Consistency for Random forests for spatially dependent data

Arkajyoti Saha[1], Sumanta Basu[2], and Abhirup Datta[1]

[1]Department of Biostatistics, Johns Hopkins Bloomberg School of Public Health

[2]Department of Statistics and Data Science, Cornell University

Appendix

A1 Consistency

In this Section we present the main theoretical result on consistency of RF-GLS for a very general class of dependent error processes.

The outline of the proof highlighting the new theoretical challenges addressed are presented in Section A2 along with some general results of independence importance. The formal proofs are provided in the Supplementary materials.

Notations

$|S|$ denotes the cardinality of any set S. $\mathbb{I}(\cdot)$ is the indicator function. The null set is $\{\}$. For any matrix $\mathbf{M}$, $\mathbf{M}^+$ denotes its generalized Moore-Penrose inverse, and $\|\mathbf{M}\|_p$, for $1 \leq p \leq \infty$, denotes its matrix $\mathbb{L}_p$ norm. For any $n \times n$ symmetric matrix $\mathbf{M}$, $\lambda_{\min} = \lambda_1 \leq \lambda_2 \leq \ldots \leq \lambda_n = \lambda_{\max}$ denotes its eigenvalues. A sequence of numbers $\{a_n\}_{n\geq 1}$ is $O(b_n)$ (or $o(b_n)$) when the sequence $|a_n/b_n|$ is bounded above (or goes to 0) as $n \to \infty$. A sequence of random variables is called $O_b(1)$ if it is uniformly bounded almost surely, $O_p(1)$ if it is bounded in probability, and $o_p(1)$ if it goes to 0 in probability. $X \sim Y$ implies X follows the same distribution as Y. $\mathbb{R}$, $\mathbb{Z}$, and $\mathbb{N}$ denote the

set of real numbers, integers, and natural numbers respectively. For $M \in \mathbb{R}^+$, T_M is the truncation operator, i.e. $T_M(u) = \max(-M, \min(u, M))$.

A1.1 Assumptions

We make the following assumption on the true dependence structure of the data:

Assumption 1 (Mixing condition)**.** $Y_i = m(X_i) + \epsilon_i$ where the error process $\{\epsilon_i\}$ is a stationary, *absolutely regular* (β-mixing) process Bradley (2005) with finite $(2+\delta)^{th}$ moment for some $\delta > 0$.

We focus on absolutely regular or β-mixing processes, since this class of stochastic processes is rich enough to accommodate many commonly used dependent error processes like ARMA Mokkadem (1988), GARCH Carrasco and Chen (2002), certain Markov processes Doukhan (2012) and Gaussian processes with Matérn covariance family. At the same time, uniform law of large numbers (ULLN) from independent processes can be extended to this dependent process under moderate restriction on the class of functions under consideration. No additional assumption is required on the decay rate of the β-mixing coefficients (which are often hard to check).

Assumption 2 (Regularity of the working precision matrix)**.** The working precision matrix $\mathbf{Q} = \mathbf{\Sigma}^{-1}$ admits a regular and sparse lower-triangular Cholesky factor $\mathbf{\Sigma}^{-\frac{1}{2}}$ such that

$$\mathbf{\Sigma}^{-\frac{1}{2}} = \begin{pmatrix} \mathbf{L}_{q\times q} & 0 & 0 & \cdots & \cdots \\ \boldsymbol{\rho}^\top_{1\times(q+1)} & & 0 & \cdots & \cdots \\ 0 & \boldsymbol{\rho}^\top_{1\times(q+1)} & & 0 & \cdots \\ \vdots & & \ddots & & \vdots \\ \cdots & 0 & 0 & \boldsymbol{\rho}^\top_{1\times(q+1)} & \end{pmatrix},$$

where $\boldsymbol{\rho} = (\rho_q, \rho_{q-1}, \cdots, \rho_0)^\top \in \mathbb{R}^{q+1}$ for some fixed $q \in \mathbb{N}$, and $\mathbf{L}$ is a fixed lower-triangular $q \times q$ matrix .

Under Assumption 2, for any two vectors $\mathbf{x}$ and $\mathbf{y}$, defining $x_i = y_i = 0$ for $i \leq 0$, we have

$$\mathbf{x}^\top \mathbf{Q} \mathbf{y} = \alpha \sum_i x_i y_i + \sum_{j \neq j' = 0}^{q} \rho_j \rho_{j'} \sum_i x_{i-j} y_{i-j'} + \sum_{i \in \tilde{\mathcal{A}}_1} \sum_{i' \in \tilde{\mathcal{A}}_2} \tilde{\gamma}_{i,i'} x_i y_{i'}, \tag{A1}$$

where $\alpha = \|\boldsymbol{\rho}\|_2^2$, $\tilde{\mathcal{A}}_1, \tilde{\mathcal{A}}_2 \subset \{1, 2, \cdots, n\}$ with $|\tilde{\mathcal{A}}_1|, |\tilde{\mathcal{A}}_2| \leq 2q$, $\tilde{\gamma}_{i,i'}$'s are fixed (independent of n) functions of $\mathbf{L}$ and $\boldsymbol{\rho}$. The expression of the quadratic form in (A1) makes it evident that $\lambda_{\max}(\mathbf{Q})$ is bounded as $n \to \infty$. As the third term is a sum of fixed (at most $4q^2$) number of terms, it is $O(1)$ as long as $\mathbf{x}$ and $\mathbf{y}$ are bounded. Such sparse and regular Cholesky factors routinely appear in time series analysis for AR(p) process. For spatial data, common families of covariance functions do not generally satisfy this assumption. However, NNGP covariance matrices satisfy this and are now commonly used as an excellent approximation to the full GP covariance matrices Datta et al. (2016). We discuss these examples in Section A1.3.

Assumption 3 (Diagonal dominance of the working precision matrix)**.** $\mathbf{Q}$ is diagonally dominant satisfying $\mathbf{Q}_{ii} - \sum_{j \neq i} |\mathbf{Q}_{ij}| > \xi$ for all i, for some constant $\xi > 0$.

Diagonal dominance implies $\lambda_{\min}(\mathbf{Q})$ is bounded away from zero as $n \to \infty$ which is needed to ensure stability of the GLS estimate. We will discuss in Section A1.3 how working correlation matrices from popular time series and spatial processes with regular design satisfy this Assumption. Note that under Assumption 2, checking that the first $(q+1)$ rows of $\mathbf{Q}$ are diagonally dominant is enough to verify Assumption 3.

Assumption 4 (Tail behavior of the error distribution)**.**

(a) $\exists \{\zeta_n\}_{n \geq 1}$ such that

$$\zeta_n \to \infty, \ \frac{t_n (\log n) \zeta_n^8}{n} \to 0, \text{ and}$$
$$\mathbb{E}\left[\left(\max_i \epsilon_i^2\right) \mathbb{I}\left(\max_i \epsilon_i^2 > \zeta_n^2\right)\right] \to 0 \text{ as } n \to \infty.$$

(b) $\exists$ constant $C_\pi > 0$ and $n_0 \in \mathbb{N}^*$ such that with probability $1 - \pi$, $\forall n > n_0$,

$$\max_i |\epsilon_i| \leqslant C_\pi \sqrt{\log n}.$$

(c) Let $\mathcal{I}_n \subseteq \{1, 2, \cdots, n\}$ with $|\mathcal{I}_n| = a_n$ and $a_n \to \infty$ as $n \to \infty$. Then $\frac{1}{a_n} |\sum_{i \in \mathcal{I}_n} \epsilon_i| > \delta$ with probability at most $C \exp(-c a_n)$ and $\frac{1}{n} |\sum_i \epsilon_i^2| > \sigma_0^2$ with probability at most $C \exp(-cn)$ for any $\delta > 0$, and some constants $c, C, \sigma_0^2 > 0$.

The scaling in Assumption 4(a) is same as the one used in Scornet et al. (2015) where ζ_n was $O(\log n)^2$ for Gaussian errors. In general, the choice of ζ_n will be dependent on the error distribution. Assumption 4(a), (b) and (c) will all be satisfied by sub-Gaussian errors.

Assumption 5 (Additive model)**.** The true mean function $m(\mathbf{x})$ is additive on the coordinates x_d of $\mathbf{x}$, i.e., $m(\mathbf{x}) = \sum_{d=1}^{D} m_d(x_d)$, where each component m_d is continuous.

As demonstrated in Scornet et al. (2015), additive models provide a rich enough environment to address the asymptotic properties of nonparametric methods like RF sans the additional complexities in controlling asymptotic variation of m in leaf nodes. Since RF is invariant to monotone transformations of covariates Friedman et al. (2001); Friedman (2006), without loss of generality, the covariates can be distribution function transformed to be Unif$[0, 1]$ distributed. Hence we can assume that the components (functions) m_d are supported on $[0, 1]$, implying m is uniformly bounded by some constant M_0.

A1.2 Main result

For the t^{th} tree, the predicted value from our method at a new point $\mathbf{x}_0$ in covariate space is denoted by $m_n(\mathbf{x}_0; \Theta_t, \mathbf{\Sigma}, \mathcal{D}_n)$ where $\mathcal{D}_n = \{(\mathbf{x}_i, y_i) \,|\, i = 1, \ldots, n\}$ denote the data, Θ_t indicates the randomness associated with each tree. In practice, Θ_t will include both the re-sampling of datapoints used in each tree as well as the choice of random splitting variable for iterative splitting in the tree. For tractability, Scornet et al. (2015) considered sub-sampling instead of re-sampling for the theoretical study. In our theoretical study, for analytical tractability of the GLS weights, we consider trees that use the entire set of samples and the randomness Θ_t in each tree is only used to choose the candidate set of features for each split. The randomness for each tree are i.i.d., i.e., $\Theta_t \overset{i.i.d}{\sim} \Theta$; $\Theta \perp\!\!\!\perp \mathcal{D}_n, \forall t \in \{1, \cdots, n_{tree}\}$. The finite RF-GLS estimate $\widehat{m}_{n,n_{tree}}(\mathbf{x}_0; \Theta_1, \cdots, \Theta_{n_{tree}}, \mathbf{\Sigma}, \mathcal{D}_n)$ that will be used in practice is given by the sample average of the individual tree estimates. Conceptually, n_{tree} can be arbitrarily large, hence following Scornet et al. (2015), we focus on "infinite" RF-GLS estimate given by $\bar{m}_n(\mathbf{x}_0; \mathbf{\Sigma}, \mathcal{D}_n) = \mathbb{E}_\Theta m_n(\mathbf{x}_0; \Theta, \mathbf{\Sigma}, \mathcal{D}_n)$ where the expectation w.r.t Θ is conditional on $\mathcal{D}_n$. For notational convenience, we hide the dependence of $m_n, \widehat{m}_{n,n_{tree}}, \bar{m}_n$ on $\mathbf{\Sigma}$ and $\mathcal{D}_n$ throughout the rest of this article. Our main result on $\mathbb{L}_2$-consistency is stated next, the proof is deferred to Section A2.1.

