sum_{k=2}^{\infty}\left[k^{-1+7\alpha_{\dagger}-2\alpha_{\xi}}\text{PolyLog}+k^{-1+(13-8s)\alpha_{\dagger}}\text{PolyLog}\right.$$
$$\left.+k^{-1+(5-4s)\alpha_{\dagger}}\right]<\infty,\qquad \text{a.s.}$$

so

$$\left\|\frac{1}{N}\sum_{k=1}^{N}\{(v)-\mathbb{E}_{k-1}(v)\}\right\|_2 = o(N^{-\frac{1}{2}}),\qquad \text{a.s.}$$

Finally, again using $\Delta\widetilde{\omega}_{k-1}=D_{k-1}+T_{k-1}$,

$$\sum_{k=2}^{\infty}\frac{1}{k}\mathbb{E}_{k-1}\left\|\left\{\frac{1}{B}\sum_{i=1}^{B}\nabla_{zz}F_0(Z_{0,i,k})\theta_0X_{i,k}^{\top}-\mathbb{E}\left[\nabla_{zz}F_0(Z_0)\theta_0X^{\top}\right]\right\}\Delta\widetilde{\theta}_{k-1}\right\|_2^2<\infty,\qquad \text{a.s.}$$

and hence its running average is $o_{\text{a.s.}}(N^{-1/2})$. This shows that

$$\frac{1}{N}\sum_{k=1}^{N}(ii)=\mathbb{E}\left[\nabla_{zz}F_0(Z_0)\theta_0X^{\top}\right]\frac{1}{N}\sum_{k=1}^{N}\Delta\widetilde{\theta}_{k-1}+o(N^{-\frac{1}{2}}),\qquad \text{a.s.}$$

We finally look at (i). Expansion leads to

$$\begin{aligned}
&(i)-\mathbb{E}[\nabla_zF_0(Z_0)]\theta_0\\
&=\mathbb{E}\left[\nabla_z\Psi_{J_{k-1}}(Z_0)^{\top}\right]\beta_{J_{k-1},0}\theta_0-\mathbb{E}[\nabla_zF_0(Z_0)]\theta_0 \qquad (i.1)\\
&\quad+\left(\frac{1}{B}\sum_{i=1}^{B}\nabla_z\Psi_{J_{k-1}}(Z_{0,i,k})^{\top}-\mathbb{E}\left[\nabla_z\Psi_{J_{k-1}}(Z_0)^{\top}\right]\right)\beta_{J_{k-1},0}\theta_0 \qquad (i.2)\\
&\quad+\mathbb{E}\left[\nabla_z\Psi_{J_{k-1}}(Z_0)^{\top}\right]\Delta\widetilde{\beta}_{k-1}\theta_0 \qquad (i.3)\\
&\quad+\mathbb{E}\left[\nabla_z\Psi_{J_{k-1}}(Z_0)^{\top}\right]\beta_{J_{k-1},0}\Delta\widetilde{\theta}_{k-1} \qquad (i.4)\\
&\quad+\mathbb{E}\left[\nabla_z\Psi_{J_{k-1}}(Z_0)^{\top}\right]\Delta\widetilde{\beta}_{k-1}\Delta\widetilde{\theta}_{k-1} \qquad (i.5)\\
&\quad+\left(\frac{1}{B}\sum_{i=1}^{B}\nabla_z\Psi_{J_{k-1}}(Z_{0,i,k})^{\top}-\mathbb{E}\left[\nabla_z\Psi_{J_{k-1}}(Z_0)^{\top}\right]\right)\Delta\widetilde{\beta}_{k-1}\theta_0 \qquad (i.6)\\
&\quad+\left(\frac{1}{B}\sum_{i=1}^{B}\nabla_z\Psi_{J_{k-1}}(Z_{0,i,k})^{\top}-\mathbb{E}\left[\nabla_z\Psi_{J_{k-1}}(Z_0)^{\top}\right]\right)\beta_{J_{k-1},0}\Delta\widetilde{\theta}_{k-1} \qquad (i.7)\\
&\quad+\left(\frac{1}{B}\sum_{i=1}^{B}\nabla_z\Psi_{J_{k-1}}(Z_{0,i,k})^{\top}-\mathbb{E}\left[\nabla_z\Psi_{J_{k-1}}(Z_0)^{\top}\right]\right)\Delta\widetilde{\beta}_{k-1}\Delta\widetilde{\theta}_{k-1} \qquad (i.8).
\end{aligned}$$

We look at (i.1)–(i.8) separately. For (i.1), by Condition 15(ii) and integration by parts,

$$\begin{aligned}\mathbb{E}\left[\nabla_z \left\{\mathbb{P}_J(F_0)(Z_0) - F_0(Z_0)\right\}\right] &= \int \nabla_z \left\{\mathbb{P}_J(F_0)(z) - F_0(z)\right\} f_Z(z)\, dz \\ &= \left\{\mathbb{P}_J(F_0)(z) - F_0(z)\right\} f_Z(z)\big|_{z=-\infty}^{z=+\infty} \\ &\quad - \int \left\{\mathbb{P}_J(F_0)(z) - F_0(z)\right\} \nabla_z f_Z(z)\, dz \\ &= -\mathbb{E}\left[\left\{\mathbb{P}_J(F_0)(Z_0) - F_0(Z_0)\right\} \ell(Z_0)\right] \\ &= -\mathbb{E}\left[\left\{\mathbb{P}_J(F_0)(Z_0) - F_0(Z_0)\right\} \left\{\ell(Z_0) - \mathbb{P}_J(\ell)(Z_0)\right\}\right],\end{aligned}$$

because $\mathbb{P}_J(\ell) \in \mathcal{S}_J$ and the projection residual is orthogonal to $\mathcal{S}_J$. Thus, under Condition 15(ii),

$$\|(i.1)\|_2 \le C\|\mathbb{P}_J(F_0) - F_0\|_\infty \left\{\mathbb{E}\left[\ell(Z_0) - \mathbb{P}_J(\ell)(Z_0)\right]^2\right\}^{1/2} \le C J_{k-1}^{-s-s_f}.$$

When $(s + s_f)\alpha_\dagger > 1/2$,

$$\left\| \frac{1}{N} \sum_{k=1}^{N} (i.1) \right\|_2 = o(N^{-1/2}).$$

For (i.2), we note that

$$\begin{aligned}&(i.2) - \frac{1}{B}\sum_{i=1}^{B} \left\{\nabla_z F_0(Z_{0,i,k}) - \mathbb{E}[\nabla_z F_0(Z_0)]\right\} \theta_0 \\ &= \frac{1}{B}\sum_{i=1}^{B} \Big[\nabla_z \left\{\mathbb{P}_{J_{k-1}}(F_0)(Z_{0,i,k}) - F_0(Z_{0,i,k})\right\} \\ &\qquad\qquad -\mathbb{E}\nabla_z \left\{\mathbb{P}_{J_{k-1}}(F_0)(Z_0) - F_0(Z_0)\right\}\Big] \theta_0.\end{aligned}$$

Using the proof strategy for sums of martingale-difference sequences, when $s > 1$, as guaranteed by Condition 14,

$$\begin{aligned}&\Bigg\| \frac{1}{N}\sum_{k=1}^{N} \frac{1}{B}\sum_{i=1}^{B} \Big[\nabla_z \left\{\mathbb{P}_{J_{k-1}}(F_0)(Z_{0,i,k}) - F_0(Z_{0,i,k})\right\} \\ &\qquad -\mathbb{E}\nabla_z \left\{\mathbb{P}_{J_{k-1}}(F_0)(Z_0) - F_0(Z_0)\right\}\Big] \theta_0 \Bigg\|_2 = o(N^{-1/2}), \qquad \text{a.s.}\end{aligned}$$

As a result,

$$\frac{1}{N}\sum_{k=1}^{N}(i.2) = \frac{1}{N}\sum_{k=1}^{N}\frac{1}{B}\sum_{i=1}^{B}\left\{\nabla_z F_0(Z_{0,i,k}) - \mathbb{E}[\nabla_z F_0(Z_0)]\right\}\theta_0 + o(N^{-1/2}), \qquad \text{a.s.}$$

Next, for (i.3), note that

$$\begin{aligned}
&\frac{1}{N}\sum_{k=1}^{N}\mathbb{E}\left[\nabla_z\Psi_{J_{k-1}}(Z_0)^\top\right]\Delta\widetilde{\beta}_{k-1}\theta_0 \\
&= \frac{N-1}{N}\mathbb{E}\left[\nabla_z\Psi_{J_{N-1}}(Z_0)^\top\right]\Delta\overline{\overline{\beta}}_{N-1}\theta_0 + \frac{1}{N}\sum_{k=1}^{N}\mathbb{E}\left[\nabla_z\left\{\mathbb{P}_{J_{N-1}}(F_0)(Z_0) - \mathbb{P}_{J_{k-1}}(F_0)(Z_0)\right\}\right]\theta_0.
\end{aligned}$$

So we only need to look at

$$\mathbb{E}\left[\nabla_z\Psi_{J_N}(Z_0)^\top\right]\Delta\overline{\overline{\beta}}_N.$$

According to Lemma 5.3,

$$\begin{aligned}
\Delta\overline{\overline{\beta}}_N = {} & \frac{1}{N}\sum_{k=1}^{N}R_{J_k,J_N}^\top\mathbb{M}_{\beta,J_k}^{-1}\frac{1}{B}\sum_{i=1}^{B}\varepsilon_{i,k}\left\{\Psi_{J_k}(Z_{0,i,k}) - P_{J_k}\nabla_z F_0(Z_{0,i,k})X_{i,k}\right\} \\
&+ \frac{1}{N}\sum_{k=1}^{N}\left(R_{J_k,J_N}^\top\beta_{J_k,0} - \beta_{J_N,0}\right) + o(N^{-1/2}), \qquad \text{a.s.}
\end{aligned}$$

We have that

$$\begin{aligned}
&\mathbb{E}\left[\nabla_z\Psi_{J_N}(Z_0)^\top\right]\Delta\overline{\overline{\beta}}_N \\
&= \frac{1}{N}\sum_{k=1}^{N}\frac{1}{B}\sum_{i=1}^{B}\varepsilon_{i,k}\mathbb{E}\left[\nabla_z\Psi_{J_k}(Z_0)^\top\right]\mathbb{M}_{\beta,J_k}^{-1}\left\{\Psi_{J_k}(Z_{0,i,k}) - P_{J_k}\nabla_z F_0(Z_{0,i,k})X_{i,k}\right\} \\
&\quad + \frac{1}{N}\sum_{k=1}^{N}\mathbb{E}\left[\nabla_z\left\{\mathbb{P}_{J_k}(F_0)(Z_0) - F_0(Z_0)\right\}\right] - \mathbb{E}\left[\nabla_z\left\{\mathbb{P}_{J_N}(F_0)(Z_0) - F_0(Z_0)\right\}\right] + o(N^{-1/2}), \qquad \text{a.s.}
\end{aligned}$$

where the last $o(N^{-1/2})$ follows from

$$\sup_J\left\|\mathbb{E}\left[\nabla_z\Psi_J(Z_0)\right]\right\|_2 < \infty.$$

Indeed, integration by parts and Cauchy–Schwarz give

$$\left\|\mathbb{E}\left[\nabla_z\Psi_J(Z_0)\right]\right\|_2 = \left\|\mathbb{E}\left[\Psi_J(Z_0)\ell(Z_0)\right]\right\|_2 \le \overline{\lambda}(\Gamma_J)^{1/2}\left\{\mathbb{E}|\ell(Z_0)|^2\right\}^{1/2} \le C$$

under Condition 7 and Condition 15(ii). Moreover, according to the proof of (i.1), the second and third lines above are bounded by

$$\frac{C}{N}\sum_{k=1}^{N} J_k^{-s-s_f} + CJ_N^{-s-s_f} = O\left(N^{-(s+s_f)\alpha_\dagger}\right) = o(N^{-1/2}).$$

Therefore,

$$\begin{aligned}
&\mathbb{E}\left[\nabla_z \Psi_{J_N}(Z_0)^\top\right] \Delta\overline{\overline{\beta}}_N \\
&= \frac{1}{N}\sum_{k=1}^{N}\frac{1}{B}\sum_{i=1}^{B}\varepsilon_{i,k}\mathbb{E}\left[\nabla_z \Psi_{J_k}(Z_0)^\top\right]\mathbb{M}_{\beta,J_k}^{-1}\left\{\Psi_{J_k}(Z_{0,i,k}) - P_{J_k}\nabla_z F_0(Z_{0,i,k})X_{i,k}\right\} \\
&\quad + o(N^{-1/2}), \qquad \text{a.s.}
\end{aligned}$$

For (i.4), we have that

$$\left\|(i.4) - \mathbb{E}[\nabla_z F_0(Z_0)]\Delta\widetilde{\theta}_{k-1}\right\|_2 \le CJ_{k-1}^{1-s}\left(\|D_{k-1}\|_2 + \|T_{k-1}\|_2\right), \qquad \text{a.s.}$$

Consequently,

$$\begin{aligned}
&\frac{1}{N}\sum_{k=1}^{N}\left\|(i.4) - \mathbb{E}[\nabla_z F_0(Z_0)]\Delta\widetilde{\theta}_{k-1}\right\|_2 \\
&\le CN^{-(s-1)\alpha_\dagger - \frac{\alpha_\xi - \alpha_\dagger}{2}}\mathrm{PolyLog} + CN^{(3-3s)\alpha_\dagger}\mathrm{PolyLog} + CN^{-1}\mathrm{PolyLog} \\
&\quad + \frac{C}{N}\left\{N^{\alpha_\xi + (\frac{1}{2}-s)\alpha_\dagger} + \log N\right\} = o(N^{-1/2}), \qquad \text{a.s.}
\end{aligned}$$

Using the PR representation of $\Delta\overline{\overline{\theta}}_N$, we obtain

$$\begin{aligned}
\frac{1}{N}\sum_{k=1}^{N}(i.4) = \frac{1}{N}\sum_{k=1}^{N}\frac{1}{B}\sum_{i=1}^{B}&\varepsilon_{i,k}\mathbb{E}[\nabla_z F_0(Z_0)]\mathbb{M}_\theta^{-1}\nabla_z F_0(Z_{0,i,k}) \\
&\times\left\{X_{i,k} - \mu_0(\theta_0, Z_{0,i,k})\right\} + o(N^{-1/2}), \qquad \text{a.s.}
\end{aligned}$$

For (i.5),

$$\begin{aligned}\frac{1}{N}\sum_{k=1}^{N}\|(i.5)\|_2 &\le \frac{C}{N}\sum_{k=1}^{N}\left(\|D_{k-1}\|_2^2+\|T_{k-1}\|_2^2\right)\\ &\le CN^{-\alpha_\xi+\alpha_\dagger}\mathrm{PolyLog}+CN^{-(4s-4)\alpha_\dagger}\mathrm{PolyLog}+CN^{-1}\mathrm{PolyLog}\\ &\quad+\frac{C}{N}\left\{N^{\alpha_\xi+(1-2s)\alpha_\dagger}+\log N\right\}=o(N^{-1/2}),\qquad \text{a.s.}\end{aligned}$$

For (i.6)–(i.8), note that

$$\mathbb{E}\left\|\nabla_z\Psi_J(Z_0)-\mathbb{E}\nabla_z\Psi_J(Z_0)\right\|_2^2\le CJ^3.$$

Using $\Delta\widetilde{\omega}_{k-1}=D_{k-1}+T_{k-1}$, (D.18), and Lemma D.5(ii), for each of these terms we have

$$\sum_{k=2}^{\infty}\frac{1}{k}\mathbb{E}_{k-1}\|(i.l)\|_2^2\le C\sum_{k=2}^{\infty}\left[k^{-1-\alpha_\xi+4\alpha_\dagger}\mathrm{PolyLog}+k^{-1+(7-4s)\alpha_\dagger}\mathrm{PolyLog}+k^{-1+(3-2s)\alpha_\dagger}\right]<\infty,\ l=6,7,8.$$

Thus their running averages are all $o_{\text{a.s.}}(N^{-1/2})$.