Theorem A1.1. Under Assumptions 1-5, RF-GLS is $\mathbb{L}_2$-consistent, i.e.,

$$\lim_{n\to\infty} \mathbb{E} \int (\bar{m}_n(X) - m(X))^2 \, dX = 0,$$

if for some $\delta > 0$, $\lim_{n\to\infty} \mathbb{E}\frac{1}{n}\sum_i |m_n(X_i)|^{2+\delta} < \infty$.

The $(2+\delta)^{th}$ moment assumption in Theorem A1.1 is needed to generalize the consistency results from the i.i.d. case to the dependent case. The following corollaries discuss three specific cases where this assumption is met. Their proofs are deferred to Section S1.3.

Corollary A1.1. For dependent errors, under Assumptions 1-5, RF-GLS is $\mathbb{L}_2$ consistent if either of the two conditions hold:

(a) Case 1: The errors are bounded.

(b) Case 2: The working precision matrix $\mathbf{Q}$ satisfies $\min_i \mathbf{Q}_{ii} > \sqrt{2}\max_i \sum_{j\neq i} |\mathbf{Q}_{ij}|$.

For bounded errors (part (a)), the $(2+\delta)^{th}$ moment-bound of Theorem A1.1 is immediately satisfied, and hence consistency can be established without further assumptions. For unbounded errors, a stronger form of diagonal dominance condition needed in Corollary A1.1 Part (b). This is used to control the $(2+\delta)^{th}$ moment of the data weights arising from the gram-matrix $(\mathbf{Z}^\top\mathbf{Q}\mathbf{Z})^{-1}$ which in turn ensures the moment-bound. We discuss examples and specific parameter choices ensuring this in Section A1.3. Also note that, the assumption of diagonal dominance is not on the true correlation matrix of the error process and hence is not a restriction on the data-generation mechanism, but rather on the working correlation matrix which is chosen by the user. One can always use parameters in the working correlation matrix that satisfies this (although enforcing this is not needed in practice).

RF is RF-GLS with $\mathbf{Q} = \mathbf{I}$, hence the Assumption of Corollary A1.1 part (b) is trivially satisfied. This proves consistency of RF under β-mixing dependence.

Corollary A1.2. RF Breiman (2001) is $\mathbb{L}_2$-consistent for dependent errors if Assumptions 1, 4 and 5 are satisfied.

To the best of our knowledge, Corollary A1.2 is the first result on consistency of RF under a dependent error process. Since RF is simply RF-GLS with the working correlation matrix $\mathbf{\Sigma} = \mathbf{I}$,

Assumptions 2 and 3 are automatically satisfied, and hence we only need the Assumptions of β-mixing process, tail bounds and additive model. The consistency result is analogous to the ordinary least squares estimate being consistent even for correlated errors. Besides its own importance, Corollary A1.2 also heuristically justifies the first step used in practical implementation of RF-GLS. The parameters in the working correlation matrix is unknown, hence we use the RF to get a preliminary estimate of m, estimate the spatial parameters using the residuals, and use these estimated parameters in the working correlation matrix for RF-GLS. This is again, analogous to feasible GLS which estimates the working correlation matrix using residuals based on OLS. Corollary A1.2 guarantees that the initial estimator used to obtain the residuals is consistent.

A1.3 Examples

In this Section, we give examples of two popular dependent error processes under which a consistent estimate of m can be obtained using RF-GLS.

A1.3.1 Spatial Matérn Gaussian processes

Our main example focuses on the spatial non-linear mixed model using Gaussian Processes. While many candidate exist for the covariance function of GP, the class of Matérn covariances enjoy hegemonic popularity in the spatial literature owing to its remarkable property of characterizing the smoothness of the spatial surface $\epsilon(\ell)$ Stein (2012). The stationary (isotropic) Matérn covariance function is specified by

$$C(\ell_i, \ell_j \,|\, \boldsymbol{\phi}) = C(\|\ell_i - \ell_j\|_2) = \sigma^2 \frac{2^{1-\nu}\left(\sqrt{2}\phi\|\ell_i - \ell_j\|_2\right)^{\nu}}{\Gamma(\nu)} \mathcal{K}_\nu\left(\sqrt{2}\phi\|\ell_i - \ell_j\|_2\right), \tag{A2}$$

where $\boldsymbol{\phi} = (\sigma^2, \phi, \nu)^\top$ and $\mathcal{K}_\nu$ is the modified Bessel function of second kind.

We consider a Matérn process sampled on one-dimensional regular lattice. This regular design is considered both for tractability of the Matérn GP likelihood but also for ensuring stationarity of the process in the sense required in Theorem A1.1 as for irregular spaced data $Cov(\epsilon_1, \epsilon_2) \neq Cov(\epsilon_2, \epsilon_3)$ whenever $\|\ell_1 - \ell_2\|_2 \neq \|\ell_2 - \ell_3\|_2$. Such assumptions on the dimensionality and/or regularity of design has been widely used for theoretical studies of spatial processes Du et al. (2009); Stein et al. (2002). By keeping the gap in the lattice fixed, we are also essentially using increasing-domain

asymptotics, as parameters are generally not identifiable in fixed domain asymptotics for Matérn GPs Zhang (2004).

The error process arising from the marginalization of (1) is the sum of a Matérn process and a nugget (random error) process. We consider half-integer $\nu \in 1.5, 2.5, \ldots$. This class of processes are popularly studied and used owing to their convenient state-space representation Hartikainen and Särkkä (2010) which in turn leads to efficient computation of these Matérn GP likelihoods. The state-space representation of half-integer Matérn GP is equivalent to that of a stable $AR(q_0)$ process on the continuous one-dimensional domain with $q_0 = \nu + 1/2$. However, unlike an $AR(q_0)$ time series, the Matérn GP when sampled on the discrete integer lattice is no longer an $AR(q_0)$ process. Consequently, covariance matrices $\mathbf{\Sigma}$ generated from Matérn GP (expect for exponential GP), do not satisfy the sparsity and regularity of the working correlation matrix of Assumption 2.

Instead, we consider the working correlation $\mathbf{\Sigma}$ to come from a Nearest Neighbor Gaussian Process (NNGP) Datta et al. (2016) based on the Matérn covariance. NNGP covariance matrices are one of the most successful surrogates for full GP covariances for large spatial data, reducing likelihood computations from $O(n^3)$ to $O(n)$. What is important for the theoretical study is that an NNGP is constructed by sequentially specifying the conditional distributions as $\epsilon_i \,|\, \epsilon_{1:i-1} \sim \epsilon_i \,|\, \epsilon_{N_q(i)}$ where $N_q(i) \subset \{1, \ldots, i-1\}$ is the set of q-nearest neighbors of ℓ_i among $\ell_1, \ldots, \ell_{i-1}$. When the locations are the integer grid, $N_q(i)$ becomes $\{i-1, \ldots, i-q\}$, and the NNGP construction is akin to an $AR(q)$ process. Consequently, the Cholesky factor $\mathbf{\Sigma}^{-1/2}$ from NNGP on an integer lattice satisfies Assumption 2 with $\boldsymbol{\rho} = (1, -\mathbf{c}^\top \mathbf{C}^{-1})^\top / \sqrt{1 - \mathbf{c}^\top \mathbf{C}^{-1} \mathbf{c}}$ and $\mathbf{L}$ such that $\mathbf{L}^\top \mathbf{L} = \mathbf{C}^{-1}$ where $\mathbf{C} = Cov(\epsilon_{1:q})$, $\mathbf{c} = Cov(\epsilon_{1:q}, \epsilon_{q+1})$ Finley et al. (2019). This ensures the following consistency result of RF-GLS fitted with NNGP for data generated using Matérn GP. The proof is in Section S1.6.

Proposition A1.1. Consider a spatial process $y(\ell_i) = m(X_i) + \epsilon(\ell_i)$ from (1) where m is an additive model as specified in Assumption 5, $\epsilon(\ell_i) = w(\ell_i) + \epsilon^*(\ell_i)$ where $\epsilon^*(\ell)$ denote i.i.d. $N(0, \tau_0^2)$ noise, and $w(\ell)$ be a Matérn GP, sampled on the integer lattice, with parameters $\boldsymbol{\phi}_0 = (\sigma_0^2, \phi_0, \nu_0)^\top$, ν_0 being a half-integer. Let $\mathbf{\Sigma}$ denote a covariance matrix from a Nearest Neighbor Gaussian Process (NNGP) derived from a Matérn covariance with parameters $\boldsymbol{\phi} = (\sigma^2, \phi, \nu)^\top$ and τ^2. Then there exists some $K > 0$ such if $\phi > K$, then RF-GLS using $\mathbf{\Sigma}$ yields an $\mathbb{L}_2$ consistent estimate of m.

One observation is central to the proof. The half-integer Matérn GP, which is an $AR(q_0)$ process in the continuous domain, when sampled on a discrete lattice becomes an ARMA process (Ihara (1993) Theorem 2.7.1). This will establish absolutely regular mixing of these Matérn processes using the result of Mokkadem (1988) on ARMA processes, and subsequently consistency of RF-GLS by Theorem A1.1.

A1.3.2 Autoregressive time series

Our main focus in this manuscript is estimation of nonlinear regression function in the spatial mixed model (1). However, the scope of our RF-GLS algorithm is much broader. It can be used for functional estimation in the general nonlinear regression model $Y_i = m(X_i) + \epsilon_i$ where ϵ_i is a dependent stochastic process with valid second moment. The method only relies on knowledge of an estimate of the residual covariance matrix $\mathbf{\Sigma} = Cov(\boldsymbol{\epsilon})$. The general consistency result (Theorem A1.1) is also not specific to the spatial GP setting, and only relies on general assumptions on the nature of the dependence, tail bounds of the error, and structure of the working correlation matrix. In this Section, we demonstrate that RF-GLS can be used for consistent function estimation for time series data, i.e., where ϵ_i models the serial (temporal) correlation. In particular, we discuss consistency of RF-GLS for Autoregressive (AR) error processes, one of the mainstays of time series studies. An $AR(q)$ model can be written as:

$$\epsilon_i = a_1\epsilon_{i-1} + a_2\epsilon_{i-2} + \ldots + a_q\epsilon_{i-q} + \eta_i \tag{A3}$$

where η_i is a realization of a white noise process at time i. AR processes are β-mixing Mokkadem (1988), they also produce a banded Cholesky factor of the precision matrix as required in Assumption 2. Hence, we have the following assertion. Its proof is in Section S1.6.

Proposition A1.2. Consider a time series $Y_i = m(X_i) + \epsilon_i$ where i is the time, m satisfies Assumption 5, ϵ_i denote a sub-Gaussian stable $AR(q_0)$ process. Let $\mathbf{\Sigma}$ denote a working correlation matrix from a stationary $AR(q)$ process. Then RF-GLS using $\mathbf{\Sigma}$ produces an $\mathbb{L}_2$ consistent estimate of m if

1. $q = 1$ and the working autocorrelation parameter ρ used in $\mathbf{\Sigma}$ satisfies $|\rho| < 1$ (for bounded errors) or $|\rho| < 1/(2\sqrt{2})$ (for unbounded errors).