Combining the above analysis, we have that

$$\begin{aligned}\Delta\overline{\overline{\tau}}_N &= \frac{1}{N}\sum_{k=1}^{N}\frac{1}{B}\sum_{i=1}^{B}\varepsilon_{i,k}\mathbb{E}\left[\nabla_{zz}F_0(Z_0)\theta_0X^\top\right]\mathbb{M}_\theta^{-1}\nabla_zF_0(Z_{0,i,k})\left\{X_{i,k}-\mu_0(\theta_0,Z_{0,i,k})\right\}\\ &\quad+\frac{1}{N}\sum_{k=1}^{N}\frac{1}{B}\sum_{i=1}^{B}\left\{\nabla_zF_0(Z_{0,i,k})-\mathbb{E}[\nabla_zF_0(Z_0)]\right\}\theta_0\\ &\quad+\frac{1}{N}\sum_{k=1}^{N}\frac{1}{B}\sum_{i=1}^{B}\varepsilon_{i,k}\theta_0\mathbb{E}\left[\nabla_z\Psi_{J_k}(Z_0)^\top\right]\mathbb{M}_{\beta,J_k}^{-1}\left\{\Psi_{J_k}(Z_{0,i,k})-P_{J_k}\nabla_zF_0(Z_{0,i,k})X_{i,k}\right\}\\ &\quad+\frac{1}{N}\sum_{k=1}^{N}\frac{1}{B}\sum_{i=1}^{B}\varepsilon_{i,k}\mathbb{E}[\nabla_zF_0(Z_0)]\mathbb{M}_\theta^{-1}\nabla_zF_0(Z_{0,i,k})\left\{X_{i,k}-\mu_0(\theta_0,Z_{0,i,k})\right\}\\ &\quad+o(N^{-1/2}),\qquad\text{a.s.}\end{aligned}$$

Therefore,

$$\Delta\overline{\overline{\tau}}_N=\frac{1}{N}\sum_{k=1}^{N}\Upsilon_k+o(N^{-1/2}),\qquad\text{a.s.}$$

Under Condition 15(iii), Cesàro summation gives

$$\frac{1}{N}\sum_{k=1}^{[Nr]}\mathbb{E}\left[\Upsilon_k\Upsilon_k^\top\right] \longrightarrow rV_\Upsilon, \qquad r\in[0,1].$$

We verify the Lindeberg condition via Lyapunov's criterion. The first, second, and fourth components of $\Upsilon_{i,k}$ have finite $(2+\delta)$-th moments for $\delta\in(0,\kappa-2)$ by Condition 6 and Condition 10. For the third component, since

$$\|\mathbb{E}\left[\nabla_z\Psi_J(Z_0)\right]\|_2 \le C,$$

the eigenvalues of $\mathbb{M}_{\beta,J}$ are uniformly bounded away from zero, $\|\Psi_J\|_\infty \le CJ^{1/2}$, and $\|P_J\|_F \le CJ^{1/2}$, its norm is bounded by

$$C|\varepsilon_{i,k}|J_k^{1/2}\left(1+\|X_{i,k}\|_2\right).$$

Hence

$$\mathbb{E}\|\Upsilon_k\|_2^{2+\delta} \le CJ_k^{1+\delta/2},$$

and

$$N^{-1-\frac{\delta}{2}}\sum_{k=1}^{N}\mathbb{E}\|\Upsilon_k\|_2^{2+\delta} \le CN^{-\frac{\delta}{2}+\left(1+\frac{\delta}{2}\right)\alpha_\dagger} \longrightarrow 0$$

whenever

$$\delta > \frac{2\alpha_\dagger}{1-\alpha_\dagger}.$$

Such

$$\delta\in\left(\frac{2\alpha_\dagger}{1-\alpha_\dagger},\kappa-2\right)$$

exists under Condition 10. The multivariate Lyapunov FCLT therefore gives

$$\frac{1}{\sqrt{N}}\sum_{k=1}^{[Nr]}\Upsilon_k \Longrightarrow V_\Upsilon^{1/2}\mathbb{W}_p(r), \qquad r\in[0,1].$$

Combining this with the almost-sure $o(N^{-1/2})$ remainder and the argument in the proof of Theorem 5.2, we obtain

$$(NV_\Upsilon)^{-1/2}[Nr]\,\Delta\overline{\overline{\tau}}_{[Nr]} \Longrightarrow \mathbb{W}_p(r), \qquad r\in[0,1].$$

E.3. Average Marginal Effects of Binary Regressors

For the binary-regressor marginal effect defined in the main text, an additional counterfactual overlap condition is necessary. Without introducing further notation, it is enough to assume directly that, for the corresponding j,

$$\sup_J \left\| \mathbb{E}\left[\Psi_J(Z_{0,-j}+\theta_{0,j}) - \Psi_J(Z_{0,-j})\right]\right\|_2 < \infty$$

and

$$\frac{1}{N}\sum_{k=1}^{N}\left|\mathbb{E}\left[\left\{\mathbb{P}_{J_{k-1}}(F_0) - F_0\right\}(Z_{0,-j}+\theta_{0,j}) - \left\{\mathbb{P}_{J_{k-1}}(F_0) - F_0\right\}(Z_{0,-j})\right]\right| = o(N^{-1/2}).$$

The second display is the direct counterfactual analogue of the small-bias condition used for the continuous average marginal effect. The centered difference between the sieve and true counterfactual signals has conditional second moment bounded by CJ_{k-1}^{-2s}, so its running average is $o_{\mathrm{a.s.}}(N^{-1/2})$ by the same martingale argument used for (i.2).

A first-order expansion of the two counterfactual evaluations gives

$$\begin{aligned}
&\frac{1}{N}\sum_{k=1}^{N}(\widetilde{\tau}_{k,j} - \tau_{0,j}) \\
&= \frac{1}{N}\sum_{k=1}^{N}\frac{1}{B}\sum_{i=1}^{B}\left[F_0\left(x_{0,i,k} + \sum_{l\neq j}\theta_{0,l}x_{l,i,k} + \theta_{0,j}\right) - F_0\left(x_{0,i,k} + \sum_{l\neq j}\theta_{0,l}x_{l,i,k}\right) - \tau_{0,j}\right] \\
&\quad + \mathbb{E}\left[\nabla_\theta\left\{F_0\left(x_0 + \sum_{l\neq j}\theta_l x_l + \theta_j\right) - F_0\left(x_0 + \sum_{l\neq j}\theta_l x_l\right)\right\}\Bigg|_{\theta=\theta_0}\right]\frac{1}{N}\sum_{k=1}^{N}\Delta\widetilde{\theta}_{k-1} \\
&\quad + \frac{1}{N}\sum_{k=1}^{N}\mathbb{E}\left[\Psi_{J_{k-1}}(Z_{0,-j}+\theta_{0,j}) - \Psi_{J_{k-1}}(Z_{0,-j})\right]^\top \Delta\widetilde{\beta}_{k-1} + o(N^{-1/2}), \qquad \text{a.s.}
\end{aligned}$$

Indeed, the second-order and centered empirical remainders are bounded by the same quantities used above for (iii), (v), and (i.5)–(i.8). In particular, using

$$\Delta\widetilde{\omega}_{k-1} = D_{k-1} + T_{k-1},$$

their D_{k-1}-parts follow from (D.18), while their T_{k-1}-parts follow from Lemma D.5(ii)–(iii), and all are $o_{\mathrm{a.s.}}(N^{-1/2})$ after averaging.

Using the PR representations of $\Delta\overline{\overline{\theta}}_N$ and $\Delta\overline{\overline{\beta}}_N$, together with the preceding counterfactual small-bias condition, we obtain

$$\overline{\overline{\tau}}_{N,j} - \tau_{0,j} = \frac{1}{N}\sum_{k=1}^{N} \Upsilon_k + o(N^{-1/2}), \qquad \text{a.s.},$$

where, for this binary-regressor marginal effect,

$$\begin{aligned}
\Upsilon_{i,k} = {} & F_0\left(x_{0,i,k} + \sum_{l\neq j}\theta_{0,l}x_{l,i,k} + \theta_{0,j}\right) - F_0\left(x_{0,i,k} + \sum_{l\neq j}\theta_{0,l}x_{l,i,k}\right) - \tau_{0,j} \\
& + \varepsilon_{i,k}\mathbb{E}\left[\nabla_\theta\left\{F_0\left(x_0 + \sum_{l\neq j}\theta_l x_l + \theta_j\right) - F_0\left(x_0 + \sum_{l\neq j}\theta_l x_l\right)\right\}\bigg|_{\theta=\theta_0}\right] \\
& \quad \times \mathbb{M}_\theta^{-1}\nabla_z F_0(Z_{0,i,k})\left\{X_{i,k} - \mu_0(\theta_0, Z_{0,i,k})\right\} \\
& + \varepsilon_{i,k}\mathbb{E}\left[\Psi_{J_k}(Z_{0,-j} + \theta_{0,j}) - \Psi_{J_k}(Z_{0,-j})\right]^\top \mathbb{M}_{\beta,J_k}^{-1} \\
& \quad \times \left\{\Psi_{J_k}(Z_{0,i,k}) - P_{J_k}\nabla_z F_0(Z_{0,i,k})X_{i,k}\right\}.
\end{aligned}$$

Thus, when Condition 15(iii) is imposed for

$$\Upsilon_k = \frac{1}{B}\sum_{i=1}^{B}\Upsilon_{i,k},$$

the same Lyapunov argument yields

$$(NV_\Upsilon)^{-1/2}[Nr]\left\{\overline{\overline{\tau}}_{[Nr],j} - \tau_{0,j}\right\} \Longrightarrow \mathbb{W}(r), \qquad r \in [0,1].$$

F. Proof of Theorem B.1 in Appendix B

This Appendix provides details for the proof of Theorem B.1. In particular, Section F.1 provides some preliminary results; Section F.2 extends the results to the case where some regressors are discrete. Section F.3 provides the asymptotic properties of the raw updates $\widehat{\theta}_N$. Finally, the last section concludes with the proof of Theorem B.1.

F.1. Preliminaries

Define

$$\Phi_k(\theta, \mathcal{W}_k) = \frac{1}{h_k \cdot B(B-1)} \cdot \sum_{i_1 \neq i_2}^{B} \mathcal{K}\left(\frac{Z_{i_1,k}(\theta) - Z_{i_2,k}(\theta)}{h_k}\right)(Y_{i_1,k} - Y_{i_2,k})(X_{i_1,k} - X_{i_2,k}), \tag{F.1}$$

$$\Phi_k(\theta) = \mathbb{E}[\Phi_k(\theta, \mathcal{W}_k)], \tag{F.2}$$

and

$$\Phi(\theta) = \lim_{k\to\infty} \Phi_k(\theta). \tag{F.3}$$

Given Condition 1 and Condition 2, we have the following lemma that describes the properties of $\Phi_k(\theta, \mathcal{W}_k)$, $\Phi_k(\theta)$, and $\Phi(\theta)$.

Lemma F.1. *Let Condition 1 and Condition 2 hold, for any choice of $\{h_k\}_{k=1}^{\infty}$ with $h_k \downarrow 0$, the following results hold:*

1. *For any $\theta \in \mathbb{R}^p$ we have:*

$$\Phi(\theta) = \int \left[F_0\left(z - X_1^\top \Delta\theta\right) - F_0\left(z - X_2^\top \Delta\theta\right)\right](X_1 - X_2) \times f\left(z - X_1^\top\theta, X_1\right) f\left(z - X_2^\top\theta, X_2\right) dz dX_1 dX_2;$$

2. *There is a finite constant $C_{\Phi,1} > 0$ such that*

$$\sup_{\theta\in\mathbb{R}^p} \|\Phi_k(\theta) - \Phi(\theta)\|_2 \leq C_{\Phi,1} h_k^{s_\mathcal{K}};$$

3. *There holds*

$$\mathbb{E}\left[(\Phi_k(\theta, \mathcal{W}_k) - \Phi_k(\theta))(\Phi_k(\theta, \mathcal{W}_k) - \Phi_k(\theta))^\top\right] = \frac{2\mathcal{V}_\mathcal{K}(\theta)}{B(B-1)h_k} + O(1),$$

where

$$\mathcal{V}_\mathcal{K}(\theta) = \int \mathcal{K}^2(t)dt \int \left(\sigma^2(z - X_1^\top\theta, X_1) + \sigma^2(z - X_2^\top\theta, X_2) + \left(F_0(z - X_1^\top\Delta\theta) - F_0(z - X_2^\top\Delta\theta)\right)^2\right) \times (X_1 - X_2)(X_1 - X_2)^\top f(z - X_1^\top\theta, X_1) f(z - X_2^\top\theta, X_2) dz dX_1 dX_2;$$

Consequently, there are finite constants $C_{\Phi,2}$ such that

$$\sup_{\theta\in\mathbb{R}^p} \mathbb{E}\left\|\Phi_k(\theta, W_k) - \Phi_k(\theta)\right\|_2^2 \le C_{\Phi,2} h_k^{-1}.$$

4. *There exist positive constants $C_{\Phi,3}$ such that*

$$\sup_{\theta\in\mathbb{R}^p} \mathbb{E}\left\|\nabla_\theta\Phi_k(\theta, W_k) - \nabla_\theta\Phi_k(\theta)\right\|_2^2 \le C_{\Phi,3} h_k^{-3}.$$

In particular, for the fixed batch size maintained in this paper this is $O(h_k^{-3})$.

5. *There is a finite constant $C_{\Phi,4}$ such that*

$$\sup_{\theta\in\mathbb{R}^p} \mathbb{E}\left\|\Phi_k(\theta, W_k) - \Phi_k(\theta)\right\|_2^4 \le C_{\Phi,4} h_k^{-3}.$$

Proof of Lemma F.1. We first prove Lemma F.1(1) and Lemma F.1(2). Obviously, under Condition 1, the expression in Lemma F.1(1) is well defined. For arbitrary x_0, X, define $Z(\theta) = x_0 + X^\top\theta$. Then

$$\mathbb{E}(Y|x_0, X) = F_0(x_0 + X^\top\theta_0) = F_0(Z(\theta) - X^\top\Delta\theta). \tag{F.4}$$

So

$$\begin{aligned}\Phi_k(\theta) &= \mathbb{E}\left[h_k^{-1}\mathcal{K}\left(h_k^{-1}\left(Z_{1,k}(\theta) - Z_{2,k}(\theta)\right)\right)\left(Y_{1,k} - Y_{2,k}\right)\left(X_{1,k} - X_{2,k}\right)\right] \\ &= \int h_k^{-1}\mathcal{K}\left(h_k^{-1}\left(z_1 - z_2\right)\right)\left(F_0(z_1 - X_1^\top\Delta\theta) - F_0(z_2 - X_2^\top\Delta\theta)\right)\times \\ &\qquad\qquad \left(X_1 - X_2\right) v(z_1, X_1|\theta) v(z_2, X_2|\theta) dz_1 dz_2 dX_1 dX_2,\end{aligned} \tag{F.5}$$

where $v(z, X|\theta)$ is the joint density of $Z(\theta)$ and X given θ. Now we derive the expression of $v(z, X|\theta)$. Since

$$\Pr(Z_{1,k}(\theta) \le z, X_{1,k} \le X) = \int_{X_{1,k}\le X}\int_{x_{0,1,k}\le z - X_{1,k}^\top\theta} f(x_{0,1,k}, X_{1,k}) dx_{0,1,k} dX_{1,k},$$

we have that

$$\frac{\partial \Pr(Z_{1,k}(\theta) \le z, X_{1,k} \le X)}{\partial z} = \int_{X_{1,k}\le X} f(z - X_{1,k}^\top\theta, X_{1,k}) dX_{1,k}$$

and

$$\frac{\partial^{p+1}P(Z_{1,k}(\theta)\le z, X_{1,k}\le X)}{\partial z\partial X}=f(z-X^\top\theta,X).$$