2. $q > 1$ and the AR(q) working precision matrix $\mathbf{Q} = \mathbf{\Sigma}^{-1}$ satisfies Assumption 3 (for bounded error) or $\min_i \mathbf{Q}_{ii} > \sqrt{2} \max_i \sum_{j \neq i} |\mathbf{Q}_{ij}|$ (for unbounded error).

We separate the results for $q = 1$ and $q \geq 2$ since, unlike $AR(1)$, for general $AR(q)$ it is challenging to derive closed form expressions of the constraints on the parameter space needed to satisfy Assumption 3 or the stronger diagonal dominance condition in Proposition A1.2 part 2 (for unbounded errors). However, verifying these conditions for a given AR precision matrix $\mathbf{Q}$ is straightforward due to stationarity and banded structure. One only needs to check the first $q+1$ rows of $\mathbf{Q}$ irrespective of the sample size n. The necessary condition for row $q+1$ is

$$\|\boldsymbol{\rho}\|^2 > 2\kappa \sum_{j=1}^{q} |\sum_{j'=j}^{q} \rho_{j'} \rho_{j'-j}| \tag{A4}$$

where κ equals 1 for bounded errors and equals $\sqrt{2}$ for unbounded errors. Additional checks are only needed for the first q rows of $\mathbf{Q}$.

In practical implementation of AR processes, the order of the autoregression is often chosen based on analysis of the auto-correlation function of the residuals and may not equal the true order of the autoregression. Proposition A1.2 accommodates this scenario by not restricting the working autoregressive covariance to be of the same order or have the same coefficients as the ones generating the data.

A2 Proof of Consistency

For studying RF-GLS with dependent data, we adopt the framework of consistency analysis of nonparametric regression (introduced in Nobel et al. (1996), generalized in Györfi et al. (2006)). In Scornet et al. (2015), the authors also adopted this framework to prove consistency of RF for i.i.d. errors. After presenting an informal outline of the consistency argument in Györfi et al. (2006), we provide a road map of how we extend different pieces of this argument for RF-GLS with dependent data, and highlight additional technical challenges that we resolved in this work. All formal proofs are provided in the Supplementary materials.

Györfi et al. (2006) consider a least squares estimator of the form

$$m_n(,\Theta) \in arg \min_{f \in \mathcal{F}_n} \frac{1}{n} \sum_i [f(X_i) - Y_i]^2$$

where $\mathcal{F}_n$ is a carefully chosen, data-dependent function class which is large enough to control *approximation error* (i.e., how well the true function is estimated by the function class), and small enough to control *estimation error* (i.e., how far the estimate m_n is from the representative of $\mathcal{F}_n$ closest to m) in the sense that a Uniform Law of Large Number (ULLN) holds on this class. m_n is consistent if both errors vanish asymptotically.

In order to use powerful exponential inequalities which hold for classes of bounded functions, the proof of $\mathbb{L}_2$-consistency in Györfi et al. (2006) uses a standard truncation argument in probability theory with a diverging sequence of truncation thresholds $\{\zeta_n\}$. They first show that if *uniformly over the class of truncated functions* in $\mathcal{F}_n$,

1. *Approximation Error* of m by a data-driven class $\mathcal{F}_n$ containing m_n is small, i.e.,

$$\mathbb{E}\left[\inf_{f \in T_{\zeta_n}\mathcal{F}_n} \mathbb{E}_X \left[f(X) - m(X)\right]^2\right] \to 0;$$

2. *Estimation Error* is small so that a ULLN holds over $\mathcal{F}_n$ for squared error loss, i.e.
$\mathbb{E}\left[\sup_{f \in T_{\zeta_n}\mathcal{F}_n} \left|\frac{1}{n}\sum_i (f(X_i) - Y_i))^2 - \mathbb{E}[f(X) - Y]^2\right|\right] \to 0;$

then the truncated estimator $T_{\zeta_n} m_n$ is $\mathbb{L}_2$-consistent for f. Then they extended the consistency guarantee from truncated to the original estimators by showing that the *truncation error* vanishes under suitable tail decay assumptions on the error distribution.

In the analysis of RF-GLS, there are two main challenges. The error process ϵ_i is no longer an i.i.d process but a stochastic process capturing the dependence. Also, we work with a quadratic loss $(\mathbf{Y} - f(\mathbf{X}))^\top \mathbf{Q} (\mathbf{Y} - f(\mathbf{X}))$ instead of $\|\mathbf{Y} - f(\mathbf{X})\|^2$ where, $f(\mathbf{X}) - \mathbf{Y} = (f(X_1) - Y_1, f(X_2) - Y_2, \cdots, f(X_n) - Y_n)^\top$. Addressing these require non-trivial generalizations of each of the above pieces.

A2.1 Consistency of quadratic loss optimizers in data-driven function classes under dependent errors

Our main technical statement (Theorem A2.1) is a generalization of (Györfi et al., 2006, Theorem 10.2) to the setting of dependent error processes and quadratic loss functions. Let $\mathcal{D}_n = \{(X_1, Y_1), \cdots, (X_n, Y_n)\}$ be the data where $Y_i = m(X_i) + \epsilon_i$. With randomness parameter Θ, let $\mathcal{F}_n = \mathcal{F}_n(\mathcal{D}_n, \Theta)$ be a class of functions. We will consider an optimal estimator $m_n \in \mathcal{F}_n$ with respect to quadratic loss:

$$m_n \in \arg\min_{f \in \mathcal{F}_n} \frac{1}{n}(f(\mathbf{X}) - \mathbf{Y})^\top \mathbf{Q}(f(\mathbf{X}) - \mathbf{Y}). \tag{A5}$$

This simply states that m_n is the GLS estimate with respect to the working precision matrix $\mathbf{Q}$ in the class $\mathcal{F}_n$. This is analogous to the OLS assumption used in Györfi et al. (2006). When $\mathcal{F}_n$ is the class of piecewise constant functions on the partitions generated by a regression tree, m_n is our GLS-style regression-tree estimate. Hence studying estimators of the form (A5) more generally suffices to prove consistency of GLS-style regression tree and henceforth of RF-GLS.

We now state a general technical result establishing $\mathbb{L}_2$-consistency of such estimators m_n under β-mixing (absolutely regular) error processes. This is a sufficiently large class of processes that includes spatial Matérn GP and autoregressive time series as discussed in Sections A1.3.1 and A1.3.2. The result is applicable beyond RF to more general nonparametric GLS estimators from dependent data using other suitable of function classes.

Theorem A2.1. Let $\{\epsilon_i\}$ be a stationary β-mixing process satisfying Assumption 1, and the matrix $\mathbf{Q}$ satisfies Assumptions 2 and 3. Let $m_n(., \Theta) : \mathbb{R}^D \to \mathbb{R}$ denote a quadratic-loss optimizer (with respect to $\mathbf{Q}$) of the form (A5) in a data-dependent function class $\mathcal{F}_n$. If m_n and $\mathcal{F}_n$ satisfies the following conditions:

(C.1) (Truncation error) $\exists$ $\{\zeta_n\}$ such that $\lim_{n\to\infty} \zeta_n = \infty$ and $\zeta_n^2/n \to 0$, such that we have,

$$\lim_{n\to\infty} \mathbb{E} \max_i \left[m_n(X_i) - T_{\zeta_n} m_n(X_i)\right]^2 = 0$$

(C.2) (Approximation error) $\lim_{n\to\infty} \mathbb{E}_\Theta \left[\inf_{f \in T_{\zeta_n}\mathcal{F}_n} \mathbb{E}_X |f(X) - m(X)|^2\right] = 0$

(C.3) (Estimation error) Let $\dot{X}_i$, $\dot{\epsilon}_i$ and $\dot{Y}_i = m(\dot{X}_i) + \dot{\epsilon}_i$ be such that $\dot{\mathcal{D}}_n = \{(\dot{X}_i, \dot{Y}_i) | i = 1, \ldots, n\}$

be identically distributed as $\mathcal{D}_n$ but independent of $\mathcal{D}_n$. Define $f(\dot{\mathbf{X}})$ and $\dot{\mathbf{Y}}$ similar to $f(\mathbf{X})$ and $\mathbf{Y}$. Then, we have for all arbitrary $L > 0$

$$\lim_{n\to\infty} \mathbb{E}\left[\sup_{f\in T_{\zeta_n}\mathcal{F}_n} |\frac{1}{n}(f(\mathbf{X})-\mathbf{Y})^\top \mathbf{Q}(f(\mathbf{X})-\mathbf{Y}) - \mathbb{E}\frac{1}{n}(f(\dot{\mathbf{X}})-\dot{\mathbf{Y}})^\top \mathbf{Q}(f(\dot{\mathbf{X}})-\dot{\mathbf{Y}})|\right] = 0.$$

Then we have

$$\lim_{n\to\infty} \mathbb{E}\left[\mathbb{E}_X(m_n(X,\Theta)-m(X))^2)\right] = 0, \text{ and}$$

$$\lim_{n\to\infty} \mathbb{E}_X(\bar{m}_n(X)-m(X))^2 = 0;$$

where $\bar{m}_n(X) = \mathbb{E}_\Theta m_n(X,\Theta)$ and X is a new sample independent of the data.

The proof is deferred to Section S1.3. Theorem A2.1 is a result of independent importance as it is a general statement on $\mathbb{L}_2$ consistency of a wide class of GLS estimates under β-mixing dependent errors. Besides data-driven-partitioning-based estimates like RF or RF-GLS, it can be used to study properties of histograms, kernel-density estimates, local polynomials, etc., under β-mixing error processes. The second part of the theorem states that the consistency also holds for an ensemble estimator $\bar{m}_n$ that averages many such estimates m_n each specified with random parameters Θ. This will be used to show consistency of the RF-GLS forest estimate subsequent to showing consistency of each RF-GLS tree estimate.

A2.2 Approximation Error

The condition **C.2** of asymptotically vanishing approximation error ensures that as sample size increases, the growing class of approximating functions (e.g. piece-wise constant functions in the case of regression trees) is rich enough to approximate the target function. In earlier works on consistency of RF, approximation error was controlled under a stringent assumption of vanishing diameter of leaf nodes. Scornet et al. (2015) replaced this by a condition that the variance of Y (or equivalently, variation of m) within a leaf node of a regression-tree vanishes asymptotically, and verified this condition for RF. There are two steps to show this.

(i) Establish a *theoretical or population-level split-criterion* — an asymptotic limit of the empirical split-criterion used in practice, such that variation of m in the leaf-nodes of a hypothetical regression tree generated using the theoretical criterion is small.

(ii) Establish stochastic equicontinuity of the empirical split-criterion, such that if two set of qualifying splits $\mathbf{Z}^{(1)}$ and $\mathbf{Z}^{(2)}$ are close, their corresponding empirical split-criterion values are close, irrespective of the location of the splits.

For our RF-GLS trees, the partitioning is driven by the DART criterion (5) and is different from the CART (2) criterion used in the RF trees. Since the GLS loss and estimator involves the matrix $\mathbf{Q}$, they are not available in simple scalar expressions unlike the OLS loss (sum of squares) and estimator (mean response within a node). So we address a number of technical challenges for steps (i) and (ii) that do not appear in the analysis of RF.

For (i), Lemma 2.1 and Theorem 2.1, establishes that the DART split-criterion remarkably has the same limit of as that for the CART criterion. Hence, variation of m in trees generated by this theoretical criterion is controlled in the same way as for RF.

For (ii), we require an entirely new and involved proof of equicontinuity for the DART-split criterion of RF-GLS as the previous arguments of Scornet et al. (2015) do not immediately generalize for RF-GLS loss function (5). We discuss the new contributions in Section A2.2.1.

A2.2.1 Equicontinuity of the split criterion

Equicontinuity of the CART-split criterion $\frac{1}{n_P}[\sum_{i=1}^{n_P}(Y_i^P - \bar{Y}^P)^2 - \sum_{i_r=1}^{n_R}(Y_i^R - \bar{Y}^R)^2 - \sum_{i_l=1}^{n_L}(Y_i^L - \bar{Y}^L)^2]$ was the center-piece of the theory in Scornet et al. (2015), requiring involved but elegant arguments on the geometry of splits. Since the CART criterion only concerns the parent node to be split and its potential two child nodes, the equicontinuity essentially boiled down to showing closeness of the respective means and variances of these 3 nodes for the two sets of splits. These 3 scalar mean and variance differences are functions of the difference in volumes of the respective nodes which goes to zero uniformly as the splits come closer.

For RF-GLS, to update each node, the entire set of node representatives get updated via the GLS-estimate (6) which is analytically intractable due to the matrix inversion. Also, the DART-split criterion (5)

$$\frac{1}{n}\left[\left(\mathbf{Y}-\mathbf{Z}^{(0)}\hat{\boldsymbol{\beta}}_{GLS}(\mathbf{Z}^{(0)})\right)^{\top}\mathbf{Q}\left(\mathbf{Y}-\mathbf{Z}^{(0)}\hat{\boldsymbol{\beta}}_{GLS}(\mathbf{Z}^{(0)})\right)-\frac{1}{n}\left(\mathbf{Y}-\mathbf{Z}\hat{\boldsymbol{\beta}}_{GLS}(\mathbf{Z})\right)^{\top}\mathbf{Q}\left(\mathbf{Y}-\mathbf{Z}\hat{\boldsymbol{\beta}}_{GLS}(\mathbf{Z})\right)\right]$$

is a quadratic form of the plugged-in GLS-estimate and thus a function of the representatives of all nodes and not just the 3 nodes as in RF. Our equicontinuity proof is built on viewing GLS

predictions as oblique projections on the design matrices corresponding to the splits.

We consider two scenarios: **R1** – where for at least one set of split, both potential child nodes have substantial volumes, and **R2** – where for each set of split, one potential child node have ignorable volume. Under **R1**, all nodes for at least one set of split have substantial representation ensuring that the gram matrix has norm bounded away from zero and equicontinuity can be established using perturbation bounds on orthogonal projection operators Chen et al. (2016). Under **R2**, this will not be the case as for both set of splits one child node will have small volume. Instead, we show that difference of the DART-split criterion at the previous level of (parent) splits is small using the same perturbation argument as the parent node volumes are bounded away from zero. Subsequently we argue that the creation of the children nodes do not change the split criterion substantially as one of the child nodes is essentially empty. This new matrix-based proof for equicontinuity also circumvents the need to invoke mathematical induction as required in Scornet et al. (2015).

Proposition A2.1 (Equicontinuity of empirical DART-split criterion)**.** Under Assumptions 2, 3, and 4(b) and 4(c), the DART split criterion (5) is stochastically equicontinuous with respect to the set of splits.

The proof is involved, and is deferred to Section S1.1 where the Proposition is more technically phrased using additional notation on splits. Subsequent to proving equicontinuity, we can prove the following result of approximation error

Proposition A2.2. Let $\mathcal{F}_n = \mathcal{F}_n(\Theta)$ is the set of all functions $f : [0,1]^D \to \mathbb{R}$, piecewise constant on each cell of the partition obtained by an RF-GLS tree. Then, under Assumptions 1 - 5, the class $T_{\zeta_n}\mathcal{F}_n$ satisfies the approximation error condition **(C.2)**.

A2.3 Estimation Error

Theorem A2.1 shows that for GLS estimators like (A5), one needs to control the quadratic form estimation error (ULLN **C.3**). A common technique for proving ULLNs under dependence is
(i) to prove an analogous ULLN for i.i.d. error processes, and then
(ii) use mixing conditions to generalize the result for the dependent process of interest.

Both steps require addressing new challenges for RF-GLS which we discuss in the next two subsections.

A2.3.1 Cross-product function classes

For step (i), it is difficult to directly state an i.i.d. analogue of the ULLN **C.3** as it is not possible to have an error process $\{\epsilon_i^*\}$ which is simultaneously i.i.d., satisfies $\epsilon_i^* \sim \epsilon_i$ and $E(\mathbf{y}^\top \mathbf{Q}\mathbf{y}) = E(\mathbf{y}^{*\top}\mathbf{Q}\mathbf{y}^*)$. To see this, simply note that as $Cov(\epsilon_i^*, \epsilon_{i-j}^*) = 0 \neq Cov(\epsilon_i, \epsilon_{i-j})$, with $y_i^* = m(X_i)+\epsilon_i^*$ we will have $E(q_{i,i-j}y_i y_{i-j}) \neq E(q_{i,i-j}y_i^* y_{i-j}^*)$ where $\mathbf{Q} = (q_{ii'})$. Instead, we create separate i.i.d. analogues of ULLN for each term in the expansion of the quadratic form, i.e., both the squared error and the cross-product terms. These ULLN are stated as condition **C.3.iid** in the Supplementary materials. Condition **C.3.iid**(a) is the standard squared error ULLN using the i.i.d error processes, and has been proved in Györfi et al. (2006) Theorem 10.2 generally, and in Scornet et al. (2015) in particular for RF.

For the cross-product terms, the ULLN is stated in Condition **C.3.iid**(b) of the Supplement, and is to our knowledge a novel strategy. We essentially construct separate ULLN for each pair of lags $(i, i-j)$ for $j = 1, \ldots, q$. As mentioned earlier, use of the univariate copy $\{\epsilon_i^*\}$ will not allow to generalize to the corresponding term in **C.3** to hold as $\epsilon_i^* \perp \epsilon_{i-j}^*$ but ϵ_i and ϵ_{i-j} are correlated. Instead, we create bivariate i.i.d. copies $(\tilde{\epsilon}_i, \ddot{\epsilon}_{i-j})$ that are identically distributed as the joint (bivariate) distribution of the pairs $(\epsilon_i, \epsilon_{i-j})$, but are independent over i. Although the entire vector $(\epsilon_1, \ldots, \epsilon_n)^\top$ is correlated, due to the banded nature of the working precision matrix $\mathbf{Q}$, it is enough to have this cross-product ULLN only for lags $1 \leq j \leq q$, which combined with the ULLN for the squared terms gives us a ULLN for the entire quadratic form.

Formulation and establishing this cross-product ULLN is a new contribution and is of independent importance for establishing vanishing limits of any estimation error that involves interaction terms. We prove this ULLN in Proposition S1.1 of the Supplement by showing that cross-product function classes has the same concentration rate as that for squared-error function classes with respect to the random $\mathbb{L}_p$ norm entropy number.

A2.3.2 ULLN for β-mixing processes

For step (ii), we need to go from Conditions **(C.3.iid)**(a) and **(C.3.iid)**(b) for i.i.d processes to their analogs for dependent error processes, which would then immediately establish **C.3**. It has been shown that the mixing condition determines the assumptions required on the class $\mathcal{F}_n$ Dehling and Philipp (2002). If we look at the "hierarchy" of dependence structures, strong-mixing or α-mixing Bradley (2005), is one of the broadest family of dependent processes accommodating dependent structures "furthest" from independence. However, existing ULLN results for α-mixing processes require the class of functions in $\mathcal{F}_n$ to be Lipschitz continuous Dehling and Philipp (2002). As regression-tree estimates are inherently discontinuous due to nature of discrete partitioning, this will not be satisfied here.

Hence we focus on absolutely regular or β-mixing process. This class of mixing processes is rich enough to include a number of commonly used spatial or time-series structures as discussed in the examples of Section A1.3. Our main challenge here is that no existing ULLN for β-mixing processes apply to the class of functions on partitions of the RF-GLS trees. ULLN for Glivenko-Cantelli classes under β-mixing was established in Nobel and Dembo (1993). Similar results have been established for a class of $\tilde{\phi}$-mixing processes in Peligrad (2001). Both results, do not need any convergence rate on the mixing coefficients, but require the class $\mathcal{F}_n$ to have an envelope (dominator) F (free of n). For data-driven partitioning based estimates like RF or RF-GLS trees such uniform envelopes are not available (as the envelop is the truncation threshold $\zeta_n \to \infty$). Instead we propose a ULLN for dependent processes that uses a weaker assumption of a n-varying envelop with a moment-bound. The proof is deferred to Section S1.2.

Proposition A2.3. (A general ULLN for β-mixing processes) Let $\{U_i\}$ be an $\mathbb{R}^d$-valued stationary β-mixing process. Let $\mathcal{G}_n(\{U_i\}_{i=0}^{n-1})$ be a class of functions $\mathbb{R}^d \to \mathbb{R}$ with envelope $G_n \geq \sup_{g\in\mathcal{G}_n} |g|$, such that G_n is "uniformly mean integrable", i.e.

$$\lim_{C\to\infty} \lim_{n\to\infty} \mathbb{E}\frac{1}{n}\sum_i |G_n(U_i)|\mathbb{I}(|G_n(U_i)| > C) = 0. \tag{A6}$$

Let $\{U_i^*\}$ be such that U_i^* is identically distributed as U_i, $\forall\, i$ and $U_i^* \perp\!\!\!\perp U_j^*$; $\forall i \neq j$. Then, $\sup_{g\in\mathcal{G}_n} \left|\frac{1}{n}\sum_i (g(U_i^*) - \mathbb{E}g(U_i^*))\right| \xrightarrow{\mathbb{L}_1} 0 \implies \sup_{g\in\mathcal{G}_n} \left|\frac{1}{n}\sum_i (g(U_i) - \mathbb{E}g(U_i))\right| \xrightarrow{\mathbb{L}_1} 0$.

Proposition A2.3 ensures that ULLN for i.i.d. errors is enough to generalize to β-mixing error processes as long as the function classes are contained within a sequence of *" mean uniform integrable"* envelopes in the sense of (A6). Next, we show that the ULLN holds for RF-GLS trees under a $(2+\delta)^{th}$ moment assumption that is sufficient for mean uniform integrability.