This implies that $v(z,X|\theta)=f(z-X^\top\theta,X)$. So

$$\begin{aligned}
\Phi_k(\theta)=&\int\mathcal{K}(t)\left(F_0(z+th_k-X_1^\top\Delta\theta)-F_0(z-X_2^\top\Delta\theta)\right)(X_1-X_2)\times\\
&\quad f(z+th_k-X_1^\top\theta,X_1)f(z-X_2^\top\theta,X_2)dtdzdX_1dX_2\\
=&\int\mathcal{K}(t)\left(F_0(z-X_1^\top\Delta\theta)-F_0(z-X_2^\top\Delta\theta)+\sum_{j=1}^{s_\mathcal{K}-1}\frac{\partial^jF_0(z-X_1^\top\Delta\theta)}{j!}t^jh_k^j\right.\\
&\quad\left.+\frac{\partial^{s_\mathcal{K}}F_0(\zeta_1)}{s_\mathcal{K}!}t^{s_\mathcal{K}}h_k^{s_\mathcal{K}}\right)\left(f\left(z-X_1^\top\theta,X_1\right)+\right.\\
&\quad\left.\sum_{j=1}^{s_\mathcal{K}-1}\frac{1}{j!}\frac{\partial^jf(z-X_1^\top\theta,X_1)}{\partial z^j}t^jh_k^j+\frac{1}{s_\mathcal{K}!}\frac{\partial^{s_\mathcal{K}}f(\zeta_2,X_1)}{\partial z^{s_\mathcal{K}}}t^{s_\mathcal{K}}h_k^{s_\mathcal{K}}\right)\times\\
&\quad(X_1-X_2)f(z-X_2^\top\theta,X_2)dtdzdX_1dX_2,
\end{aligned}$$

where ζ_1 lies somewhere between $z-X_1^\top\Delta\theta$ and $z+th_k-X_1^\top\Delta\theta$, and ζ_2 lies somewhere between $z-X_1^\top\theta$ and $z+th_k-X_1^\top\theta$. Since $\int\mathcal{K}(t)dt=1$, we have that

$$\begin{aligned}
\Phi_k(\theta)=\Phi(\theta)&+\int\mathcal{K}(t)\left(F_0(z-X_1^\top\Delta\theta)-F_0(z-X_2^\top\Delta\theta)\right)(X_1-X_2)f(z-X_2^\top\theta,X_2)\times\\
&\left(\sum_{j=1}^{s_\mathcal{K}-1}\frac{1}{j!}\frac{\partial^jf(z-X_1^\top\theta,X_1)}{\partial z^j}t^jh_k^j+\frac{1}{s_\mathcal{K}!}\frac{\partial^{s_\mathcal{K}}f(\zeta_2,X_1)}{\partial z^{s_\mathcal{K}}}t^{s_\mathcal{K}}h_k^{s_\mathcal{K}}\right)dtdzdX_1dX_2\\
&+\int\mathcal{K}(t)(X_1-X_2)f(z-X_1^\top\theta,X_1)f(z-X_2^\top\theta,X_2)\times\\
&\left(\sum_{j=1}^{s_\mathcal{K}-1}\frac{\partial^jF_0(z-X_1^\top\Delta\theta)}{j!}t^jh_k^j+\frac{\partial^{s_\mathcal{K}}F_0(\zeta_1)}{s_\mathcal{K}!}t^{s_\mathcal{K}}h_k^{s_\mathcal{K}}\right)dtdzdX_1dX_2\\
&+\int\mathcal{K}(t)(X_1-X_2)f(z-X_2^\top\theta,X_2)\left(\sum_{j=1}^{s_\mathcal{K}-1}\frac{\partial^jF_0(z-X_1^\top\Delta\theta)}{j!}t^jh_k^j+\frac{\partial^{s_\mathcal{K}}F_0(\zeta_1)}{s_\mathcal{K}!}t^{s_\mathcal{K}}h_k^{s_\mathcal{K}}\right)\times\\
&\left(\sum_{j=1}^{s_\mathcal{K}-1}\frac{1}{j!}\frac{\partial^jf(z-X_1^\top\theta,X_1)}{\partial z^j}t^jh_k^j+\frac{1}{s_\mathcal{K}!}\frac{\partial^{s_\mathcal{K}}f(\zeta_2,X_1)}{\partial z^{s_\mathcal{K}}}t^{s_\mathcal{K}}h_k^{s_\mathcal{K}}\right)dtdzdX_1dX_2.
\end{aligned}$$

Condition 1 and Condition 2 immediately lead to Lemma F.1(2) and Lemma F.1(1) is proved because $h_k\downarrow 0$.

To prove Lemma F.1(3), let $\mathbb{K}_{k,\theta}(W_1,W_2)=\mathcal{K}(h_k^{-1}(Z_1(\theta)-Z_2(\theta))(Y_1-Y_2)(X_1-X_2)$ denote the symmetric order-two kernel in (F.1). The exact Hoeffding covariance identity is

$$\begin{aligned}\mathrm{Cov}\{\Phi_k(\theta, W_k)\} = &\frac{4(B-2)}{B(B-1)}\mathrm{Cov}\{\mathbb{E}[\mathbb{K}_{k,\theta}(W_1,W_2)|W_1] - \mathbb{E}[\mathbb{K}_{k,\theta}(W_1,W_2)]\}\\ &+ \frac{2}{B(B-1)}\mathrm{Cov}\{\mathbb{K}_{k,\theta}(W_1,W_2)\}.\end{aligned}$$

In the calculations below, $\mathbb{E}_1$ denotes integration w.r.t. the first observation. Note that

$$\begin{aligned}&\mathbb{E}_1\left[\frac{1}{h_k}\mathcal{K}\left(\frac{Z_{1,k}(\theta)-Z_{2,k}(\theta)}{h_k}\right)(Y_{1,k}-Y_{2,k})(X_{1,k}-X_{2,k})\right]\\ &= \int \frac{1}{h_k}\mathcal{K}\left(\frac{z-Z_{2,k}(\theta)}{h_k}\right)\left[F_0(z-X_1^\top\Delta\theta)-Y_{2,k}\right](X_1-X_{2,k})\,f(z-X_1^\top\theta, X_1)dzdX_1\\ &= \int \mathcal{K}(t)\left(F_0\left(Z_{2,k}(\theta)+th_k-X_1^\top\Delta\theta\right)-Y_{2,k}\right)(X_1-X_{2,k})f(Z_{2,k}(\theta)+th_k-X_1^\top\theta, X_1)dtdX_1\\ &= \int \left(F_0\left(Z_{2,k}(\theta)-X_1^\top\Delta\theta\right)-Y_{2,k}\right)(X_1-X_{2,k})f(Z_{2,k}(\theta)-X_1^\top\theta, X_1)dX_1\\ &+ O\left(h_k^{s_{\mathcal{K}}}(1+|Y_{2,k}|)(1+\|X_{2,k}\|_2)\right).\end{aligned}$$

So the above term is bounded by $(1+|Y_{2,k}|)(1+\|X_{2,k}\|_2)$ up to some constant. This implies that

$$\begin{aligned}&\mathrm{var}\left(\mathbb{E}_1\left[\frac{1}{h_k}\cdot\mathcal{K}\left(\frac{Z_{1,k}(\theta)-Z_{2,k}(\theta)}{h_k}\right)(Y_{1,k}-Y_{2,k})(X_{1,k}-X_{2,k})\right]\right)\\ &\le C\mathbb{E}\left[(1+|Y_{2,k}|)^2(1+\|X_{2,k}\|_2)^2\right] < \infty\end{aligned}$$

due to the fact that $\sup_{x_0,X}\mathbb{E}(Y^2|x_0,X) \le \sup_{x_0,X} F_0^2(x_0+X^\top\theta_0) + \sup_{x_0,X}\sigma^2(x_0,X) < \infty$ and $\mathbb{E}\|X\|_2^2 < \infty$. To show the remaining parts, we note that $\Phi_k(\theta) = \Phi(\theta) + O(h_k^{s_{\mathcal{K}}})$ and $\Phi(\theta)$ is uniformly bounded. So

$$\begin{aligned}&\mathrm{var}\left(\frac{1}{h_k}\cdot\mathcal{K}\left(\frac{Z_{1,k}(\theta)-Z_{2,k}(\theta)}{h_k}\right)(Y_{1,k}-Y_{2,k})(X_{1,k}-X_{2,k})\right)\\ &= h_k^{-2}\mathbb{E}\left(\mathcal{K}^2\left(\frac{Z_{1,k}(\theta)-Z_{2,k}(\theta)}{h_k}\right)(Y_{1,k}-Y_{2,k})^2(X_{1,k}-X_{2,k})(X_{1,k}-X_{2,k})^\top\right) + O(1).\end{aligned}$$

Note that

$$
\begin{aligned}
&h_k^{-2}\mathbb{E}\left(\mathcal{K}^2\left(\frac{Z_{1,k}(\theta)-Z_{2,k}(\theta)}{h_k}\right)(Y_{1,k}-Y_{2,k})^2\left(X_{1,k}-X_{2,k}\right)\left(X_{1,k}-X_{2,k}\right)^\top\right)\\
&= h_k^{-2}\mathbb{E}\left(\mathcal{K}^2\left(\frac{Z_{1,k}(\theta)-Z_{2,k}(\theta)}{h_k}\right)\left(F_0(Z_{1,k}(\theta)-X_{1,k}^\top\Delta\theta)-F_0(Z_{2,k}(\theta)-X_{2,k}^\top\Delta\theta)\right)^2\right.\\
&\qquad\qquad\qquad\left.\times\left(X_{1,k}-X_{2,k}\right)\left(X_{1,k}-X_{2,k}\right)^\top\right)\\
&+ h_k^{-2}\mathbb{E}\left(\mathcal{K}^2\left(\frac{Z_{1,k}(\theta)-Z_{2,k}(\theta)}{h_k}\right)\left(\sigma^2(Z_{1,k}(\theta)-X_{1,k}^\top\theta,X_{1,k})+\sigma^2(Z_{2,k}(\theta)-X_{2,k}^\top\theta,X_{2,k})\right)\right.\\
&\qquad\qquad\left.\times\left(X_{1,k}-X_{2,k}\right)\left(X_{1,k}-X_{2,k}\right)^\top\right)\\
&= h_k^{-2}\int\mathcal{K}^2\left(\frac{z_1-z_2}{h_k}\right)\left(F_0(z_1-X_1^\top\Delta\theta)-F_0(z_2-X_2^\top\Delta\theta)\right)^2\left(X_1-X_2\right)\left(X_1-X_2\right)^\top\times\\
&\qquad\qquad\qquad\qquad f(z_1-X_1^\top\theta,X_1)f(z_2-X_2^\top\theta,X_2)dz_1dz_2dX_1dX_2\\
&+ h_k^{-2}\int\mathcal{K}^2\left(\frac{z_1-z_2}{h_k}\right)\left(\sigma^2(z_1-X_1^\top\theta,X_1)+\sigma^2(z_2-X_2^\top\theta,X_2)\right)\left(X_1-X_2\right)\left(X_1-X_2\right)^\top\times\\
&\qquad\qquad\qquad\qquad f(z_1-X_1^\top\theta,X_1)f(z_2-X_2^\top\theta,X_2)dz_1dz_2dX_1dX_2
\end{aligned}
$$

Following the previous proofs, we can show that the terms on the RHS of the last equality can be written as

$$
\begin{aligned}
&h_k^{-1}\int\mathcal{K}^2(t)dt\int\left(F_0(z-X_1^\top\Delta\theta)-F_0(z-X_2^\top\Delta\theta)\right)^2\left(X_1-X_2\right)\left(X_1-X_2\right)^\top\times\\
&\qquad\qquad\qquad\qquad f(z-X_1^\top\theta,X_1)f(z-X_2^\top\theta,X_2)dzdX_1dX_2\\
&+h_k^{-1}\int\mathcal{K}^2(t)dt\int\left(\sigma^2(z-X_1^\top\theta,X_1)+\sigma^2(z-X_2^\top\theta,X_2)\right)\left(X_1-X_2\right)\left(X_1-X_2\right)^\top\times\\
&\qquad\qquad\qquad\qquad f(z-X_1^\top\theta,X_1)f(z-X_2^\top\theta,X_2)dzdX_1dX_2+O(1).
\end{aligned}
$$

This proves the result.

For Lemma F.1(4), note that

$$
\nabla_\theta\Phi_k\left(\theta,\mathcal{W}_k\right)=\frac{1}{h_k^2B(B-1)}\sum_{i\neq j}^{B}\nabla\mathcal{K}\left(h_k^{-1}\left(Z_{i,k}(\theta)-Z_{j,k}(\theta)\right)\right)\left(Y_{i,k}-Y_{j,k}\right)\left(X_{i,k}-X_{j,k}\right)\left(X_{i,k}-X_{j,k}\right)^\top
$$

Again using the Hoeffding decomposition for a symmetric second-order U-statistic, the first projection and the degenerate remainder contribute, respectively,

$$\begin{aligned}
&\mathbb{E}\,\|\nabla_\theta \Phi_k(\theta, \mathcal{W}_k) - \nabla_\theta \Phi_k(\theta)\|_2^2 \\
&\quad \le \frac{C}{B} + \frac{C}{B(B-1)h_k^3} \int (\nabla \mathcal{K}(t))^2 \Big[\big\{F_0(z + th_k - X_1^\top \Delta\theta) - F_0(z - X_2^\top \Delta\theta)\big\}^2 \\
&\qquad\qquad\qquad\qquad\qquad\quad + \sigma^2(z + th_k - X_1^\top \theta, X_1) + \sigma^2(z - X_2^\top \theta, X_2)\Big]\, \|X_1 - X_2\|_2^4 \\
&\qquad\qquad\qquad\qquad\qquad\quad \times f(z + th_k - X_1^\top \theta, X_1) f(z - X_2^\top \theta, X_2)\, dt\, dz\, dX_1\, dX_2 \\
&\quad \le \frac{C}{B} + \frac{C}{B(B-1)h_k^3}.
\end{aligned}$$

The last integral is uniformly bounded by Condition 1 and Condition 2.

For part (5), write the centered U-statistic as a finite sum of centered ordered-pair kernels. Since B is fixed, the inequality $\|\sum_{r=1}^{B(B-1)} a_r\|_2^4 \le C_B \sum_r \|a_r\|_2^4$ reduces the claim to one ordered pair. Conditional fourth moments of $Y_1 - Y_2$ are uniformly bounded by Condition 1(iv), and the fourth moment of $X_1 - X_2$ is integrable by Condition 1(iii). Consequently, uniformly in θ,

$$\begin{aligned}
&\mathbb{E}\left[h_k^{-4} \mathcal{K}^4\left(\frac{Z_1(\theta) - Z_2(\theta)}{h_k}\right) |Y_1 - Y_2|^4 \|X_1 - X_2\|_2^4\right] \\
&\quad \le C h_k^{-4} \int \mathcal{K}^4\left(\frac{z_1 - z_2}{h_k}\right) (\|X_1\|_2^4 + \|X_2\|_2^4) f(z_1 - X_1^\top \theta, X_1) f(z_2 - X_2^\top \theta, X_2)\, dz_1 dz_2 dX_1 dX_2 \\
&\quad \le C h_k^{-3} \int \mathcal{K}^4(t)\, dt.
\end{aligned}$$

This proves part (5) and the lemma. □

Given Lemma F.1, we can briefly discuss the intuition of our algorithm. Note that (3.1) leads to

$$\begin{aligned}
\widehat{\theta}_k &= \widehat{\theta}_{k-1} + \gamma_k \cdot \Phi_k(\widehat{\theta}_{k-1}, \mathcal{W}_k) \\
&= \underbrace{\widehat{\theta}_{k-1} + \gamma_k \cdot \Phi(\widehat{\theta}_{k-1})}_{(I)} - \underbrace{\gamma_k \cdot \Big(\Phi(\widehat{\theta}_{k-1}) - \Phi_k(\widehat{\theta}_{k-1})\Big)}_{(II)} - \underbrace{\gamma_k \cdot \Big(\Phi_k(\widehat{\theta}_{k-1}) - \Phi_k(\widehat{\theta}_{k-1}, \mathcal{W}_k)\Big)}_{(III)}.
\end{aligned} \tag{F.6}$$

In the above update, term (III) has zero expectation conditioned on observations up to period $k-1$, and term (II) will vanish as k increases if we choose $h_k \downarrow 0$ according to Lemma F.1.