Proposition A2.4 (Estimation error for RF-GLS)**.** Let $\mathcal{F}_n = \mathcal{F}_n(\Theta)$ is the set of all functions $f : [0,1]^D \to \mathbb{R}$, piecewise constant on each cell of the partition obtained by an RF-GLS tree. If any subset $\tilde{F}_n \subseteq \mathcal{F}_n$ satisfies the following condition:

(C.4) (Moment bound:) $\exists$ an envelope $F_n \geqslant \sup_{f \in \tilde{\mathcal{F}}_n} |f|$, such that $\lim_{n\to\infty} \mathbb{E}\frac{1}{n}\sum_i |F_n(X_i)|^{2+\delta} < \infty$ for some $\delta > 0$,

then $T_{\zeta_n}\tilde{\mathcal{F}}_n$ satisfies the ULLN **(C3)** for β-mixing error processes, and under Assumption 2.

The proof is in Section S1.2. The $(2+\delta)^{th}$ moment condition **C.4** is easier to verify for RF or RF-GLS as discussed in Corollaries A1.1 and A1.2.

A2.4 Proof of Theorem A1.1

Equipped with Theorem A2.1, we can prove Theorem A1.1 by showing that RF-GLS meets the conditions **C.1** - **C.3** holds. Proposition S1.3 in the Supplement shows that the truncation error condition **C.1** is met for any ζ_n satisfying the scalings of Assumption 4(a) required for proving the approximation and estimation error conditions.

We have already shown conditions **C.2** (Proposition A2.2) and **C.3** (Proposition A2.4) for two separate choices of function classes. The last step of the proof is choosing the function class that satisfies both conditions. Let $\mathcal{F}_n = \mathcal{F}_n(\Theta)$ be the set of all functions $f : [0,1]^D \to \mathbb{R}$ piecewise constant on each cell of the partition $\mathcal{P}_n(\Theta)$ created by an RF-GLS tree with data $\mathcal{D}_n$ and randomization Θ. We have already shown in in Proposition A2.2 that $\mathcal{F}_n$ satisfies **C.2**. To apply Proposition A2.4, $\mathcal{F}_n$ also needs to satisfy the moment condition of **C.4**. As $T_{\zeta_n}\mathcal{F}_n$ is only bounded by ζ_n which goes to ∞, clearly $\mathcal{F}_n$ will not satisfy **C.4** .

We carefully carve out a subclass $\tilde{\mathcal{F}}_n \subseteq \mathcal{F}_n$ which is still wide enough to satisfy the approximation error condition **(C.2)**, while satisfying the additional restriction **(C.4)**. For a given partition $\mathcal{P}_n(\Theta)$, we define $\tilde{\mathcal{F}}_n$ as follows:

$$\tilde{\mathcal{F}}_n = \tilde{\mathcal{F}}_n(\Theta) = \{m_n\} \cup \left\{ \cup_{\mathbf{x}_\mathcal{B} \in \mathcal{B} \in \mathcal{P}_n(\Theta)} \sum_{\mathcal{B} \in \mathcal{P}_n(\Theta)} m(\mathbf{x}_\mathcal{B}) \mathbb{I}(\mathbf{x} \in \mathcal{B}) \right\} \subseteq \mathcal{F}_n(\Theta). \tag{A7}$$

Since by construction of RF-GLS, m_n is the optimiser over a much larger set $\mathcal{F}_n(\Theta)$, trivially m_n is also the optimiser in $\tilde{\mathcal{F}}_n$. The first step of the proof of Proposition A2.2 makes it evident why Condition **C.2** will also hold for this smaller class $\tilde{\mathcal{F}}_n$. To apply Proposition A2.4 and show Condition **C.3**, the final piece is to show that the condition **(C.4)** is satisfied by $\mathcal{F}_n$. Since apart from m_n, $\mathcal{F}_n$ consists of functions that are bounded by M_0, we can have the envelope to be $F_n = |m_n| + M_0$. Hence, for Condition **(C.4)** to hold, it is enough to show $\lim_{n\to\infty} \frac{1}{n} \sum_i \mathbb{E}|m_n(X_i)|^{2+\delta} < \infty$ which is an assumption of the Theorem. □

Supplementary Materials

S1 Proofs of main results

S1.1 Approximation Error

Proof of Lemma 2.1. For any k, denote the participating leaf nodes in the partition under consideration as $\mathcal{C}_1, \mathcal{C}_2, \cdots, \mathcal{C}_{g^{(k)}+1}$. Using Assumption 2, and letting $\alpha = \|\boldsymbol{\rho}\|^2$, we have from (A1)

$$\frac{1}{n}\left(\mathbf{Z}^\top \mathbf{Q}\mathbf{Z}\right)_{l,l'} = \frac{1}{n}\mathbf{Z}_{.,l}^\top \mathbf{Q}\mathbf{Z}_{.,l'} = \frac{1}{n}\left[\alpha \sum_i \mathbf{Z}_{i,l}\mathbf{Z}_{i,l'} + \sum_{j\neq j'=0}^{q} \rho_j \rho_{j'} \sum_i \mathbf{Z}_{i-j,l}\mathbf{Z}_{i-j',l'} + O_b(1)\right] \tag{S1}$$

Here the last term is $O_b(1)$ since $\mathbf{Z}_{i,l}$'s are independent and identically distributed for $i = 1, \ldots, n$, and this term is a sum of $O(q^2)$ terms of the form $\mathbf{Z}_{i,l}\mathbf{Z}_{i',l'}$. Using strong law of large numbers,

$$\lim_{n\to\infty} \frac{1}{n}\sum_i \mathbf{Z}_{i,l}\mathbf{Z}_{i,l'} \overset{a.s.}{=} \mathbb{E}\left(\mathbf{Z}_{1,l}\mathbf{Z}_{1,l'}\right) = \mathbb{P}(X \in \mathcal{C}_l)\mathbb{I}(l = l') = Vol(\mathcal{C}_l)\mathbb{I}(l = l').$$

The last equality follows from the fact that X is uniformly distributed over $[0,1]^D (\supseteq \mathcal{C}_l)$.

Next note that the interaction terms $t_i = \mathbf{Z}_{i-j,l}\mathbf{Z}_{i-j',l}$ are identically distributed but are not independent. However, as $0 \leq j, j' \leq q$, t_i is independent of $t_{i'}$ for $i - i' > q$. Hence, $\{t_i\}$ is an *m-dependent process*, with maximum lag q Hoeffding et al. (1948). Following the hierarchy of mixing conditions (Bradley (2005), p. 112) m-dependence $\Rightarrow$ ψ-mixing $\Rightarrow$ ψ^*-mixing $\Rightarrow$ information regularity $\Rightarrow$ absolute regularity. Hence, $\{t_i\}$ is a stationary β-mixing process. Using Theorem 1 in Nobel and Dembo (1993) with the class $\mathcal{F}$ equal to the identity function, and the fact that $|\mathbf{Z}_{i-j,l}\mathbf{Z}_{i-j',l'}| \leqslant 1$, we have the following for $j \neq j'$:

$$\lim_{n\to\infty} \frac{1}{n}\sum_i \mathbf{Z}_{i-j,l}\mathbf{Z}_{i-j',l'} \overset{a.s.}{=} \mathbb{E}(\mathbf{Z}_{i-j,l}\mathbf{Z}_{i-j',l'}) \overset{ind.}{=} \mathbb{E}(\mathbf{Z}_{i-j,l})\mathbb{E}(\mathbf{Z}_{i-j',l'}) = Vol(\mathcal{C}_l)Vol(\mathcal{C}_l').$$

Letting $n \to \infty$ in both sides of (S1), we have

$$\lim_{n\to\infty} \frac{1}{n}\left(\mathbf{Z}^\top \mathbf{Q}\mathbf{Z}\right)_{l,l'} \overset{a.s.}{=} \alpha Vol(\mathcal{C}_l)\mathbb{I}(l = l') + \sum_{j\neq j'=0}^{q} \rho_j \rho_{j'} Vol(\mathcal{C}_l)Vol(\mathcal{C}_{l'}). \tag{S2}$$

Since $\frac{1}{n}\mathbf{Z}^\top\mathbf{Q}\mathbf{Z}$ is finite dimensional (i.e. $(g^{(k)}+1)\times(g^{(k)}+1)$, does not depend on n), we also have

$$\lim_{n\to\infty}\frac{1}{n}\mathbf{Z}^\top\mathbf{Q}\mathbf{Z}\overset{a.s.}{=}\alpha\times diag(\mathbf{c})+\left(\sum_{j\neq j'=0}^{q}\rho_j\rho_{j'}\right)\mathbf{c}\mathbf{c}^\top$$

where $\mathbf{c}=(c_1,c_2,\cdots,c_{g^{(k)}+1}), c_l=Vol(\mathcal{C}_l)$ and $\mathbf{1}^\top\mathbf{c}=1$.

Defining $\mathbf{c}^{-1}:=(Vol(\mathcal{C}_1)^{-1},Vol(\mathcal{C}_2)^{-1},\cdots,Vol(\mathcal{C}_{g^{(k)}+1})^{-1})$ and using Sherman–Morrison–Woodbury identity we have,

$$\begin{aligned}\left(\lim_{n\to\infty}\frac{1}{n}\mathbf{Z}^\top\mathbf{Q}\mathbf{Z}\right)^{-1}&=\alpha^{-1}diag(\mathbf{c}^{-1})-\frac{\alpha^{-2}diag(\mathbf{c}^{-1})\left(\sum_{j\neq j'=0}^{q}\rho_j\rho_{j'}\right)\mathbf{c}\mathbf{c}^\top diag(\mathbf{c}^{-1})}{1+\alpha^{-1}\left(\sum_{j\neq j'=0}^{q}\rho_j\rho_{j'}\right)\mathbf{c}^\top diag(\mathbf{c}^{-1})\mathbf{c}}\\&=\alpha^{-1}diag(\mathbf{c}^{-1})-\frac{\alpha^{-1}\left(\sum_{j\neq j'=0}^{q}\rho_j\rho_{j'}\right)\mathbf{1}\mathbf{1}^\top}{\alpha+\left(\sum_{j\neq j'=0}^{q}\rho_j\rho_{j'}\right)\mathbf{1}^\top\mathbf{c}}\\&=\alpha^{-1}\left[diag(\mathbf{c}^{-1})-\frac{\left(\sum_{j\neq j'=0}^{q}\rho_j\rho_{j'}\right)}{\alpha+\sum_{j\neq j'=0}^{q}\rho_j\rho_{j'}}\mathbf{1}\mathbf{1}^\top\right].\end{aligned}$$