So we can regard the update (3.1) as being mainly driven by term (I). Define

$$\mathbb{H}(\theta,\tau) \equiv \int \nabla_z F_0(z - X_2^\top \Delta\theta - \tau(X_1 - X_2)^\top \Delta\theta) f(z - X_1^\top \theta, X_1) f(z - X_2^\top \theta, X_2) \times (X_1 - X_2)(X_1 - X_2)^\top \, dz dX_1 dX_2, \tag{F.7}$$

then for any θ, we obviously have that

$$\Phi(\theta) = -\left[\int_0^1 \mathbb{H}(\theta,\tau)d\tau\right]\Delta\theta \tag{F.8}$$

and

$$\lambda_{\min}\left(\int_0^1 \mathbb{H}(\theta,\tau)d\tau\right) \geq 0. \tag{F.9}$$

If additional conditions are imposed on the data generating process so that $\lambda_{\min}(\int_0^1 \mathbb{H}(\theta,\tau)d\tau) > 0$ for any θ, (I) is a contraction mapping and the same applies to the kernel-based warm-start update as $k \to \infty$. Consequently, our method guarantees global stability, meaning that we do not require any assumptions over the initial starting point of the update. Such strict lower boundedness of the smallest eigenvalue is formally stated in the following lemma. Note that the following lemma relies on the fact that all regressors are continuous; we relax this requirement in the following Section F.2.

Lemma F.2. *If Condition 1 holds, then there exists $C_{\Phi,4} > 0$ such that*

$$\sup_{\theta\in\mathbb{R}^p} \sup_{\tau\in[0,1]} \lambda_{max}(\mathbb{H}(\theta,\tau)) \leq C_{\Phi,4}.$$

If Condition 3 holds, then

$$\inf_{\tau\in[0,1]} \lambda_{\min}(\mathbb{H}(\theta,\tau)) \geq c_{\mathbb{H}} \cdot \left(\frac{\overline{z} - \underline{z}}{12(1 + \|\Delta\theta\|_2 + \|\theta_0\|_2)} \wedge r_X\right)^{2p+2},$$

where $c_{\mathbb{H}}$ is a positive constant that does not depend on the choice of θ.

Proof of Lemma F.2. Lemma F.2(i) is obvious if we note that

$$\begin{aligned}\lambda_{\max}(\mathbb{H}(\theta,\tau)) \leq \|\mathbb{H}(\theta,\tau)\|_F &\leq \|\nabla_z F_0\|_\infty \int \|X_1 - X_2\|_2^2 f(z - X_1^\top\theta, X_1) f(z - X_2^\top\theta, X_2) dz dX_1 dX_2 \\ &\leq 4\|\nabla_z F_0\|_\infty \int \|X_1\|_2^2 f(z - X_1^\top\theta, X_1) f(z - X_2^\top\theta, X_2) dz dX_1 dX_2 \\ &\leq 4\|\nabla_z F_0\|_\infty \int \|X_1\|_2^2 g_0(X_1) dX_1 \int f(z - X_2^\top\theta, X_2) dz dX_2\end{aligned}$$

which is uniformly bounded according to Condition 1.

To prove Lemma F.2(ii), consider the choices of z, X_1, and X_2 such that

$$\frac{3\underline{z}+\overline{z}}{4} \leq z \leq \frac{\underline{z}+3\overline{z}}{4},\ \|X_1\|_2 \leq \frac{\overline{z}-\underline{z}}{12(1+\|\Delta\theta\|_2+\|\theta_0\|_2)} \wedge r_X,\ \|X_2\|_2 \leq \frac{\overline{z}-\underline{z}}{12(1+\|\Delta\theta\|_2+\|\theta_0\|_2)} \wedge r_X.$$

In this case,

$$z - X_2^\top \Delta\theta - \tau\left(X_1 - X_2\right)^\top \Delta\theta - \underline{z} \geq \frac{\overline{z}-\underline{z}}{4} - 2\|X_2\|_2\|\Delta\theta\|_2 - \|X_1\|_2\|\Delta\theta\|_2 \geq 0,$$

and

$$z - X_2^\top \Delta\theta - \tau\left(X_1 - X_2\right)^\top \Delta\theta - \overline{z} \leq -\frac{\overline{z}-\underline{z}}{4} + 2\|X_2\|_2\|\Delta\theta\|_2 + \|X_1\|_2\|\Delta\theta\|_2 \leq 0,$$

so $\nabla F_0\left(z - X_2^\top \Delta\theta - \tau\left(X_1 - X_2\right)^\top \Delta\theta\right) \geq \underline{c}_F$ according to Condition 3. On the other side, for z, X_1, X_2 satisfying the above requirement, we also have that for $i = 1, 2$, $z - X_i^\top\theta \geq z - \|X_i\|_2\|\theta\|_2 \geq z - \|X_i\|_2(\|\theta_0\|_2 + \|\Delta\theta\|_2) \geq \underline{z}$ and similarly $z - X_i^\top\theta \leq \overline{z}$. This implies that $f(z - X_i^\top\theta, X_i) \geq \underline{c}_f$ according again to Condition 3.

Then define area $\Omega = \{(z, X_1, X_2) : \frac{3\underline{z}+\overline{z}}{4} \leq z \leq \frac{\underline{z}+3\overline{z}}{4}, \|X_1\|_2 \leq \frac{\overline{z}-\underline{z}}{12(1+\|\Delta\theta\|_2+\|\theta_0\|_2)} \wedge r_X, \|X_2\|_2 \leq \frac{\overline{z}-\underline{z}}{12(1+\|\Delta\theta\|_2+\|\theta_0\|_2)} \wedge r_X\}$. We have that

$$\begin{aligned}
&\lambda_{\min}(\mathbb{H}(\theta,\tau)) \\
&\geq \lambda_{\min}\left(\int_\Omega \partial F_0\left(z - X_2^\top\Delta\theta - \tau\left(X_1-X_2\right)^\top\Delta\theta\right) f(z - X_1^\top\theta, X_1) f(z - X_2^\top\theta, X_2)\right. \\
&\qquad\qquad \left. \times \left(X_1 - X_2\right)\left(X_1-X_2\right)^\top dz dX_1 dX_2\right) \\
&\geq \underline{c}_F \underline{c}_f^2 \lambda_{\min}\left(\int_\Omega \left(X_1 - X_2\right)\left(X_1-X_2\right)^\top dz dX_1 dX_2\right) \\
&= (\overline{z}-\underline{z})\underline{c}_F\underline{c}_f^2 c_p \cdot \left(\frac{\overline{z}-\underline{z}}{12(1+\|\Delta\theta\|_2+\|\theta_0\|_2)} \wedge r_X\right)^p \cdot \lambda_{\min}\left(\int_{\|X\|_2 \leq \frac{\overline{z}-\underline{z}}{12(1+\|\Delta\theta\|_2+\|\theta_0\|_2)} \wedge r_X} XX^\top dX\right) \\
&= (\overline{z}-\underline{z})\underline{c}_F\underline{c}_f^2 c_p \cdot \left(\frac{\overline{z}-\underline{z}}{12(1+\|\Delta\theta\|_2+\|\theta_0\|_2)} \wedge r_X\right)^p \cdot \int_{\|X\|_2 \leq \frac{\overline{z}-\underline{z}}{12(1+\|\Delta\theta\|_2+\|\theta_0\|_2)} \wedge r_X} x_1^2 dX \\
&= \frac{c_p^2(\overline{z}-\underline{z})\underline{c}_F\underline{c}_f^2}{p+2}\left(\frac{\overline{z}-\underline{z}}{12(1+\|\Delta\theta\|_2+\|\theta_0\|_2)} \wedge r_X\right)^{2p+2},
\end{aligned}$$

where c_p is the volume of p-dimensional unit ball. This shows the result. $\square$

The lower bound of the smallest eigenvalue as a function of θ in Lemma F.2 is crucial because, for all sufficiently large k, it makes term (I) in (F.6) a strict contraction. The same result also guarantees point identification of θ_0.

F.2. The Case with Discrete Regressors

This section verifies the curvature and identification ingredient needed to extend the preceding global-stability argument when X contains both continuous and discrete components. The stochastic convergence conclusions also require the corresponding moment, noise, kernel, and step-size assumptions from Condition 1, Condition 2, and Condition 4; the lemma below does not replace those assumptions. We require throughout that x_0 be continuously distributed. Decompose $X = (X_c^\top, X_d^\top)^\top$, where $X_c \in \mathbb{R}^{p_c}$ collects the continuous regressors and $X_d \in \mathcal{X}_d \subseteq \mathbb{R}^{p_d}$ the discrete ones, $p = p_c + p_d$, with $\mathcal{X}_d$ countable. Let $p(x_d) = \mathbb{P}(X_d = x_d)$, and let $f(\cdot, \cdot \mid x_d)$ denote the conditional joint density of $(Z_0, X_c^\top)^\top$ given $X_d = x_d$, where $Z_0 = x_0 + X^\top\theta_0$ is the true index.

Define

$$\begin{aligned}
\mathbb{H}(\theta,\tau) \;=\; \sum_{X_{d,1}\in\mathcal{X}_d}\sum_{X_{d,2}\in\mathcal{X}_d} p(X_{d,1})\,p(X_{d,2}) \int \nabla_z F_0\big(z - X_2^\top\Delta\theta - \tau(X_1 - X_2)^\top\Delta\theta\big)& \\
\times\; f(z - X_1^\top\Delta\theta, X_{c,1} \mid X_{d,1})\, f(z - X_2^\top\Delta\theta, X_{c,2} \mid X_{d,2})& \\
\times\,(X_1 - X_2)(X_1 - X_2)^\top\, dz\, dX_{c,1}\, dX_{c,2},&
\end{aligned} \tag{F.10}$$

indexed by $\tau \in [0,1]$. The object of this section is a strictly positive lower bound on the minimum eigenvalue of $\mathbb{H}(\theta,\tau)$, which substitutes for the corresponding step in the continuous-regressor proofs.

Condition 1′. *(i) F_0 is bounded and has uniformly bounded derivatives up to order $s_\mathcal{K}$; (ii) for each $x_d \in \mathcal{X}_d$, $f(\cdot, X_c \mid x_d)$ has partial derivatives with respect to its first argument up to order $s_\mathcal{K}$, and these are uniformly bounded over $x_d \in \mathcal{X}_d$; (iii) there exist non-negative functions $g_0, \ldots, g_{s_\mathcal{K}+1}$ on $\mathbb{R}^{p_c}$ such that, uniformly over $x_d \in \mathcal{X}_d$ and $0 \le j \le s_\mathcal{K} + 1$, $\sup_{z\in\mathbb{R}} |\partial^j f(z, X_c \mid x_d)/\partial z^j| \le g_j(X_c)$ and $\int (1 + \|X_c\|_2^4)\, g_j(X_c)\, dX_c < \infty$; moreover, $\mathbb{E}\|X\|_2^4 < \infty$; (iv) $\sup_{x_0,X} \mathbb{E}(\varepsilon^4 \mid x_0, X) < \infty$ is uniformly bounded and $\sigma^2(x_0, X)$ has a uniformly bounded derivative with respect to x_0.*

Condition 3′. *(i) $\nabla_z F_0(z) > 0$ for all $z \in \mathbb{R}$; (ii) there exist finitely many pairs of support* *points* $(x_d^{(l)}, \tilde{x}_d^{(l)}) \in \mathcal{X}_d \times \mathcal{X}_d$*,* $l = 1, \dots, L$*, each with* $p(x_d^{(l)}) > 0$ *and* $p(\tilde{x}_d^{(l)}) > 0$*, such that*

$$\operatorname{span}\left\{ x_d^{(1)} - \tilde{x}_d^{(1)}, \ \dots, \ x_d^{(L)} - \tilde{x}_d^{(L)} \right\} \ = \ \mathbb{R}^{p_d};$$

(iii) there exists $r_X > 0$ *such that, for every support point* x_d *appearing in the pairs of part (ii), the map* $(z, X_c) \mapsto f(z, X_c \mid x_d)$ *is continuous and, for every compact* $I \subset \mathbb{R}$*,*

$$\inf_{z \in I, \ \|X_c\|_2 \le r_X} f(z, X_c \mid x_d) \ > \ 0.$$

Remark F.1. *Condition 3′(ii) is the natural discrete-regressor analogue of the full-rank requirement in the continuous case, and is in the spirit of the identification conditions for maximum-rank-correlation estimators (Han 1987). It accommodates mutually exclusive dummy variables: if* X_d *consists of group indicators with an omitted reference category (as with the race and stop-reason dummies in Stanford Open Policing Project application in the main text), the pairs* $(e_j, \mathbf{0}_{p_d})$*,* $j = 1, \dots, p_d$ *— "group* j *versus the reference group" — easily satisfy the spanning requirement whenever each category has positive probability.*

Lemma F.3. *Suppose Conditions 1′ and 3′ hold and* x_0 *is continuously distributed given* X*. Then for every* $\theta \in \mathbb{R}^p$*,*

$$\inf_{\tau \in [0,1]} \lambda_{\min}\big(\mathbb{H}(\theta, \tau)\big) \ > \ 0.$$

Moreover, for every compact $\Theta \subset \mathbb{R}^p$*,* $\inf_{\theta \in \Theta} \inf_{\tau \in [0,1]} \lambda_{\min}(\mathbb{H}(\theta, \tau)) > 0$*.*

Proof. We first verify that $\mathbb{H}(\theta, \tau)$ is finite and jointly continuous in (θ, τ). Each entry of the integrand in (F.10) is bounded in absolute value by $\|\nabla_z F_0\|_\infty \, \|X_1 - X_2\|_2^2 \, f_1 f_2$ with $f_l := f(z - X_l^\top \Delta\theta, X_{c,l} \mid X_{d,l})$. Integrating z first against f_2 for fixed $(X_{c,2}, X_{d,2})$ gives the marginal conditional density value $f_{X_c|X_d}(X_{c,2} \mid X_{d,2})$, while $f_1 \le g_0(X_{c,1})$ by Condition 1′(iii). Hence, by $\|X_1 - X_2\|_2^2 \le 2\|X_1\|_2^2 + 2\|X_2\|_2^2$, the total mass is bounded by

$$C\|\nabla_z F_0\|_\infty \left[\int (1 + \|X_c\|_2^2) \, g_0(X_c) \, dX_c + \mathbb{E}\|X\|_2^2 + \sum_{x_d} p(x_d) \|x_d\|_2^2 \right] < \infty,$$

using Condition 1′(iii) and $\mathbb{E}\|X\|_2^4 < \infty$ (which implies $\sum_{x_d} p(x_d)\|x_d\|_2^2 < \infty$). The integrand is continuous in (θ, τ) for almost every $(z, X_{c,1}, X_{c,2})$ and every $(X_{d,1}, X_{d,2})$, by continuity of $\nabla_z F_0$ (Condition 1′(i)) and of f in its first argument (Condition 1′(ii)); the display above provides an integrable envelope that is locally uniform in θ. Dominated convergence therefore yields joint continuity of $(\theta, \tau) \mapsto \mathbb{H}(\theta, \tau)$.