Next we focus on $\frac{1}{n}\mathbf{Z}^\top\mathbf{Q}\mathbf{Y}$. Proceeding similarly, we have

$$\begin{aligned}\frac{1}{n}(\mathbf{Z}^\top\mathbf{Q}\mathbf{Y})_l&=\frac{1}{n}\left[\alpha\sum_i\mathbf{Z}_{i,l}Y_i+\sum_{j\neq j'=0}^{q}\rho_j\rho_{j'}\sum_i\mathbf{Z}_{i-j,l}Y_{i-j'}+O_b(1)\right]\\&=\alpha\frac{\sum_i\mathbf{Z}_{i,l}(m(X_i)+\epsilon_i)}{n}+\sum_{j\neq j'=0}^{q}\rho_j\rho_{j'}\frac{\sum_i\mathbf{Z}_{i-j,l}(m(X_{i-j'})+\epsilon_{i-j'})}{n}+O_b(1/n)\\&=\alpha\frac{\sum_i\mathbf{Z}_{i,l}m(X_i)}{n}+\sum_{j\neq j'=0}^{q}\rho_j\rho_{j'}\frac{\sum_i\mathbf{Z}_{i-j,l}m(X_{i-j'})}{n}\\&\quad+\alpha\frac{\sum_i\mathbf{Z}_{i,l}\epsilon_i}{n}+\sum_{j\neq j'=0}^{q}\rho_j\rho_{j'}\frac{\sum_i\mathbf{Z}_{i-j,l}\epsilon_{i-j'}}{n}+O_b(1/n)\end{aligned}\tag{S3}$$

Under Assumption 1, Lemma S3.1 with $B_i=\mathbf{Z}_{i,l}$ implies

$$\lim_{n\to\infty}\frac{1}{n}\sum_i\mathbf{Z}_{i,l}\epsilon_i\overset{a.s.}{=}0\text{ and }\lim_{n\to\infty}\frac{\sum_i\mathbf{Z}_{i-j,l}\epsilon_{i-j'}}{n}\overset{a.s.}{=}0.$$

Since $\mathbf{Z}_{i,l}m(X_i)$'s are i.i.d copies of each other, using strong law of large numbers, we have

$$\begin{aligned}
&\lim_{n\to\infty}\frac{1}{n}\sum_i \mathbf{Z}_{i,l}m(X_i)\\
&\overset{a.s.}{=}\mathbb{E}\left(\mathbf{Z}_{1,l}m(X_1)\right)\\
&=\mathbb{E}\left(\mathbf{Z}_{1,l}m(X_1)\right)+\mathbb{E}\left(\mathbf{Z}_{1,l}\epsilon_1\right);\ [X\perp\!\!\!\perp\epsilon,\mathbb{E}\left(\mathbf{Z}_{1,l}\epsilon_1\right)=\mathbb{E}\left(\mathbf{Z}_{1,l}\right)\mathbb{E}\left(\epsilon_1\right)=0]\\
&=\mathbb{E}\left(\mathbf{Z}_{1,l}(m(X_1)+\epsilon_1)\right))\\
&=\mathbb{E}\left(\mathbf{Z}_{1,l}Y_1\right)\\
&=\mathbb{E}\left(Y|X\in\mathcal{C}_l\right)Vol(\mathcal{C}_l).
\end{aligned}$$

Similarly, for $j\neq j'$, using *m-dependence* we can show

$$\lim_{n\to\infty}\frac{1}{n}\sum_i \mathbf{Z}_{i-j,l}m(X_{i-j'})\overset{a.s.}{=}\mathbb{E}\left(Y\right)Vol(\mathcal{C}_l).$$

Applying all the limits to (S3), we have

$$\begin{aligned}
\hat{r}_l:=\lim_{n\to\infty}\frac{1}{n}(\mathbf{Z}^\top\mathbf{Q}\mathbf{Y})_l&\overset{a.s.}{=}\alpha\mathbb{E}(Y|X\in\mathcal{C}_l)Vol(\mathcal{C}_l)+Vol(\mathcal{C}_l)\mathbb{E}(Y)\sum_{j\neq j'=0}^{q}\rho_j\rho_{j'}\\
&=\alpha\mathbb{E}(Y\mathbb{I}(X\in\mathcal{C}_l))+Vol(\mathcal{C}_l)\mathbb{E}(Y)\sum_{j\neq j'=0}^{q}\rho_j\rho_{j'}.
\end{aligned}\tag{S4}$$

Finally, defining $\hat{\mathbf{r}}:=\left(\hat{r}_1,\cdots,\hat{r}_{g^{(k)}+1}\right)$, and noting that the dimension of $\frac{1}{n}\mathbf{Z}^\top\mathbf{Q}\mathbf{Z}$ does not grow with n, we have

$$\begin{aligned}
\lim_{n\to\infty}\hat{\boldsymbol{\beta}}&=\lim_{n\to\infty}\left(\frac{1}{n}\mathbf{Z}^\top\mathbf{Q}\mathbf{Z}\right)^{-1}\lim_{n\to\infty}\frac{1}{n}\left(\mathbf{Z}^\top\mathbf{Q}\mathbf{Y}\right)\\
&=\alpha^{-1}\left[diag(\mathbf{c}^{-1})-\frac{\left(\sum_{j\neq j'=0}^{q}\rho_j\rho_{j'}\right)}{\alpha+\sum_{j\neq j'=0}^{q}\rho_j\rho_{j'}}\mathbf{1}\mathbf{1}^\top\right]\hat{\mathbf{r}}\\
&=\alpha^{-1}\left[diag(\mathbf{c}^{-1})\hat{\mathbf{r}}-\frac{\left(\sum_{j\neq j'=0}^{q}\rho_j\rho_{j'}\right)}{\alpha+\sum_{j\neq j'=0}^{q}\rho_j\rho_{j'}}\mathbf{1}\mathbf{1}^\top\hat{\mathbf{r}}\right]\\
&=\alpha^{-1}\left[\hat{\mathbf{o}}-\frac{\left(\sum_{j\neq j'=0}^{q}\rho_j\rho_{j'}\right)}{\alpha+\left(\sum_{j\neq j'=0}^{q}\rho_j\rho_{j'}\right)}\mathbf{1}\left(\alpha\mathbb{E}(Y)+\mathbb{E}(Y)\sum_{j\neq j'=0}^{q}\rho_j\rho_{j'}\right)\right],
\end{aligned}$$

where $\hat{\mathbf{o}} := \left(\hat{o}_1, \hat{o}_2, \cdots, \hat{o}_{g^{(k)}+1}\right); \hat{o}_l = \hat{r}_l/c_l = \alpha\mathbb{E}(Y|X \in \mathcal{C}_l) + \mathbb{E}(Y)\sum_{j\neq j'=0}^{q}\rho_j\rho_{j'}$. Hence

$$\begin{aligned}\lim_{n\to\infty}\hat{\boldsymbol{\beta}}_l &\overset{a.s.}{=} \alpha^{-1}\hat{o}_l - \mathbb{E}(Y)\frac{(\sum_{j\neq j'=0}^{q}\rho_j\rho_{j'})(\alpha+\sum_{j\neq j'=0}^{q}\rho_j\rho_{j'})}{\alpha(\alpha+\sum_{j\neq j'=0}^{q}\rho_j\rho_{j'})}\\ &= \mathbb{E}(Y|X\in\mathcal{C}_l) + \mathbb{E}(Y)\left(\frac{\sum_{j\neq j'=0}^{q}\rho_j\rho_{j'}}{\alpha} - \frac{\sum_{j\neq j'=0}^{q}\rho_j\rho_{j'}}{\alpha}\right)\\ &= \mathbb{E}(Y|X\in\mathcal{C}_l).\end{aligned}$$

This completes the proof of Lemma 2.1. □

Proof of Theorem 2.1. With $k = k+1$, let $\mathcal{C}_l = \mathcal{C}_l^{(k)}$ for $l < g^{(k)}$; $\mathcal{C}_{g^{(k)}} = \mathcal{C}_{g^{(k)}}^{(k+1)}$, $\mathcal{C}_{g^{(k)}+1} = \mathcal{C}_{g^{(k)}+1}^{(k+1)}$. Since $\left(\mathbf{Y} - \mathbf{Z}\hat{\boldsymbol{\beta}}(\mathbf{Z})\right)^\top \mathbf{Q}\left(\mathbf{Y} - \mathbf{Z}\hat{\boldsymbol{\beta}}(\mathbf{Z})\right) = \mathbf{Y}^\top\mathbf{Q}\mathbf{Y} - \mathbf{Y}^\top\mathbf{Q}\mathbf{Z}\hat{\boldsymbol{\beta}}(\mathbf{Z})$, we have

$$v_{n,\mathbf{Q}}(\mathfrak{C}^{(k)}, l_1, (d,c)) = \frac{1}{n}\left(\mathbf{Y}^\top\mathbf{Q}\mathbf{Z}\hat{\boldsymbol{\beta}}(\mathbf{Z}) - \mathbf{Y}^\top\mathbf{Q}\mathbf{Z}^{(0)}\hat{\boldsymbol{\beta}}(\mathbf{Z}^{(0)})\right)$$

where $\mathbf{Z}_{i,l}^{(0)} = \mathbb{I}\left(\mathbf{x}_i \in \mathcal{C}_l\right)$; for $l = 1, 2, \cdots, g^{(k)} - 1$, and $\mathbf{Z}_{i,g^{(k)}}^{(0)} = \mathbb{I}\left(\mathbf{x}_i \in \mathcal{C}_{g^{(k)}} \cup \mathcal{C}_{g^{(k)}+1}\right)$. We use the notations $\mathcal{B} = \mathcal{B}^L \cup \mathcal{B}^R$, where $\mathcal{B}^L = \mathcal{C}_{g^{(k)}}$ and $\mathcal{B}^R = \mathcal{C}_{g^{(k)}+1}$.

From Lemma 2.1, we have $\lim_{n\to\infty}\frac{1}{n}\mathbf{Y}^\top\mathbf{Q}\mathbf{Z}\hat{\boldsymbol{\beta}}(\mathbf{Z}) \overset{a.s.}{=} \hat{\mathbf{r}}^\top\hat{\mathbf{b}}$, where $\hat{\mathbf{b}} := (\hat{b}_1, \hat{b}_2, \cdots, \hat{b}_{g^{(k)}+1}); \hat{b}_l = \mathbb{E}(Y|X\in\mathcal{C}_l)$ and $\hat{\mathbf{r}}$ is defined in (S4). Now

$$\begin{aligned}\hat{\mathbf{r}}^\top\hat{\mathbf{b}} &= \sum_{l=1}^{g^{(k)}+1}\mathbb{E}(Y|X\in\mathcal{C}_l)\left(\alpha\mathbb{E}(Y|X\in\mathcal{C}_l)Vol(\mathcal{C}_l) + Vol(\mathcal{C}_l)\mathbb{E}(Y)\sum_{j\neq j'=0}^{q}\rho_j\rho_{j'}\right)\\ &= \sum_{l=1}^{g^{(k)}+1}\left(\alpha\mathbb{E}(Y|X\in\mathcal{C}_l)^2Vol(\mathcal{C}_l) + Vol(\mathcal{C}_l)\mathbb{E}(Y)\mathbb{E}(Y|X\in\mathcal{C}_l)\sum_{j\neq j'=0}^{q}\rho_j\rho_{j'}\right)\\ &= \sum_{l=1}^{g^{(k)}+1}\alpha\mathbb{E}(Y|X\in\mathcal{C}_l)^2Vol(\mathcal{C}_l) + \mathbb{E}(Y)^2\sum_{j\neq j'=0}^{q}\rho_j\rho_{j'}.\end{aligned}$$

Substituting this in the expression of asymptotic value of $v_{n,\mathbf{Q}}(\mathfrak{C}^{(k)}, l_1, (d,c))$, we have

$$
\begin{aligned}
&\lim_{n\to\infty} v_{n,\mathbf{Q}}(\mathfrak{C}^{(k)}, l_1, (d,c)) \\
&\overset{a.s.}{=} \sum_{l=g^{(k)}}^{g^{(k)}+1} \alpha \mathbb{E}(Y|X \in \mathcal{C}_l)^2 Vol(\mathcal{C}_l) - \alpha \mathbb{E}(Y|X \in \mathcal{C}_{g^{(k)}} \cup \mathcal{C}_{g^{(k)}+1})^2 Vol(\mathcal{C}_{g^{(k)}} \cup \mathcal{C}_{g^{(k)}+1}) \\
&= \alpha \left(\sum_{l=g^{(k)}}^{g^{(k)}+1} \mathbb{E}(Y|X \in \mathcal{C}_l)^2 Vol(\mathcal{C}_l) - \mathbb{E}(Y|X \in \mathcal{B})^2 Vol(\mathcal{B}) \right) \\
&= \alpha \Bigg(\left[\mathbb{E}(Y^2|X \in \mathcal{B}) - \mathbb{E}(Y|X \in \mathcal{B})^2 \right] Vol(\mathcal{B}) \\
&\qquad - \left[\mathbb{E}(Y^2|X \in \mathcal{B}) Vol(\mathcal{B}) - \sum_{l=g^{(k)}}^{g^{(k)}+1} \mathbb{E}(Y|X \in \mathcal{C}_l)^2 Vol(\mathcal{C}_l) \right] \Bigg) \\
&= \alpha \left(\mathbb{V}(Y|X \in \mathcal{B}) Vol(\mathcal{B}) - \left[\mathbb{E}(Y^2 \mathbb{I}(X \in \mathcal{B})) - \sum_{l=g^{(k)}}^{g^{(k)}+1} \mathbb{E}(Y|X \in \mathcal{C}_l)^2 Vol(\mathcal{C}_l) \right] \right) \\
&= \alpha \Bigg(\mathbb{V}(Y|X \in \mathcal{B}) Vol(\mathcal{B}) \\
&\qquad - \left[\mathbb{E}(Y^2 \mathbb{I}(X \in \mathcal{C}_{g^{(k)}} \cup \mathcal{C}_{g^{(k)}+1})) - \sum_{l=g^{(k)}}^{g^{(k)}+1} \mathbb{E}(Y|X \in \mathcal{C}_l)^2 Vol(\mathcal{C}_l) \right] \Bigg) \\
&= \alpha \left(\mathbb{V}(Y|X \in \mathcal{B}) Vol(\mathcal{B}) - \sum_{l=g^{(k)}}^{g^{(k)}+1} \mathbb{V}(Y|X \in \mathcal{C}_l) Vol(\mathcal{C}_l) \right) \\
&= \alpha Vol(\mathcal{B}) \Big[\mathbb{V}(Y|\mathbf{X} \in \mathcal{B}) - \mathbb{P}(\mathbf{X} \in \mathcal{B}^R|\mathbf{X} \in \mathcal{B}) \mathbb{V}(Y|\mathbf{X} \in \mathcal{B}^R) \\
&\qquad - \mathbb{P}(\mathbf{X} \in \mathcal{B}^L|\mathbf{X} \in \mathcal{B}) \mathbb{V}(Y|\mathbf{X} \in \mathcal{B}^L) \Big].
\end{aligned}
$$

This completes the proof of Theorem 2.1. □

Proof of Proposition A2.1. We first introduce some additional notation on splits.

Notation of splits

We introduce some additional notations of splits. The split associated with a specific node (a subset of the feature space) indicates the direction and the cutoff associated with its partition. A split is denoted by $s = (d, c)$, where d denotes the direction of the split (the feature along which the aforementioned split is performed) $\in \{1, 2, \cdots, D\}$ and c is the cutoff value of the split. Let the

complete set of nodes in level k is $\mathfrak{C}^{(k)} = \{\mathcal{C}_1^{(k)}, \cdots, \mathcal{C}_{g^{(k)}}^{(k)}\}$. Let the split of the l^{th} node of k^{th} level i.e. $\mathcal{C}_l^{(k)}$ be denoted by $s_l^{(k)}$. We observe that $\mathfrak{C}^{(k+1)}$ is determined by $\mathfrak{C}^{(k)}$ and $\mathcal{S}^{(k)}$ where $\mathcal{S}^{(k)} = \{s_1^{(k)}, s_2^{(k)}, \cdots, s_{g^{(k)}}^{(k)}\}$ is the set of splits on the partitions in level k to create the partitions at level $k+1$. This in turn implies that $\mathfrak{C}^{(k+1)}$ is determined by $\tilde{\mathbf{s}}_k = \{\mathcal{S}^{(1)}, \mathcal{S}^{(2)}, \cdots, \mathcal{S}^{(k)}\}$, as by definition, $\mathcal{C}^{(1)} = [0,1]^D$. We define the set of all possible such $\tilde{\mathbf{s}}_k$ to be $\tilde{\mathfrak{S}}_k$.

For any fixed $\mathbf{x} \in [0,1]^D$ and $k \geqslant 1$, $\mathfrak{S}_k(\mathbf{x})$ to is the set of all possible splits that built the node containing $\mathbf{x}$ in $k{+}1^{th}$ level. Members of $\mathfrak{S}_k(\mathbf{x})$ are denoted as $\mathbf{s}_k := \mathbf{s}_k(\mathbf{x}) = (\mathcal{S}^{(1)}, \mathcal{S}^{(2)}, \cdots, \mathcal{S}^{(k-1)}, s_{(\mathbf{x})}^{(k)}) = (\tilde{\mathbf{s}}_{k-1}, s_{(\mathbf{x})}^{(k)})$, where $\tilde{\mathbf{s}}_{k-1} \in \tilde{\mathfrak{S}}_{k-1}$ and $s_{(\mathbf{x})}^{(k)}$ denotes the split associated with the node at level k, containing $\mathbf{x}$.

The node at level $k+1$, containing $\mathbf{x}$, built with $\mathbf{s}_k$ is $\mathcal{B}(\mathbf{x}, \mathbf{s}_k)$. The node containing $\mathbf{x}$ in a RF-GLS tree built with random parameter Θ and data $\mathcal{D}_n$ is denoted by $\mathcal{B}_n(\mathbf{x}, \Theta)$. The optimal splits obtained from empirical DART-split criterion, that build the node containing $\mathbf{x}$ at level $k+1$ of RF-GLS tree (built with n points and randomness Θ) is denoted as $\hat{\mathbf{s}}_{k,n}(\mathbf{x}, \Theta)$. The distance between $\mathbf{s}_k^{(1)}, \mathbf{s}_k^{(2)} \in \mathfrak{S}_k(\mathbf{x})$ is the $\mathbb{L}_\infty$ norm of their difference; i.e. $\|\mathbf{s}_k^{(1)} - \mathbf{s}_k^{(2)}\|_\infty$. The distance between $\mathbf{s}_k \in \mathfrak{S}_k(\mathbf{x})$ and $\acute{\mathfrak{S}}_k \subseteq \mathfrak{S}_k(\mathbf{x})$ is $dist(\mathbf{s}_k, \acute{\mathfrak{S}}_k) = \inf_{\mathbf{s} \in \acute{\mathfrak{S}}_k} \|\mathbf{s}_k - \mathbf{s}\|_\infty$.

Next, for a fixed $\mathbf{x} \in [0,1]^D$, and any k levels of split $\mathbf{s}_k$, we define $v_{n,k,\mathbf{Q}}(\mathbf{x}, \mathbf{s}_k)$ to be the DART split criterion (5) to maximise in $s^{(k)}$ of $\mathbf{s}_k$, i.e. the final k^{th} level split of $\mathcal{B}(\mathbf{x}, \mathbf{s}_{k-1})$. For all $\varepsilon > 0$, we define $\tilde{\mathfrak{S}}_{k-1}^{\varepsilon} \subset \tilde{\mathfrak{S}}_{k-1}$, the set of all splits of the $(k-1)$ level nodes, such that each node in $\mathfrak{C}^{(k)}$ contains a hypercube of edge length ε. Additionally, we define $\bar{\mathfrak{S}}_k^{\varepsilon}(\mathbf{x}) = \{\mathbf{s}_k := \mathbf{s}_k(\mathbf{x}) : \tilde{\mathbf{s}}_{k-1} \in \tilde{\mathfrak{S}}_{k-1}^{\varepsilon}\}$.

Equipped with the notation. We state a more technical version of equicontinuity result in Proposition A2.1. The proof is deferred to Section S1.1.

Technical statement of Proposition A2.1: Under Assumptions 2, 3, and 4(b) and 4(c), for fixed $\mathbf{x}$, $k \in \mathbb{N}$ and $\varepsilon > 0$, $v_{n,k,\mathbf{Q}}(\mathbf{x}, \mathbf{s}_k(\mathbf{x}))$ is stochastically equicontinuous with respect to $\mathbf{s}_k$ on $\bar{\mathfrak{S}}_k^{\varepsilon}(\mathbf{x})$, i.e. $\forall \phi, \pi > 0$, $\exists\, \delta > 0$, i.e.,

$$\lim_{n\to\infty} \mathbb{P}\left[\sup_{\substack{\|\mathbf{s}_k^{(1)} - \mathbf{s}_k^{(2)}\|_\infty \leqslant \delta \\ \mathbf{s}_k^{(1)}, \mathbf{s}_k^{(2)} \in \bar{\mathfrak{S}}_k^{\varepsilon}(\mathbf{x})}} |v_{n,k,\mathbf{Q}}(\mathbf{x}, \mathbf{s}_k^{(1)}) - v_{n,k,\mathbf{Q}}(\mathbf{x}, \mathbf{s}_k^{(2)})| > \phi\right] \leqslant \pi.$$

We will show that $v_{n,k,\mathbf{Q}}(\mathbf{x}, \mathbf{s}_k)$ (defined in Section S1.1) is stochastically equicontinuous with respect to $\mathbf{s}_k$ for all $\mathbf{x} \in [0,1]^D$, provided the volumes of leaf nodes in the previous level are not arbitrarily close to 0.

By the Glivenko-Cantelli theorem Loeve (1977), $\exists n_3 \in \mathbb{N}$, such that for all $n > n_3$, and $\mathcal{C} \subseteq [0,1]^D$, we have with probability at least $1 - \pi/4$,

$$Vol(\mathcal{C}) - \delta^2 \leqslant \frac{1}{n}\sum_i \mathbb{I}(X_i \in \mathcal{C}) \leqslant Vol(\mathcal{C}) + \delta^2. \tag{S5}$$

As each split can be chosen from at-most n^D candidates, the collection $\mathcal{P}$ of all possible nodes created by splits upto level k is of polynomial-in-n cardinality. Since Assumptions 4(b) holds for some n_2 and $\pi = \pi/8$, and 4(c) holds, then Lemma S3.2 holds for $\mathcal{P}$ for some n_3 and $\pi = \pi/4$. For the rest of the proof, we consider $n > \max\{n_0, n_2, n_3\}$ and restrict ourselves to the set Ω_n where all of these assertions (Equation S5, Assumption 4(b) and Lemma S3.2) hold, which occurs with probability at-least $1 - 3\pi/4$.