Now fix θ and $\tau \in [0,1]$, and let $v = (v_c^\top, v_d^\top)^\top \in \mathbb{R}^p$ with $\|v\|_2 = 1$. Since $\nabla_z F_0 > 0$ and all densities and weights are nonnegative, every term of the double sum in $v^\top \mathbb{H}(\theta,\tau)v$ is nonnegative, so

$$v^\top \mathbb{H}(\theta,\tau)v \;\geq\; p(x_d)\, p(\tilde{x}_d) \int_{\mathcal{R}} \nabla_z F_0(\zeta)\, f_1 f_2 \left[(X_1 - X_2)^\top v\right]^2 dz\, dX_{c,1}\, dX_{c,2} \tag{F.11}$$

for any fixed pair $(x_d, \tilde{x}_d)$ of support points and any measurable region $\mathcal{R}$, where $\zeta := z - X_2^\top \Delta\theta - \tau(X_1 - X_2)^\top \Delta\theta$, $X_1 = (X_{c,1}^\top, x_d^\top)^\top$, $X_2 = (X_{c,2}^\top, \tilde{x}_d^\top)^\top$. Take $\mathcal{R} = \mathcal{R}(r) := \{(z, X_{c,1}, X_{c,2}) : z \in [-1,1],\ \|X_{c,1}\|_2 \leq r,\ \|X_{c,2}\|_2 \leq r\}$ for an $r \in (0, r_X]$ to be chosen. On $\mathcal{R}(r)$, the argument ζ lies in the compact interval

$$\mathcal{A}_\theta := [-1 - (r + M_d)\|\Delta\theta\|_2 - 2(r + M_d)\|\Delta\theta\|_2,\ 1 + (r + M_d)\|\Delta\theta\|_2 + 2(r + M_d)\|\Delta\theta\|_2]$$

for all $\tau \in [0,1]$, where $M_d := \max_l \max\{\|x_d^{(l)}\|_2, \|\tilde{x}_d^{(l)}\|_2\}$. By Condition 3′(i) and continuity of $\nabla_z F_0$,

$$c_0(\theta) := \inf_{u \in \mathcal{A}_\theta} \nabla_z F_0(u) > 0,$$

and this bound holds uniformly over $\tau \in [0,1]$ because $\zeta \in \mathcal{A}_\theta$ for every τ. Similarly, the arguments $z - X_l^\top \Delta\theta$ of f_1, f_2 lie in a compact interval I_θ on $\mathcal{R}(r)$, so Condition 3′(iii) gives

$$\delta(\theta) := \min_{x_d \in \{x_d^{(l)}, \tilde{x}_d^{(l)} : l \leq L\}} \ \inf_{z \in I_\theta,\, \|X_c\|_2 \leq r_X} f(z, X_c \mid x_d) > 0.$$

Now consider v with $v_d \neq 0$. By Condition 3′(ii), there exists l with $a := (x_d^{(l)} - \tilde{x}_d^{(l)})^\top v_d \neq 0$; take $(x_d, \tilde{x}_d) = (x_d^{(l)}, \tilde{x}_d^{(l)})$ in (F.11) and set $r = \min\{r_X, |a|/4\}$. On $\mathcal{R}(r)$,

$$\left|(X_1 - X_2)^\top v\right| \;=\; \left|(X_{c,1} - X_{c,2})^\top v_c + a\right| \;\geq\; |a| - 2r\|v_c\|_2 \;\geq\; |a| - 2r \;\geq\; \frac{|a|}{2},$$

so $[(X_1 - X_2)^\top v]^2 \geq a^2/4$ pointwise. Consequently, we have that

$$v^\top \mathbb{H}(\theta,\tau)v \;\geq\; p(x_d^{(l)})\, p(\tilde{x}_d^{(l)})\, c_0(\theta)\, \delta(\theta)^2\, \frac{a^2}{4} \cdot 2\,\mathrm{vol}(B_r)^2 \;>\; 0,$$

where $\mathrm{vol}(B_r)$ is the volume of the r-ball in $\mathbb{R}^{p_c}$ and the factor 2 is the length of $[-1,1]$.

We finally consider the case of v with $v_d = 0$, $v_c \neq 0$, so $\|v_c\|_2 = 1$. Take any single support point from the pairs, say $x_d = \tilde{x}_d = x_d^{(1)}$, and $r = r_X$ in (F.11). Then $(X_1 - X_2)^\top v = (X_{c,1} - X_{c,2})^\top v_c$, and by symmetry of $B_r \times B_r$ (the cross term vanishes and the ball is isotropic),

$$\int_{B_r}\int_{B_r} \left[(X_{c,1}-X_{c,2})^\top v_c\right]^2 dX_{c,1}\, dX_{c,2} \;=\; 2\,\mathrm{vol}(B_r)\int_{B_r}(u^\top v_c)^2\, du \;=\; 2\,\mathrm{vol}(B_r)\,\kappa(r)\,\|v_c\|_2^2 \;=:\; c_1(r) > 0,$$

with $\kappa(r) = \int_{B_r}(u^\top e)^2 du > 0$ independent of the unit vector e. Hence

$$v^\top \mathbb{H}(\theta,\tau)v \;\geq\; p(x_d^{(1)})^2\, c_0(\theta)\,\delta(\theta)^2\; \cdot 2\, c_1(r_X) \;>\; 0.$$

The above discussion shows that $v^\top \mathbb{H}(\theta,\tau)v > 0$ for every unit v, i.e., $\mathbb{H}(\theta,\tau)$ is positive definite; since it is a finite symmetric matrix, $\lambda_{\min}(\mathbb{H}(\theta,\tau)) > 0$. Finally, to show the uniform boundedness, we only need to note that $(\theta,\tau) \mapsto \lambda_{\min}(\mathbb{H}(\theta,\tau))$ is continuous on any compact set. □

Remark F.2. *As in the continuous case, Condition 3′(i) and the full-support requirement in Condition 3′(iii) (positivity of $f(\cdot, X_c \mid x_d)$ for all $z \in \mathbb{R}$) supply the curvature condition used for* global *stability and are stronger than what point identification requires. Together with the remaining stochastic assumptions stated above, they permit the continuous-regressor stability proof to be repeated conditional on the discrete support points.*

F.3. Convergence Rate of Raw Iterate $\widehat{\theta}_N$

This section gives the sharp rate and distribution of $\widehat{\theta}_N$, which will be frequently used in the proof of Theorem B.1. Consider the following decomposition of $\mathbb{H}_0$:

$$\mathbb{H}_0 = \mathcal{P}_{\mathbb{H}_0}^\top \Lambda_{\mathbb{H}_0}\mathcal{P}_{\mathbb{H}_0}, \qquad \mathcal{P}_{\mathbb{H}_0}\mathbb{H}_0\mathcal{P}_{\mathbb{H}_0}^\top = \Lambda_{\mathbb{H}_0}.$$

Choose a deterministic k_0 so large that $0 < \gamma_k \lambda_{\max}(\mathbb{H}_0) < 1$ for $k \geq k_0$, and define

$$A_{k_0-1} = \mathbb{I}_p, \qquad A_k = \prod_{\ell=k_0}^{k}(\mathbb{I}_p - \gamma_\ell \mathbb{H}_0)^{-1}, \qquad \zeta_{0,k} = \Phi_k(\theta_0, \mathcal{W}_k) - \Phi_k(\theta_0).$$

Define

$$\Sigma_N = \mathrm{Var}\left(\sum_{k=k_0}^{N} \gamma_k \mathcal{P}_{\mathbb{H}_0} A_k \zeta_{0,k}\right), \qquad s_{N,j}^2 = e_j^\top \Sigma_N e_j, \qquad D_N = \mathrm{Diag}(s_{N,1},\ldots,s_{N,p}).$$

We have the following results.

Theorem F.1. *Let Condition 1–4 hold. Then for every fixed coordinate j,*

$$\limsup_{N\to\infty} \frac{e_j^\top \mathcal{P}_{\mathbb{H}_0} A_N \Delta\widehat{\theta}_N}{\sqrt{2s_{N,j}^2 \log\log(s_{N,j}^2)}} = 1, \qquad \liminf_{N\to\infty} \frac{e_j^\top \mathcal{P}_{\mathbb{H}_0} A_N \Delta\widehat{\theta}_N}{\sqrt{2s_{N,j}^2 \log\log(s_{N,j}^2)}} = -1 \quad a.s. \tag{F.12}$$

Theorem F.1 provides the sharp rate and coordinate-wise LIL under Conditions 1–4. In particular, Lemma F.5 gives, for every fixed j,

$$s_{N,j}^2 \asymp N^{-\alpha_\gamma+\alpha_h} \exp\left(\frac{2\gamma_0\lambda_j}{1-\alpha_\gamma} N^{1-\alpha_\gamma}\right), \qquad e_j^\top \mathcal{P}_{\mathbb{H}_0} A_N \mathcal{P}_{\mathbb{H}_0}^\top e_j \asymp \exp\left(\frac{\gamma_0\lambda_j}{1-\alpha_\gamma} N^{1-\alpha_\gamma}\right).$$

Thus (F.12) implies $\|\Delta\widehat{\theta}_N\|_2 = O_{\text{a.s.}}(N^{-(\alpha_\gamma-\alpha_h)/2}\sqrt{\log N})$. Detailed supporting lemmas follow in Section F.3.1.

F.3.1. Development of Theorem F.1

This section provides detailed development of Theorem F.1. Define

$$\begin{aligned} V_N =& \gamma_N(\Phi_N(\widehat{\theta}_{N-1}, \mathcal{W}_N) - \Phi_N(\widehat{\theta}_{N-1})) + \gamma_{N-1}(\mathbb{I}_p - \gamma_N\mathbb{H}_0)(\Phi_{N-1}(\widehat{\theta}_{N-2}, \mathcal{W}_{N-1}) - \Phi_{N-1}(\widehat{\theta}_{N-2})) \\ &+ \cdots + \gamma_1 (\mathbb{I}_p - \gamma_N\mathbb{H}_0)\cdots(\mathbb{I}_p - \gamma_2\mathbb{H}_0)(\Phi_1(\widehat{\theta}_0, \mathcal{W}_1) - \Phi_1(\widehat{\theta}_0)). \end{aligned}$$

The following lemma describes a preliminary a.s. convergence rate of $\|V_N\|_2^2$.

Lemma F.4. *Let Condition 1–Condition 4 hold, then for any $r > 1$ we have that*

$$\|V_N\|_2^2 = o\left(N^{-2\alpha_\gamma+\alpha_h+1}\log^r(N)\right), \ a.s.$$

Proof of Lemma F.4. According to Lemma F.2, we have that $\lambda_{\min}(\mathbb{H}_0) > 0$. Note that

$$V_N = \gamma_N(\Phi_N(\widehat{\theta}_{N-1}, \mathcal{W}_N) - \Phi_N(\widehat{\theta}_{N-1})) + (\mathbb{I}_p - \gamma_N\mathbb{H}_0)V_{N-1},$$

and $\mathbb{E}_{N-1}[V_{N-1}^\top(\Phi_N(\widehat{\theta}_{N-1}, \mathcal{W}_N) - \Phi_N(\widehat{\theta}_{N-1}))] = 0$. Then according to Lemma F.1(3), we have that

$$\begin{aligned} \mathbb{E}_{N-1}\|V_N\|_2^2 &= \|(\mathbb{I}_p - \gamma_N\mathbb{H}_0)V_{N-1}\|_2^2 + \gamma_N^2\mathbb{E}_{N-1}\|\Phi_N(\widehat{\theta}_{N-1}, W_N) - \Phi_N(\widehat{\theta}_{N-1})\|_2^2 \\ &\leq (1 - C\gamma_N)\|V_{N-1}\|_2^2 + C\gamma_N^2 h_N^{-1} \leq \left(1 - CN^{-\alpha_\gamma}\right)\|V_{N-1}\|_2^2 + CN^{-2\alpha_\gamma+\alpha_h}. \end{aligned}$$

Define $a_N = N^{2\alpha_\gamma - \alpha_h - 1} \log^{-r}(N)$ for any $r > 1$, we have

$$\mathbb{E}_{N-1}\left[a_N \|V_N\|_2^2\right] \leq \frac{a_N}{a_{N-1}}\left(1 - CN^{-\alpha_\gamma}\right)\left[a_{N-1}\|V_{N-1}\|_2^2\right] + C(N\log^r(N))^{-1}.$$

When $2\alpha_\gamma - \alpha_h - 1 > 0$, for N sufficiently large we have

$$\frac{N^{2\alpha_\gamma-\alpha_h-1}}{(N-1)^{2\alpha_\gamma-\alpha_h-1}} = \left(1 + \frac{1}{N-1}\right)^{2\alpha_\gamma-\alpha_h-1} \leq 1 + \frac{C}{N},$$

and

$$\frac{\log^r(N)}{\log^r(N-1)} \leq \left(1 + \frac{1}{(N-1)\log(N-1)}\right)^r \leq 1 + \frac{C}{N}.$$

So

$$\frac{a_N}{a_{N-1}}\left(1 - CN^{-\alpha_\gamma}\right) \leq \left(1 + CN^{-1}\right)\left(1 - CN^{-\alpha_\gamma}\right) \leq 1 - CN^{-\alpha_\gamma}$$

for N sufficiently large, then we have that for k sufficiently large,

$$\mathbb{E}_{N-1}\left[a_N \|V_N\|_2^2\right] \leq \left(1 - CN^{-\alpha_\gamma}\right)\left[a_{N-1}\|V_{N-1}\|_2^2\right] + C(N\log^r(N))^{-1}.$$

Since $r > 1$, $\sum_{N=1}^{\infty}(N\log^r(N))^{-1} < \infty$, thus we have that $a_N \|V_N\|_2^2$ a.s. converges to a finite random variable by Theorem 1 of Robbins and Siegmund (1971). Furthermore, we have that $\sum_{N=1}^{\infty} N^{-\alpha_\gamma}\left[a_{N-1}\|V_{N-1}\|_2^2\right] < \infty$ a.s. holds. This shows that $a_{N-1}\|V_{N-1}\|_2^2 \to 0$ a.s. holds, which proves the result. □

As will be seen later, the convergence rate provided in Lemma F.4 is quite rough, but it will be used to provide a preliminary convergence rate for $\widehat{\theta}_N$. To do this, we first provide a technical lemma.

Lemma F.5. *Let $\frac{1}{2} < \alpha < 1$ and $0 < C_0 < 1$, then (i) the following exists*

$$\lim_{j\to\infty} \frac{\prod_{l=1}^{j}(1 - C_0 l^{-\alpha})}{\exp\left(-\frac{C_0}{1-\alpha} j^{1-\alpha}\right)};$$

(ii) further, let $\varsigma > 0$, and let $b(x)$ be a strictly positive, continuously differentiable function for all sufficiently large x such that $b(x)\exp(\frac{\varsigma C_0}{1-\alpha}x^{1-\alpha})$ is increasing for $x \geq x^*$, $\lim_{x\to\infty} b'(x)x^\alpha/b(x) = 0$, and $\int_1^k (x^{\alpha-1}b(x) + x^\alpha b'(x))\exp(\frac{\varsigma C_0}{1-\alpha}x^{1-\alpha})dx \uparrow \infty$, then the following limit exists:

$$\lim_{k\to\infty} \frac{\sum_{j=1}^{k} b(j)\left(\prod_{l=1}^{j}(1 - C_0 l^{-\alpha})\right)^{-\varsigma}}{k^\alpha b(k)\exp\left(\frac{\varsigma C_0}{1-\alpha}k^{1-\alpha}\right)}.$$

Proof of Lemma F.5. To show (i), note that for $x \in (0,1)$, $\log(1-x) \le -x$. So sequence

$$\log\left(\prod_{l=1}^{j}(1-C_0 l^{-\alpha})\right) + C_0\sum_{l=1}^{j} l^{-\alpha} = \sum_{l=1}^{j}\left(\log(1-C_0 l^{-\alpha}) + C_0 l^{-\alpha}\right)$$

is non-increasing with respect to j. On the other side, Taylor expansion leads to

$$\left|\log\left(\prod_{l=1}^{j}(1-C_0 l^{-\alpha})\right) + C_0\sum_{l=1}^{j} l^{-\alpha}\right| \le C\sum_{l=1}^{j} l^{-2\alpha},$$

where the RHS is upper bounded uniformly w.r.t. j because $2\alpha > 1$. This implies that

$$\lim_{j\to\infty}\left[\log\left(\prod_{l=1}^{j}(1-C_0 l^{-\alpha})\right) + C_0\sum_{l=1}^{j} l^{-\alpha}\right]$$

exists. Also note that $\sum_{l=1}^{j} l^{-\alpha} - \frac{1}{1-\alpha}j^{1-\alpha}$ is bounded and decreasing, so the limit also exists. This leads to that

$$\lim_{j\to\infty}\left[\log\left(\prod_{l=1}^{j}(1-C_0 l^{-\alpha})\right) + \frac{C_0}{1-\alpha}j^{1-\alpha}\right]$$

exists. Denote such limit as C_1, we have that

$$\lim_{j\to\infty}\exp\left(\log\left(\prod_{l=1}^{j}(1-C_0 l^{-\alpha})\right) + \frac{C_0}{1-\alpha}j^{1-\alpha}\right) = \exp(C_1).$$

This proves the result.

To prove (ii), note that from (i), for any fixed ς,

$$\delta_j \equiv \frac{\left(\prod_{l=1}^{j}(1-C_0 l^{-\alpha})\right)^{-\varsigma}}{\exp\left(\frac{\varsigma C_0}{1-\alpha}j^{1-\alpha}\right)} \to \delta_\infty$$

so

$$\sum_{j=1}^{k} b(j)\left(\prod_{l=1}^{j}(1-C_0 l^{-\alpha})\right)^{-\varsigma} = \delta_\infty\sum_{j=1}^{k} b(j)\exp\left(\frac{\varsigma C_0}{1-\alpha}j^{1-\alpha}\right) + \sum_{j=1}^{k}(\delta_j-\delta_\infty)b(j)\exp\left(\frac{\varsigma C_0}{1-\alpha}j^{1-\alpha}\right).$$

Since $b(x)\exp(\frac{\varsigma C_0}{1-\alpha}x^{1-\alpha})$ eventually increases for sufficiently large x, we have that

$$\frac{\sum_{j=1}^{k}(\delta_j-\delta_\infty)b(j)\exp\left(\frac{\varsigma C_0}{1-\alpha}j^{1-\alpha}\right)}{\sum_{j=1}^{k}b(j)\exp\left(\frac{\varsigma C_0}{1-\alpha}j^{1-\alpha}\right)}\to 0,$$

and we only need to look at the ratio

$$\frac{\sum_{j=1}^{k}b(j)\exp\left(\frac{\varsigma C_0}{1-\alpha}j^{1-\alpha}\right)}{k^\alpha b(k)\exp(\frac{\varsigma C_0}{1-\alpha}k^{1-\alpha})}.$$

Note that for $k\geq x^*$, $b(x)\exp\left(\frac{\varsigma C_0}{1-\alpha}x^{1-\alpha}\right)$ is increasing, so

$$\begin{aligned}\int_{[x^*]}^{k}b(x)\exp\left(\frac{\varsigma C_0}{1-\alpha}x^{1-\alpha}\right)dx &\leq \sum_{j=[x^*]+1}^{k}b(j)\exp\left(\frac{\varsigma C_0}{1-\alpha}j^{1-\alpha}\right)\\ &\leq \int_{[x^*]}^{k}b(x)\exp\left(\frac{\varsigma C_0}{1-\alpha}x^{1-\alpha}\right)dx+b(k)\exp\left(\frac{\varsigma C_0}{1-\alpha}k^{1-\alpha}\right).\end{aligned}$$

For $\int_{[x^*]}^{k}b(x)\exp(\frac{\varsigma C_0}{1-\alpha}x^{1-\alpha})dx$, integration by part leads to

$$\begin{aligned}\int_{[x^*]}^{k}b(x)\exp\left(\frac{\varsigma C_0}{1-\alpha}x^{1-\alpha}\right)dx &= \int_{[x^*]}^{k}\frac{x^\alpha b(x)}{\varsigma C_0}\varsigma C_0 x^{-\alpha}\exp\left(\frac{\varsigma C_0}{1-\alpha}x^{1-\alpha}\right)dx\\ &= \frac{k^\alpha b(k)}{\varsigma C_0}\exp\left(\frac{\varsigma C_0}{1-\alpha}k^{1-\alpha}\right)-C(x^*)\\ &\quad -\frac{1}{\varsigma C_0}\int_{[x^*]}^{k}(\alpha x^{\alpha-1}b(x)+x^\alpha b'(x))\exp\left(\frac{\varsigma C_0}{1-\alpha}x^{1-\alpha}\right)dx,\end{aligned}$$

where $C(x^*)$ is a constant depending only on x^*. Using L'Hospital's rule, we have that

$$\lim_{k\to\infty}\frac{\int_{[x^*]}^{k}(\alpha x^{\alpha-1}b(x)+x^\alpha b'(x))\exp\left(\frac{\varsigma C_0}{1-\alpha}x^{1-\alpha}\right)dx}{k^\alpha b(k)\exp(\frac{\varsigma C_0}{1-\alpha}k^{1-\alpha})}=\lim_{k\to\infty}\frac{\alpha k^{\alpha-1}b(k)+k^\alpha b'(k)}{\alpha k^{\alpha-1}b(k)+k^\alpha b'(k)+\varsigma C_0 q(k)}=0.$$

So the following exists

$$\lim_{k\to\infty}\frac{\int_{[x^*]}^{k}b(x)\exp\left(\frac{\varsigma C_0}{1-\alpha}x^{1-\alpha}\right)dx}{k^\alpha b(k)\exp\left(\frac{\varsigma C_0}{1-\alpha}k^{1-\alpha}\right)}.$$

This shows the result. □

Combine Lemma F.4 and Lemma F.5, we can immediately provide an a.s. convergence rate for $\|\Delta\widehat{\theta}_N\|_2^2$. But to highlight how the convergence rate of $\|\Delta\widehat{\theta}_N\|_2^2$ can be accelerated as we will do later, we first provide the following general results.

Lemma F.6. *Let Condition 1–Condition 4 hold. Then for any $b(x) \downarrow 0$ satisfying all the requirements in Lemma F.5 with $\alpha = \alpha_\gamma$, if $\|V_N\|_2^2 = O(b(N))$ a.s. holds, there holds*

$$\|\Delta\widehat{\theta}_N\|_2^2 = O(b(N) \vee N^{-2s_\mathcal{K}\alpha_h}),\ a.s.$$

Proof of Lemma F.6. Note that we can decompose the dynamics of $\widehat{\theta}_N$ as

$$\Delta\widehat{\theta}_N - V_N = (\mathbb{I}_p - \gamma_N\mathbb{H}_0)(\Delta\widehat{\theta}_{N-1} - V_{N-1}) + \gamma_N\delta_{1,N} + \gamma_N\delta_{2,N}$$

where $\delta_{1,N} = \Phi_N(\widehat{\theta}_{N-1}) - \Phi(\widehat{\theta}_{N-1})$, and $\delta_{1,N} = \int_0^1(\mathbb{H}(\theta_0,\tau) - \mathbb{H}(\widehat{\theta}_{N-1},\tau))\Delta\widehat{\theta}_{N-1}d\tau$. Note that $\gamma_N\|\delta_{1,N}\|_2 \le CN^{-\alpha_\gamma - s_\mathcal{K}\alpha_h}$. Under Condition 1, we have that $\mathbb{H}(\theta,\tau)$ is Lipschitz with respect to θ uniformly for all τ, that is, there exists a positive constant C such that for any θ_1 and θ_2, there holds $\sup_{\tau\in[0,1]}\|\mathbb{H}(\theta_1,\tau) - \mathbb{H}(\theta_2,\tau)\|_F \le C\|\theta_1 - \theta_2\|_2$. So $\|\delta_{2,N}\|_2 \le C\|\Delta\widehat{\theta}_{N-1}\|_2^2$. Define

$$\mathcal{A}_N = \Delta\widehat{\theta}_N - V_N, \tag{F.13}$$

then $\|\delta_{2,N}\|_2 \le C\|\mathcal{A}_{N-1} + V_{N-1}\|_2^2 \le 2C\|\mathcal{A}_{N-1}\|_2^2 + 2C\|V_{N-1}\|_2^2$. Since $\Delta\widehat{\theta}_N$ and V_N are both $o(1)$ a.s., $\mathcal{A}_N = o(1)$ a.s. Let $c_0 > 0$ be the eventual linear contraction constant of $I - \gamma_N\mathbb{H}_0$. On each probability-one path choose N_0 so large that $2C\|\mathcal{A}_{N-1}\|_2 \le c_0/2$ for $N \ge N_0$. Then the quadratic term satisfies $2C\gamma_N\|\mathcal{A}_{N-1}\|_2^2 \le (c_0/2)\gamma_N\|\mathcal{A}_{N-1}\|_2$ and is absorbed by the linear contraction. Thus, for N sufficiently large on that path,

$$\|\mathcal{A}_N\|_2 \le (1 - CN^{-\alpha_\gamma})\|\mathcal{A}_{N-1}\|_2 + CN^{-\alpha_\gamma - s_\mathcal{K}\alpha_h} + CN^{-\alpha_\gamma}b(N)$$

because we assume that $\|V_{N-1}\|_2^2 = O(b(N))$ a.s. holds. Then for almost all paths, there exists a (path-specific) N_0, such that

$$\begin{aligned}\|\mathcal{A}_N\|_2 \le &\left(\prod_{k=N_0}^{N}(1 - Ck^{-\alpha_\gamma})\right) C\sum_{k=N_0}^{N} k^{-\alpha_\gamma}(k^{-s_\mathcal{K}\alpha_h} + b(k))\left(\prod_{j=N_0}^{k}(1 - Cj^{-\alpha_\gamma})\right)^{-1} \\ &+ \left(\prod_{k=N_0+1}^{N}(1 - Ck^{-\alpha_\gamma})\right)\|\mathcal{A}_{N_0}\|_2.\end{aligned}$$

So $\|\mathcal{A}_N\|_2 = O(N^{-s_\mathcal{K}\alpha_h} \vee b(N))$ a.s. holds according to Lemma F.5. Finally, note that $\|\Delta\widehat{\theta}_N\|_2 \le \|\mathcal{A}_N\|_2 + \|V_N\|_2 = O(N^{-s_\mathcal{K}\alpha_h} \vee \sqrt{b(N)})$ a.s. holds. This proves the result. $\square$

Obviously, if we specify $b(N) = N^{-2\alpha_\gamma+\alpha_h+1}\log^r(N)$, then Lemma F.4 and Lemma F.6

immediately lead to that $\|\Delta\widehat{\theta}_N\|_2^2 = O(N^{-2\alpha_\gamma+\alpha_h+1}\log^r(N))$ a.s. for any $r > 1$. Again, we point out that this rate is slow. However, as we have demonstrated before, such preliminary rate can be used to build the sharp rate for $\Delta\widehat{\theta}_N$. To achieve sharp rate, we need a more delicate decomposition for V_N. In particular, decompose $V_N = V_{1,N} + V_{2,N}$, where

$$
\begin{aligned}
V_{1,N} =& \gamma_N \left(\Phi_N(\theta_0, \mathcal{W}_N) - \Phi_N(\theta_0)\right) + \gamma_{N-1}\left(\mathbb{I}_p - \gamma_N \mathbb{H}_0\right)\left(\Phi_{N-1}(\theta_0, \mathcal{W}_{N-1}) - \Phi_{N-1}(\theta_0)\right) + \\
& \cdots + \gamma_1 \left(\mathbb{I}_p - \gamma_N \mathbb{H}_0\right) \cdots \left(\mathbb{I}_p - \gamma_2 \mathbb{H}_0\right)\left(\Phi_1(\theta_0, \mathcal{W}_1) - \Phi_1(\theta_0)\right).
\end{aligned}
$$

and

$$
\begin{aligned}
V_{2,N} =& \gamma_N \int_0^1 \partial_\theta \left(\Phi_N(\theta_0 + \tau\Delta\widehat{\theta}_{N-1}, \mathcal{W}_N) - \Phi_N(\theta_0 + \tau\Delta\widehat{\theta}_{N-1})\right) d\tau \Delta\widehat{\theta}_{N-1} \\
& + \gamma_{N-1}\left(\mathbb{I}_p - \gamma_N \mathbb{H}_0\right) \int_0^1 \partial_\theta \left(\Phi_{N-1}(\theta_0 + \tau\Delta\widehat{\theta}_{N-2}, \mathcal{W}_{N-1}) - \Phi_{N-1}(\theta_0 + \tau\Delta\widehat{\theta}_{N-2})\right) d\tau \Delta\widehat{\theta}_{N-2} \\
& + \cdots + \gamma_1 \left(\mathbb{I}_p - \gamma_N \mathbb{H}_0\right) \cdots \left(\mathbb{I}_p - \gamma_2 \mathbb{H}_0\right) \int_0^1 \partial_\theta \left(\Phi_1(\theta_0 + \tau\Delta\widehat{\theta}_0, \mathcal{W}_1) - \Phi_1(\theta_0 + \tau\Delta\widehat{\theta}_0)\right) d\tau \Delta\widehat{\theta}_0.
\end{aligned}
$$

We present the asymptotic behaviors of $V_{1,N}$ and $V_{2,N}$.

Lemma F.7. *Let Condition 1–Condition 4 hold. With A_N, Σ_N, and $s_{N,j}^2$ defined above, for every fixed $j = 1, \ldots, p$,*

$$
\limsup_{N\to\infty} \frac{e_j^\top \mathcal{P}_{\mathbb{H}_0} A_N V_{1,N}}{\sqrt{2s_{N,j}^2 \log\log(s_{N,j}^2)}} = 1, \qquad \liminf_{N\to\infty} \frac{e_j^\top \mathcal{P}_{\mathbb{H}_0} A_N V_{1,N}}{\sqrt{2s_{N,j}^2 \log\log(s_{N,j}^2)}} = -1 \quad a.s.
$$

Proof of Lemma F.7. Choose k_0 as above. According to Lemma F.1,

$$
\mathbb{E}\left[\mathcal{P}_{\mathbb{H}_0}\left(\Phi_N(\theta_0, W_N) - \Phi_N(\theta_0)\right)\left(\Phi_N(\theta_0, W_N) - \Phi_N(\theta_0)\right)^\top \mathcal{P}_{\mathbb{H}_0}^\top\right] = \frac{2\mathcal{P}_{\mathbb{H}_0}\mathcal{V}_{\mathcal{K},0}\mathcal{P}_{\mathbb{H}_0}^\top}{B(B-1)h_N} + O(1).
$$

Recall that $\mathbb{H}_0 = \mathcal{P}_{\mathbb{H}_0}^\top \Lambda_{\mathbb{H}_0} \mathcal{P}_{\mathbb{H}_0}$. For coordinate j, let λ_j be the corresponding diagonal element, let $\widetilde{e}_{k,j}$ be coordinate j of $\mathcal{P}_{\mathbb{H}_0}\zeta_{0,k}$, and put

$$
a_{k,j} = \prod_{\ell=k_0}^{k} (1 - \lambda_j \gamma_\ell)^{-1}, \qquad e_{k,j} = \gamma_k a_{k,j} \widetilde{e}_{k,j}.
$$

Cancellation of the tail factors gives the exact identity

$$
e_j^\top \mathcal{P}_{\mathbb{H}_0} A_N V_{1,N} = C_{j,k_0-1} + \sum_{k=k_0}^{N} e_{k,j},
$$

where C_{j,k_0-1} is the fixed contribution of the finitely many innovations before k_0. We give the scalar calculation for $j = 1$ and write $e_k = e_{k,1}$ and $s_N^2 = \sum_{k=k_0}^N \mathbb{E}e_k^2$. Then

$$\mathbb{E}e_k^2 = \frac{\gamma_k^2}{\left[\prod_{j=k_0}^k (1-\lambda_{\mathbb{H}_0,1}\gamma_j)\right]^2} \left\{ \frac{2[\mathcal{P}_{\mathbb{H}_0}\mathcal{V}_{\mathcal{K},0}\mathcal{P}_{\mathbb{H}_0}^\top](1,1)}{B(B-1)h_k} + O(1) \right\}$$

where $[\mathcal{P}_{\mathbb{H}_0}\mathcal{V}_{\mathcal{K},0}\mathcal{P}_{\mathbb{H}_0}^\top](1,1)$ refers to the element in first column and first row of $\mathcal{P}_{\mathbb{H}_0}\mathcal{V}_{\mathcal{K},0}\mathcal{P}_{\mathbb{H}_0}^\top$. This leads to that for any N,

$$\left| s_N^2 - \sum_{k=k_0}^N \frac{2[\mathcal{P}_{\mathbb{H}_0}\mathcal{V}_{\mathcal{K},0}\mathcal{P}_{\mathbb{H}_0}^\top](1,1)\gamma_0^2 h_0^{-1} k^{-2\alpha_\gamma+\alpha_h}}{B(B-1)\left[\prod_{j=k_0}^k (1-\gamma_0\lambda_{\mathbb{H}_0,1}j^{-\alpha_\gamma})\right]^2} \right| \le C \sum_{k=k_0}^N \frac{k^{-2\alpha_\gamma}}{\left[\prod_{j=k_0}^k (1-\gamma_0\lambda_{\mathbb{H}_0,1}j^{-\alpha_\gamma})\right]^2}.$$

Using Lemma F.5, we can show that for N sufficiently large, we have that

$$s_N^{-2} \sum_{k=k_0}^N \frac{2[\mathcal{P}_{\mathbb{H}_0}\mathcal{V}_{\mathcal{K},0}\mathcal{P}_{\mathbb{H}_0}^\top](1,1)\gamma_0^2 h_0^{-1} k^{-2\alpha_\gamma+\alpha_h}}{B(B-1)\left[\prod_{j=k_0}^k (1-\gamma_0\lambda_{\mathbb{H}_0,1}j^{-\alpha_\gamma})\right]^2} \to 1$$

and

$$0 < C_1 \le \frac{s_N^2}{N^{-\alpha_\gamma+\alpha_h}\exp(C^* N^{1-\alpha_\gamma})} \le C_2 < \infty$$

for some positive constant C^*. So given the choice of α_γ and α_h, we have that $s_N^2 \to \infty$.

Moreover, we can similarly show that

$$\frac{\log^2(\log(s_N^2))\mathbb{E}e_N^4}{s_N^4} \le \frac{C\log^2(N)N^{-4\alpha_\gamma+3\alpha_h}\exp(2C^*N^{1-\alpha_\gamma})}{N^{-2\alpha_\gamma+2\alpha_h}\exp(2C^*N^{1-\alpha_\gamma})} = C\log^2(N)N^{-2\alpha_\gamma+\alpha_h}.$$

This implies that $\sum_{N\ge\max\{k_0,3\}} s_N^{-4}\log^2(\log(s_N^2))\mathbb{E}e_N^4 < \infty$ because $2\alpha_\gamma - \alpha_h > 1$, so

$$\left| s_N^{-1}\sqrt{\log(\log(s_N^2))}e_N \right| < C, \ a.s.$$

holds for any $C > 0$, implying that $s_N^{-1}\sqrt{\log(\log(s_N^2))}e_N \to 0$ a.s. holds. Using the LIL from Theorem 1.1 of Tomkins (1983), we have that

$$H_0 \le \overline{\lim}_{N\to\infty} \frac{\sum_{k=k_0}^N e_k}{\sqrt{2s_N^2\log(\log(s_N^2))}} \le H_1, \ a.s.,$$

and

$$-H_1^- \leq \underline{\lim}_{N\to\infty} \frac{\sum_{k=k_0}^N e_k}{\sqrt{2s_N^2 \log(\log(s_N^2))}} \leq -H_0^-, \ a.s.,$$

for some $0 \leq H_0 \leq H_1 \leq 1$ and $0 \leq H_0^- \leq H_1^- \leq 1$.

We next show that $H_0 = H_1 = 1$; the proof that $H_0^- = H_1^- = 1$ is identical. To show the results, following Tomkins (1983) we define $H_N(x) = s_N^{-2} \sum_{k=k_0}^N \mathbb{E}\left(e_k^2 \mathbf{1}(|e_k| \leq x s_k t_k^{-1})\right)$ with $t_k = \sqrt{2\log(\log(s_k^2))}$. For any x, we have that $H_N(x) = 1 - s_N^{-2} \sum_{k=k_0}^N \mathbb{E}\left(e_k^2 \mathbf{1}(|e_k| > x s_k t_k^{-1})\right)$, where

$$\mathbb{E}\left(e_k^2 \mathbf{1}(|e_k| > x s_k t_k^{-1})\right) \leq \sqrt{\mathbb{E} e_k^4} \sqrt{\mathbb{E}\mathbf{1}(|e_k| > x s_k t_k^{-1})}.$$

According to our previous analysis, we know that

$$\sqrt{\mathbb{E} e_k^4} \leq C k^{-2\alpha_\gamma + 3/2\alpha_h} \exp(C^* k^{1-\alpha_\gamma}),$$

$$\sqrt{\mathbb{E}\mathbf{1}(|e_k| > x s_k t_k^{-1})} \leq C \sqrt{\frac{t_k^4 \mathbb{E} e_k^4}{x^4 s_k^4}} \leq C x^{-2} \log(k) k^{-\alpha_\gamma + \alpha_h/2}$$

So $\sqrt{\mathbb{E} e_k^4} \sqrt{\mathbb{E}\mathbf{1}(|e_k| > x s_k t_k^{-1})} \leq C x^{-2} \log(k) k^{-3\alpha_\gamma + 2\alpha_h} \exp(C^* k^{1-\alpha_\gamma})$. Then

$$\begin{aligned}\sum_{k=k_0}^N \mathbb{E}\left(e_k^2 \mathbf{1}(|e_k| > x s_k t_k^{-1})\right) &\leq C x^{-2} \sum_{k=k_0}^N \log(k) k^{-3\alpha_\gamma + 2\alpha_h} \exp(C^* k^{1-\alpha_\gamma}) \\ &\leq C x^{-2} \log(N) N^{-2\alpha_\gamma + 2\alpha_h} \exp(C^* N^{1-\alpha_\gamma}).\end{aligned}$$

Since $s_N^2 \geq C N^{-\alpha_\gamma + \alpha_h} \exp(C^* N^{1-\alpha_\gamma})$, we have that $s_N^{-2} \sum_{k=k_0}^N \mathbb{E}\{e_k^2 \mathbf{1}(|e_k| > x s_k t_k^{-1})\} \to 0$, and consequently, $\lim_{N\to\infty} H_N(x) = 1$. According to Tomkins (1983), we have that

$$H_0 = \lim \inf_{N\to\infty} H_N(x), \quad H_1 = \lim \sup_{N\to\infty} H_N(x),$$

This leads to $H_0 = H_1 = 1$, and similarly $H_0^- = H_1^- = 1$. Based on the above analysis, we have that

$$\limsup_{N\to\infty} \frac{\sum_{k=k_0}^N e_k}{\sqrt{2s_N^2 \log\log(s_N^2)}} = 1, \qquad \liminf_{N\to\infty} \frac{\sum_{k=k_0}^N e_k}{\sqrt{2s_N^2 \log\log(s_N^2)}} = -1, \quad \text{a.s.}$$

The same argument applies to every coordinate and proves the coordinatewise LIL using only Condition 1–Condition 4. □

Next we introduce the lemma that clearly demonstrates how the convergence of $\|\Delta\widehat{\theta}_N\|^2$ accelerates the convergence rate of $\|V_{2,N}\|^2$, which, in turn, accelerates the convergence rate of $\|\Delta\widehat{\theta}_N\|^2$ itself.

Lemma F.8. *Let Condition 1–Condition 4 hold. If $\|\Delta\widehat{\theta}_N\|^2 = O(b(N))$ a.s. for some $b(N) \downarrow 0$ with $b(N-1)/b(N) \leq 1 + C/N$, then, for every $r > 1$,*

$$\|V_{2,N}\|_2^2 = o\left(b(N)N^{-2\alpha_\gamma+3\alpha_h+1}\log^r(N)\right), \ a.s.$$

Proof of Lemma F.8. Note that

$$V_{2,N} = \gamma_N \int_0^1 \partial_\theta \left(\Phi_N(\theta_0 + \tau\Delta\widehat{\theta}_{N-1}, W_N) - \Phi_N(\theta_0 + \tau\Delta\widehat{\theta}_{N-1})\right) d\tau\Delta\widehat{\theta}_{N-1} + (\mathbb{I}_p - \gamma_N\mathbb{H}_0)\, V_{2,N-1}.$$

Then according to Lemma F.1(4),

$$\mathbb{E}_{N-1}\|V_{2,N}\|_2^2 \leq (1 - C\gamma_N)\,\|V_{2,N-1}\|_2^2 + C\gamma_N^2 h_N^{-3}\|\Delta\widehat{\theta}_{N-1}\|_2^2,$$

and

$$\begin{aligned}
&\mathbb{E}_{N-1}\left[b^{-1}(N)N^{2\alpha_\gamma-3\alpha_h-1}\log^{-r}(N)\|V_{2,N}\|_2^2\right]\\
&\leq \left(1 - CN^{-\alpha_\gamma}\right)\frac{b(N-1)\log^r(N-1)}{b(N)\log^r(N)}\left[b^{-1}(N-1)(N-1)^{2\alpha_\gamma-3\alpha_h-1}\log^{-r}(N-1)\|V_{2,N-1}\|_2^2\right]\\
&+ Cb(N-1)b^{-1}(N)(N\log^r(N))^{-1}b^{-1}(N-1)\|\Delta\widehat{\theta}_{N-1}\|_2^2\\
&\leq \left(1 - CN^{-\alpha_\gamma}\right)\left[b^{-1}(N-1)(N-1)^{2\alpha_\gamma-3\alpha_h-1}\log^{-r}(N-1)\|V_{2,N-1}\|_2^2\right]\\
&+ Cb(N-1)b^{-1}(N)(N\log^r(N))^{-1}b^{-1}(N-1)\|\Delta\widehat{\theta}_{N-1}\|_2^2.
\end{aligned}$$

Since $\|\Delta\widehat{\theta}_N\|_2^2 = O(b(N))$ a.s. and $b(N-1)/b(N) \leq C$, we have that

$$\sum_{N=1}^{\infty} b(N-1)b^{-1}(N)(N\log^r(N))^{-1}b^{-1}(N-1)\|\Delta\widehat{\theta}_{N-1}\|_2^2 < \infty, \ a.s.$$

Using the method that we use in the proof of Lemma F.4, we have that

$$b^{-1}(N)N^{2\alpha_\gamma-3\alpha_h-1}\log^{-r}(N)\|V_{2,N}\|_2^2 = o(1), \ a.s.$$

So $\|V_{2,N}\|_2^2 = o\left(b(N)N^{-2\alpha_\gamma+3\alpha_h+1}\log^r(N)\right)$ a.s. holds. □

F.3.2. Proof of Theorem F.1

Define $d = 2\alpha_\gamma - 3\alpha_h - 1 > 0, \delta_0 = 2\alpha_\gamma - \alpha_h - 1, \delta_\star = \alpha_\gamma - \alpha_h, \delta_{m+1} = \min\{\delta_\star, \delta_m + d\}$. The preliminary bounds in Lemma F.4 and Lemma F.6 give the exponent δ_0. If at some stage $\|\Delta\widehat{\theta}_N\|_2^2 = O_{\text{a.s.}}(N^{-\delta_m} \operatorname{polylog} N)$, Lemma F.8 improves the nonlinear martingale by d, while Lemma F.6 transfers that improvement back to the iterate. Because $d > 0$, after finitely many repetitions the recursion reaches $\delta_\star$. If the final exponent step hits $\delta_\star$ exactly, the bound can still carry an arbitrary fixed polylogarithmic factor. Apply Lemma F.8 once more to that bound. Its factor N^{-d} makes the nonlinear martingale polynomially smaller than $N^{-\delta_\star}$, so Lemma F.7, Lemma F.6, and the deterministic bias bound yield

$$\|\Delta\widehat{\theta}_N\|_2^2 = O_{\text{a.s.}}(b_N), \qquad b_N = N^{-\delta_\star} \log N = N^{-\alpha_\gamma + \alpha_h} \log N.$$

A final application of Lemma F.8 yields $\|V_{2,N}\|_2^2 = o_{\text{a.s.}}(b_N N^{-d} (\log N)^r)$ for every $r > 1$; in particular, $\|V_{2,N}\|_2 = o_{\text{a.s.}}(N^{-\delta_\star/2})$ because $d > 0$. The deterministic linearization remainder from Lemma F.6 satisfies

$$\|\mathcal{A}_N\|_2 = O_{\text{a.s.}}(b_N \vee N^{-s_K \alpha_h}) = o_{\text{a.s.}}(N^{-\delta_\star/2}),$$

where the last equality follows from Condition 4. Hence

$$\Delta\widehat{\theta}_N = V_{1,N} + o_{\text{a.s.}}(N^{-\delta_\star/2}).$$

By the coordinatewise LIL in Lemma F.7 and the fact that p is fixed,

$$\|V_{1,N}\|_2 = O_{\text{a.s.}}\left(N^{-\delta_\star/2}\sqrt{\log N}\right).$$

Together with the preceding display, this proves the displayed raw-iterate rate.

For the linear recursion, commutativity of the factors gives

$$\mathcal{P}_{\mathbb{H}_0} A_N V_{1,N} = C_{k_0-1} + \sum_{k=k_0}^{N} \gamma_k \mathcal{P}_{\mathbb{H}_0} A_k \zeta_{0,k},$$

where C_{k_0-1} is a fixed finite prefix. Lemma F.7 proves from Condition 1–Condition 4 that each $s_{N,j}$ diverges and that the independent triangular array satisfies the scalar LIL. Thus the prefix is negligible coordinatewise and (F.12) follows. The same scalar variance estimates give

$$\frac{\|e_j^\top \mathcal{P}_{\mathbb{H}_0} A_N\|_2}{s_{N,j}} = O(N^{\delta_\star/2}),$$

so the $o_{\text{a.s.}}(N^{-\delta_\star/2})$ nonlinear and deterministic remainders are negligible under every coordinate LIL normalization.

F.4. Proof of Theorem B.1

The raw kernel-based updates lead to $\widehat{\theta}_k = \widehat{\theta}_{k-1} + \gamma_k \Phi_k(\widehat{\theta}_{k-1}, \mathcal{W}_k)$, then we can decompose the update as follows

$$\widehat{\theta}_k = \widehat{\theta}_{k-1} + \gamma_k \Phi(\widehat{\theta}_{k-1}) + \gamma_k(\Phi_k(\widehat{\theta}_{k-1}) - \Phi(\widehat{\theta}_{k-1})) + \gamma_k(\Phi_k(\widehat{\theta}_{k-1}, \mathcal{W}_k) - \Phi_k(\widehat{\theta}_{k-1})).$$

Simple calculation leads to

$$\begin{aligned}
\|\Delta\widehat{\theta}_k\|_2^2 = {} & \|\Delta\widehat{\theta}_{k-1} + \gamma_k\Phi(\widehat{\theta}_{k-1})\|_2^2 + \gamma_k^2\|\Phi_k(\widehat{\theta}_{k-1}) - \Phi(\widehat{\theta}_{k-1})\|_2^2 + \gamma_k^2\|\Phi_k(\widehat{\theta}_{k-1}, \mathcal{W}_k) - \Phi_k(\widehat{\theta}_{k-1})\|_2^2 \\
& + 2\gamma_k(\Delta\widehat{\theta}_{k-1} + \gamma_k\Phi(\widehat{\theta}_{k-1}))^\top(\Phi_k(\widehat{\theta}_{k-1}) - \Phi(\widehat{\theta}_{k-1})) \\
& + 2\gamma_k(\Delta\widehat{\theta}_{k-1} + \gamma_k\Phi(\widehat{\theta}_{k-1}))^\top(\Phi_k(\widehat{\theta}_{k-1}, \mathcal{W}_k) - \Phi_k(\widehat{\theta}_{k-1})) \\
& + 2\gamma_k^2(\Phi_k(\widehat{\theta}_{k-1}) - \Phi(\widehat{\theta}_{k-1}))^\top(\Phi_k(\widehat{\theta}_{k-1}, \mathcal{W}_k) - \Phi_k(\widehat{\theta}_{k-1})).
\end{aligned} \tag{F.14}$$

Recall that we use $\mathbb{E}_{k-1}$ to denote the expectation conditioned on the first $k-1$ batches (and let $\mathbb{E}_0$ denote the unconditional expectation). We verify the conditional expectation for all the terms on the right-hand side of (F.14). The population Hessian is uniformly bounded above under Condition 1; hence choose a deterministic k_0 such that $\gamma_k \sup_{\theta,\tau} \lambda_{\max}\{\mathbb{H}(\theta,\tau)\} \leq 1$ for $k \geq k_0$. All contraction displays below are asserted for $k \geq k_0$; the finite prefix changes only constants. Since $\Phi(\theta) = -\int_0^1 \mathbb{H}(\theta,\tau)d\tau\Delta\theta$ for any θ, we have that $\Phi(\widehat{\theta}_{k-1}) = -\int_0^1 \mathbb{H}(\widehat{\theta}_{k-1},\tau)d\tau\Delta\widehat{\theta}_{k-1}$. So Lemma F.2 leads to

$$\begin{aligned}
\mathbb{E}_{k-1}\|\Delta\widehat{\theta}_{k-1} + \gamma_k\Phi(\widehat{\theta}_{k-1})\|_2^2 &= \Delta\widehat{\theta}_{k-1}^\top\left(\mathbb{I}_p - \gamma_k\int_0^1 \mathbb{H}(\widehat{\theta}_{k-1},\tau)d\tau\right)^2\Delta\widehat{\theta}_{k-1} \\
&\leq \left(1 - C\gamma_k\left(\frac{\overline{z} - \underline{z}}{12(1 + \|\Delta\widehat{\theta}_{k-1}\|_2 + \|\theta_0\|_2)} \wedge r_X\right)^{2p+2}\right)\|\Delta\widehat{\theta}_{k-1}\|_2^2.
\end{aligned}$$

Lemma F.1(2) leads to $\mathbb{E}_{k-1}\gamma_k^2\|\Phi_k(\widehat{\theta}_{k-1})-\Phi(\widehat{\theta}_{k-1})\|_2^2 \le C\gamma_k^2 h_k^{2s\kappa}$, and Lemma F.1(3) leads to $\gamma_k^2\mathbb{E}_{k-1}\|\Phi_k(\widehat{\theta}_{k-1},W_k)-\Phi_k(\widehat{\theta}_{k-1})\|_2^2 \le C\gamma_k^2 h_k^{-1}$. Moreover, we have that

$$\begin{aligned}
&\left|\mathbb{E}_{k-1}\left[2\gamma_k(\Delta\widehat{\theta}_{k-1}+\gamma_k\Phi(\widehat{\theta}_{k-1}))^\top(\Phi_k(\widehat{\theta}_{k-1})-\Phi(\widehat{\theta}_{k-1}))\right]\right| \\
&\le Ch_k^{2s\kappa}\left(1-C\gamma_k\left(\frac{\overline{z}-\underline{z}}{12(1+\|\Delta\widehat{\theta}_{k-1}\|_2+\|\theta_0\|_2)}\wedge r_X\right)^{2p+2}\right)\|\Delta\widehat{\theta}_{k-1}\|_2^2+C\gamma_k^2,
\end{aligned}$$

$$\mathbb{E}_{k-1}\left[2\gamma_k(\Delta\widehat{\theta}_{k-1}+\gamma_k\Phi(\widehat{\theta}_{k-1}))^\top(\Phi_k(\widehat{\theta}_{k-1},W_k)-\Phi_k(\widehat{\theta}_{k-1}))\right]=0,$$

and

$$\mathbb{E}_{k-1}\left[2\gamma_k^2(\Phi_k(\widehat{\theta}_{k-1})-\Phi(\widehat{\theta}_{k-1}))^\top(\Phi_k(\widehat{\theta}_{k-1},W_k)-\Phi_k(\widehat{\theta}_{k-1}))\right]=0.$$

The above together leads to

$$\begin{aligned}
\mathbb{E}_{k-1}\|\Delta\widehat{\theta}_k\|_2^2 \le \left(1+Ch_k^{2s\kappa}\right)\left\|\Delta\widehat{\theta}_{k-1}\right\|_2^2+C\gamma_k^2h_k^{-1} \\
-C\gamma_k\left(\frac{\overline{z}-\underline{z}}{12(1+\|\Delta\widehat{\theta}_{k-1}\|_2+\|\theta_0\|_2)}\wedge r_X\right)^{2p+2}\|\Delta\widehat{\theta}_{k-1}\|_2^2.
\end{aligned}$$

Then since $\sum_{k=1}^\infty h_k^{2s\kappa}<\infty$ and $\sum_{k=1}^\infty \gamma_k^2h_k^{-1}<\infty$, according to Theorem 1 of Robbins and Siegmund (1971), we have that $\|\Delta\widehat{\theta}_{k-1}\|_2^2$ converges a.s. to a finite random variable, and that

$$\sum_{k=1}^\infty \gamma_k\left(\frac{\overline{z}-\underline{z}}{12(1+\|\Delta\widehat{\theta}_{k-1}\|+\|\theta_0\|)}\wedge r_X\right)^{2p+2}\|\Delta\widehat{\theta}_{k-1}\|^2<\infty, \qquad \text{a.s.}$$

Since $\sum_{k=1}^\infty\gamma_k=\infty$, if $\|\Delta\widehat{\theta}_{k-1}\|_2^2$ converges but not to zero, then

$$\sum_{k=1}^\infty \gamma_k\left(\frac{\overline{z}-\underline{z}}{12(1+\|\Delta\widehat{\theta}_{k-1}\|_2+\|\theta_0\|_2)}\wedge r_X\right)^{2p+2}\|\Delta\widehat{\theta}_{k-1}\|_2^2=+\infty,$$

which leads to a contradiction. This implies that $\|\Delta\widehat{\theta}_k\|_2^2\to 0$ must a.s. hold. This proves the convergence of $\widehat{\theta}_k$ and $\overline{\theta}_k$.

We further decompose the PR average. Note that

$$\begin{aligned}
\Delta\widehat{\theta}_N &= \Delta\widehat{\theta}_{N-1}+\gamma_N\Phi_N(\widehat{\theta}_{N-1},\mathcal{W}_N) \\
&= \Delta\widehat{\theta}_{N-1}-\gamma_N\mathbb{H}_0\Delta\widehat{\theta}_{N-1}+\gamma_N\left[\delta_{1,N}+\delta_{2,N}+\Phi_N(\widehat{\theta}_{N-1},\mathcal{W}_N)-\Phi_N(\widehat{\theta}_{N-1})\right],
\end{aligned}$$

so

$$\mathbb{H}_0\Delta\widehat{\theta}_{N-1} = \gamma_N^{-1}(\Delta\widehat{\theta}_{N-1} - \Delta\widehat{\theta}_N) + \left[\delta_{1,N} + \delta_{2,N} + \Phi_N(\widehat{\theta}_{N-1}, \mathcal{W}_N) - \Phi_N(\widehat{\theta}_{N-1})\right],$$

where recall that $\delta_{1,N} = \Phi_N(\widehat{\theta}_{N-1}) - \Phi(\widehat{\theta}_{N-1})$ and $\delta_{2,N} = \int_0^1(\mathbb{H}(\theta_0,\tau) - \mathbb{H}(\widehat{\theta}_{N-1},\tau))d\tau\Delta\widehat{\theta}_{N-1}$. Then

$$\sum_{k=1}^N \mathbb{H}_0\Delta\widehat{\theta}_{k-1} = \sum_{k=1}^N \gamma_k^{-1}(\Delta\widehat{\theta}_{k-1} - \Delta\widehat{\theta}_k) + \sum_{k=1}^N \left[\delta_{1,N} + \delta_{2,N} + \Phi_k(\widehat{\theta}_{k-1}, \mathcal{W}_k) - \Phi_k(\widehat{\theta}_{k-1})\right]$$

Because $\overline{\theta}_N = N^{-1}\sum_{k=1}^N \widehat{\theta}_k$,

$$N\mathbb{H}_0\Delta\overline{\theta}_N = \sum_{k=1}^N \mathbb{H}_0\Delta\widehat{\theta}_{k-1} + \mathbb{H}_0(\Delta\widehat{\theta}_N - \Delta\widehat{\theta}_0).$$

The last boundary term is $O(1)$ a.s. after consistency and hence is $o(N^{(1+\alpha_h)/2})$ a.s. For the first term, we have that

$$\begin{aligned}\sum_{k=1}^N \gamma_k^{-1}(\Delta\widehat{\theta}_{k-1} - \Delta\widehat{\theta}_k) &= \sum_{k=1}^N \gamma_k^{-1}\Delta\widehat{\theta}_{k-1} - \sum_{k=1}^N \gamma_k^{-1}\Delta\widehat{\theta}_k = \sum_{k=0}^{N-1} \gamma_{k+1}^{-1}\Delta\widehat{\theta}_k - \sum_{k=1}^N \gamma_k^{-1}\Delta\widehat{\theta}_k \\ &= \gamma_1^{-1}\Delta\widehat{\theta}_0 + \sum_{k=1}^{N-1}(\gamma_{k+1}^{-1} - \gamma_k^{-1})\Delta\widehat{\theta}_k - \gamma_N^{-1}\Delta\widehat{\theta}_N\end{aligned}$$

Under Condition 4, we have that $\gamma_{k+1}^{-1} - \gamma_k^{-1} = \gamma_0^{-1}((k+1)^{\alpha_\gamma} - k^{\alpha_\gamma}) \le Ck^{\alpha_\gamma - 1}$. So

$$\left\|\sum_{k=1}^N \gamma_k^{-1}(\Delta\widehat{\theta}_{k-1} - \Delta\widehat{\theta}_k)\right\|_2 \le C + C\sum_{k=1}^{N-1} k^{\alpha_\gamma-1}\left\|\Delta\widehat{\theta}_k\right\|_2 + \gamma_0^{-1}N^{\alpha_\gamma}\left\|\Delta\widehat{\theta}_N\right\|_2$$

According to Theorem F.1, we have that

$$\sum_{k=1}^{N-1} k^{\alpha_\gamma-1}\left\|\Delta\widehat{\theta}_k\right\|_2 \le C\sum_{k=1}^{N-1} k^{\frac{\alpha_\gamma+\alpha_h}{2}-1}\sqrt{\log k},\ a.s.$$

and that

$$\gamma_0^{-1}N^{\alpha_\gamma}\left\|\Delta\widehat{\theta}_N\right\| \le CN^{\frac{\alpha_\gamma+\alpha_h}{2}}\sqrt{\log N}, \qquad \text{a.s.}$$

This implies that

$$\left\|\sum_{k=1}^{N}\gamma_k^{-1}(\Delta\widehat{\theta}_{k-1}-\Delta\widehat{\theta}_k)\right\|_2 = O\left(N^{\frac{\alpha_\gamma+\alpha_h}{2}}\sqrt{\log N}\right), \qquad \text{a.s.},$$

and as a result,

$$\left(N^{1+\alpha_h}\right)^{-\frac{1}{2}}\left\|\sum_{k=1}^{N}\gamma_k^{-1}(\Delta\widehat{\theta}_{k-1}-\Delta\widehat{\theta}_k)\right\|_2 \to 0, \qquad \text{a.s.}$$

On the other side, we have that $\|\delta_{1,N}\|_2 \le CN^{-s_{\mathcal{K}}\alpha_h}$ and $\|\delta_{2,N}\| \le C\|\Delta\widehat{\theta}_{N-1}\|_2^2$. So $\|\sum_{k=1}^{N}\delta_{2,N}\|_2 = O(N^{-\alpha_\gamma+\alpha_h+1}\log N)$ a.s. and, by the exact power-sum bound,

$$\left\|\sum_{k=1}^{N}\delta_{1,N}\right\|_2 \le C\sum_{k=1}^{N}k^{-s_{\mathcal{K}}\alpha_h} = O\big(1+N^{(1-s_{\mathcal{K}}\alpha_h)_+}\log N\big)\,.$$

The latter is $o(N^{(1+\alpha_h)/2})$ because $s_{\mathcal{K}}\alpha_h > 1/2$ under Condition 4; the former has that order because $\alpha_\gamma > (1+\alpha_h)/2$. Thus both terms are $o(\sqrt{N^{1+\alpha_h}})$ a.s.

We finally look at $\sum_{k=1}^{N}(\Phi_k(\widehat{\theta}_{k-1},\mathcal{W}_k)-\Phi_k(\widehat{\theta}_{k-1}))$. Obviously, according to our previous decomposition, we have that

$$\begin{aligned}\sum_{k=1}^{N}\left(\Phi_k(\widehat{\theta}_{k-1},\mathcal{W}_k)-\Phi_k(\widehat{\theta}_{k-1})\right) &= \sum_{k=1}^{N}\left(\Phi_k(\theta_0,\mathcal{W}_k)-\Phi_k(\theta_0)\right)\\ &\quad+\sum_{k=1}^{N}\int_0^1\nabla_\theta\left(\Phi_k(\theta_0+\tau\Delta\widehat{\theta}_{k-1},\mathcal{W}_k)-\Phi_k(\theta_0+\tau\Delta\widehat{\theta}_{k-1})\right)d\tau\Delta\widehat{\theta}_{k-1}.\end{aligned}$$

We next show that

$$\sum_{k=1}^{N}\int_0^1\nabla_\theta\left(\Phi_k(\theta_0+\tau\Delta\widehat{\theta}_{k-1},\mathcal{W}_k)-\Phi_k(\theta_0+\tau\Delta\widehat{\theta}_{k-1})\right)d\tau\Delta\widehat{\theta}_{k-1} = o(\sqrt{N^{1+\alpha_h}}), \qquad \text{a.s.}$$

Let

$$M_N = \sum_{k=1}^{N}\int_0^1\nabla_\theta\Big\{\Phi_k(\theta_0+\tau\Delta\widehat{\theta}_{k-1},\mathcal{W}_k)-\Phi_k(\theta_0+\tau\Delta\widehat{\theta}_{k-1})\Big\}\,d\tau\Delta\widehat{\theta}_{k-1}.$$

We have that

$$\mathbb{E}_{N-1}\|M_N\|_2^2 \le \|M_{N-1}\|_2^2 + Ch_N^{-3}\|\Delta\widehat{\theta}_{N-1}\|_2^2 \le \|M_{N-1}\|_2^2 + CN^{-\alpha_\gamma+4\alpha_h}\log N, \qquad \text{a.s.}$$

where the last equality uses the raw-iterate rate from Theorem F.1. Then using the previous proofs leads to $\|M_N\|_2 = O(\frac{1+4\alpha_h-\alpha_\gamma}{2}\sqrt{\log^{1+r}(N)})$ a.s. for any $r > 1$. This is $o_{\text{a.s.}}(N^{(1+\alpha_h)/2})$ because $\alpha_\gamma > 3\alpha_h$, which follows from $2\alpha_\gamma - 3\alpha_h > 1$ and $\alpha_\gamma < 1$.

The above analysis implies that

$$\left\| N \left(\frac{2\mathcal{V}_{\mathcal{K},0}}{B(B-1)} \right)^{-\frac{1}{2}} \mathbb{H}_0 \Delta\overline{\theta}_N - \left(\frac{2\mathcal{V}_{\mathcal{K},0}}{B(B-1)} \right)^{-\frac{1}{2}} \sum_{k=1}^{N} \left(\Phi_k(\theta_0, \mathcal{W}_k) - \Phi_k(\theta_0) \right) \right\|_2 = o\left(\sqrt{N^{1+\alpha_h}}\right), \ a.s.$$

Note that

$$\text{var}\left[\left(\frac{2\mathcal{V}_{\mathcal{K},0}}{B(B-1)} \right)^{-\frac{1}{2}} \left(\Phi_k(\theta_0, \mathcal{W}_k) - \Phi_k(\theta_0) \right) \right] = h_k^{-1}\mathbb{I}_p + O(1) = h_0^{-1}k^{\alpha_h}\mathbb{I}_p + O(1),$$

and

$$\sum_{k=1}^{N} \text{var}\left[\left(\frac{2\mathcal{V}_{\mathcal{K},0}}{B(B-1)} \right)^{-\frac{1}{2}} \left(\Phi_k(\theta_0, \mathcal{W}_k) - \Phi_k(\theta_0) \right) \right] = \frac{N^{1+\alpha_h}\mathbb{I}_p}{h_0(1+\alpha_h)} + o(N^{1+\alpha_h}).$$

For a fixed coordinate j, write

$$\zeta_{k,j} = e_j^\top \left(\frac{2\mathcal{V}_{\mathcal{K},0}}{B(B-1)} \right)^{-1/2} \{\Phi_k(\theta_0, \mathcal{W}_k) - \Phi_k(\theta_0)\}, \qquad s_{N,j}^2 = \sum_{k \le N} \mathbb{E}\zeta_{k,j}^2.$$

The preceding variance calculation gives $s_{N,j}^2 \asymp N^{1+\alpha_h}$, while Lemma F.1(5) gives $\mathbb{E}|\zeta_{k,j}|^4 \le Ch_k^{-3} \le Ck^{3\alpha_h}$. Therefore

$$\sum_{N \ge 3} \frac{\{\log\log(s_{N,j}^2)\}^2 \mathbb{E}|\zeta_{N,j}|^4}{s_{N,j}^4} \le C \sum_{N \ge 3} N^{-2+\alpha_h} (\log\log N)^2 < \infty.$$

Here $\alpha_h < 1/3$ follows from $2\alpha_\gamma - 3\alpha_h > 1$ and $\alpha_\gamma < 1$. The scalar Lindeberg LIL of Tomkins (1983) applies and yields the two coordinatewise limits in Theorem B.1.

To prove the CLT result, part (5) of Lemma F.1 gives

$$\sum_{k=1}^{N} \mathbb{E} \left\| \Phi_k(\theta_0, W_k) - \Phi_k(\theta_0) \right\|_2^4 \le C \sum_{k=1}^{N} h_k^{-3} \le CN^{1+3\alpha_h}.$$

so

$$\frac{\sum_{k=1}^{N} \mathbb{E} \left\| \Phi_k(\theta_0, W_k) - \Phi_k(\theta_0) \right\|_2^4}{N^{2+2\alpha_h}} \to 0.$$

The multivariate Lyapunov central limit theorem therefore gives the stated result.