From the definition of the distance in $\mathfrak{S}_k$, if $\mathbf{s}_k^{(1)}, \mathbf{s}_k^{(2)} \in \mathfrak{S}_k$ satisfy $\|\mathbf{s}_k^{(1)} - \mathbf{s}_k^{(2)}\|_\infty < 1$, then the split directions are identical. So we can always consider $\delta < 1$. Since we consider two sets of splits $\mathbf{s}_k^{(1)}$ and $\mathbf{s}_k^{(2)}$, for convenience of notation, we use $\ddot{\mathcal{C}}_l^{(h)}$ to denote the l^{th} leaf node of the partition induced by $\mathbf{s}_k^{(h)}$, for $h = 1, 2$ and $l = 1, \ldots, g^{(k)} + 1$, i.e, $\ddot{\mathcal{C}}_l^{(h)}$ is the node $\mathcal{C}_l^{(k)}$ for the h^{th} set of splits. Also, we will be using the notation $\mathbf{Z}_1$ and $\mathbf{Z}_2$ instead of $\mathbf{Z}(\mathbf{s}_k^{(1)})$ and $\mathbf{Z}(\mathbf{s}_k^{(2)})$ respectively. Along the same line as proof of Theorem 2.1, we can write

$$\begin{aligned} |v_{n,k,\mathbf{Q}}(\mathbf{x}, \mathbf{s}_k^{(1)}) - v_{n,k,\mathbf{Q}}(\mathbf{x}, \mathbf{s}_k^{(2)})| = &\frac{1}{n}|\mathbf{Y}^\top \mathbf{Q}\mathbf{Z}_1 \hat{\boldsymbol{\beta}}(\mathbf{Z}_1) - \mathbf{Y}^\top \mathbf{Q}\mathbf{Z}_2 \hat{\boldsymbol{\beta}}(\mathbf{Z}_2)| \\ &+ \frac{1}{n}|\mathbf{Y}^\top \mathbf{Q}\mathbf{Z}_1^{(0)} \hat{\boldsymbol{\beta}}(\mathbf{Z}_1^{(0)}) - \mathbf{Y}^\top \mathbf{Q}\mathbf{Z}_2^{(0)} \hat{\boldsymbol{\beta}}(\mathbf{Z}_2^{(0)})| \end{aligned}, \tag{S6}$$

where $\mathbf{Z}_{h_{i,l}}^{(0)} = \mathbb{I}\left(\mathbf{x}_i \in \ddot{\mathcal{C}}_l^{(h)}\right); l = 1, 2, \cdots, g^{(k)} - 1$, and $\mathbf{Z}_{h_{i,g^{(k)}}}^{(0)} = \mathbb{I}\left(\mathbf{x}_i \in \ddot{\mathcal{C}}_{g^{(k)}}^{(h)} \cup \ddot{\mathcal{C}}_{g^{(k)}+1}^{(h)}\right)$ for $h = 1, 2$.

We first focus on $\frac{1}{n}|\mathbf{Y}^\top \mathbf{Q}\mathbf{Z}_1 \hat{\boldsymbol{\beta}}(\mathbf{Z}_1) - \mathbf{Y}^\top \mathbf{Q}\mathbf{Z}_2 \hat{\boldsymbol{\beta}}(\mathbf{Z}_2)| = \frac{1}{n}|\mathbf{Y}^\top \mathbf{Q}^{\frac{1}{2}} \left[\dot{\mathbf{P}}_{\mathbf{Z}_1} - \dot{\mathbf{P}}_{\mathbf{Z}_2}\right] \mathbf{Q}^{\frac{\top}{2}} \mathbf{Y}|$ where,

$$\mathbf{Q}^{\frac{1}{2}} = \boldsymbol{\Sigma}^{-\frac{T}{2}}; \; \dot{\mathbf{P}}_{\mathbf{Z}} = \mathbf{Q}^{\frac{\top}{2}} \mathbf{Z} \left[\mathbf{Z}^\top \mathbf{Q}\mathbf{Z}\right]^{-1} \mathbf{Z}^\top \mathbf{Q}^{\frac{1}{2}}; \; \mathbf{Z} \in \{\mathbf{Z}_1, \mathbf{Z}_2\},$$

and consider the two possible scenarios:

- **R1**: $\max\left(\frac{\min\left(Vol(\ddot{\mathcal{C}}^{(1)}_{g^{(k)}}),Vol(\ddot{\mathcal{C}}^{(1)}_{g^{(k)}+1})\right)}{Vol(\ddot{\mathcal{C}}^{(1)}_{g^{(k)}})+Vol(\ddot{\mathcal{C}}^{(1)}_{g^{(k)}+1})}, \frac{\min\left(Vol(\ddot{\mathcal{C}}^{(2)}_{g^{(k)}}),Vol(\ddot{\mathcal{C}}^{(2)}_{g^{(k)}+1})\right)}{Vol(\ddot{\mathcal{C}}^{(2)}_{g^{(k)}})+Vol(\ddot{\mathcal{C}}^{(2)}_{g^{(k)}+1})}\right) \geqslant \sqrt{\delta}$
- **R2**: $\max\left(\frac{\min\left(Vol(\ddot{\mathcal{C}}^{(1)}_{g^{(k)}}),Vol(\ddot{\mathcal{C}}^{(1)}_{g^{(k)}+1})\right)}{Vol(\ddot{\mathcal{C}}^{(1)}_{g^{(k)}})+Vol(\ddot{\mathcal{C}}^{(1)}_{g^{(k)}+1})}, \frac{\min\left(Vol(\ddot{\mathcal{C}}^{(2)}_{g^{(k)}}),Vol(\ddot{\mathcal{C}}^{(2)}_{g^{(k)}+1})\right)}{Vol(\ddot{\mathcal{C}}^{(2)}_{g^{(k)}})+Vol(\ddot{\mathcal{C}}^{(2)}_{g^{(k)}+1})}\right) < \sqrt{\delta}$

Scenario **R1** happens when at least for one of the two sets of splits, both the new child nodes are "significantly different" from their parent node; i.e. their volumes are bounded away from the volume of the parent node and zero. Here, we will show equicontinuity by exploiting perturbation bounds on orthogonal projections Chen et al. (2016). The other possibility is Scenario **R2** where for both the set of splits, the volume of the larger child node is arbitrary close to that of parent node. Here, we prove equicontinuity by showing that the DART-split criterion value asymptotically vanishes.

Without loss of generality, we consider $\delta > 0$ small enough such that, $\sqrt{\delta} < 1 - \sqrt{\delta}$. Under **R1**, we have

$$\frac{1}{n}|\mathbf{Y}^\top \mathbf{Q}^{\frac{1}{2}}\left[\dot{\mathbf{P}}_{\mathbf{Z}_1} - \dot{\mathbf{P}}_{\mathbf{Z}_2}\right]\mathbf{Q}^{\frac{\top}{2}}\mathbf{Y}| \leqslant \frac{1}{n}\mathbf{Y}^\top\mathbf{Q}\mathbf{Y}\|\dot{\mathbf{P}}_{\mathbf{Z}_1} - \dot{\mathbf{P}}_{\mathbf{Z}_2}\|_2. \tag{S7}$$

Defining $\mathbf{H}_{\mathbf{Z}} = \left[\mathbf{Z}^\top\mathbf{Q}\mathbf{Z}\right]^{-1}\mathbf{Z}^\top\mathbf{Q}^{\frac{1}{2}}$ for $\mathbf{Z} = \mathbf{Z}_1, \mathbf{Z}_2$ and using the perturbation bounds on projection operators from Chen et al. (2016) Theorem 1.2 (1.8), we have

$$\|\dot{\mathbf{P}}_{\mathbf{Z}_1} - \dot{\mathbf{P}}_{\mathbf{Z}_2}\|_2 \leqslant \min\{\|\mathbf{H}_{\mathbf{Z}_1}\|_2, \|\mathbf{H}_{\mathbf{Z}_2}\|_2\}\|\mathbf{Q}^{\frac{\top}{2}}\mathbf{Z}_1 - \mathbf{Q}^{\frac{\top}{2}}\mathbf{Z}_2\|_2. \tag{S8}$$

By definition of matrix $\mathbb{L}_2$ norm, $\|\mathbf{Q}^{\frac{\top}{2}}\mathbf{Z}_1 - \mathbf{Q}^{\frac{\top}{2}}\mathbf{Z}_2\|_2 \leq \lambda^{\frac{1}{2}}_{\max}(\mathbf{Q})\|\mathbf{Z}_1 - \mathbf{Z}_2\|_2$. Since Assumption 2 implies $\lambda_{\max}(\mathbf{Q})$ is bounded, we focus on $\ddot{\mathbf{D}} = \mathbf{Z}_1 - \mathbf{Z}_2$. By Gershgorin circle theorem,

$$\begin{aligned}
\lambda_{\max}(\ddot{\mathbf{D}}^2) &\leqslant \max_{1\leqslant l_1\leqslant g^{(k)}+1}\sum_{l_2=1}^{g^{(k)}+1}|(\ddot{\mathbf{D}}^2)_{l_1,l_2}| \\
&= \max_{1\leqslant l_1\leqslant g^{(k)}+1}\sum_{l_2=1}^{g^{(k)}+1}\sum_i|\ddot{\mathbf{D}}_{i,l_1}\ddot{\mathbf{D}}_{i,l_2}| \\
&\leqslant \max_{1\leqslant l_1\leqslant g^{(k)}+1}\sum_i|\ddot{\mathbf{D}}_{i,l_1}|\sum_{l_2=1}^{g^{(k)}+1}|\ddot{\mathbf{D}}_{i,l_2}| \\
&\leqslant 2\max_{1\leqslant l_1\leqslant g^{(k)}+1}\sum_i|\ddot{\mathbf{D}}_{i,l_1}|.
\end{aligned}$$

The last inequality follows from the fact that the $\mathbf{Z}_1$ and $\mathbf{Z}_2$ are binary matrices whose row sums are 1. Now

$$\sum_i |\ddot{\mathbf{D}}_{i,l}| = \sum_i \mathbb{I}(\mathbf{x}_i \in \ddot{\mathcal{C}}_l^{(1)} \triangle \ddot{\mathcal{C}}_l^{(2)}) \text{ where } \ddot{\mathcal{C}}_l^{(1)} \triangle \ddot{\mathcal{C}}_l^{(2)} = \left(\ddot{\mathcal{C}}_l^{(1)} \cup \ddot{\mathcal{C}}_l^{(2)}\right) \cap \left(\ddot{\mathcal{C}}_l^{(1)} \cap \ddot{\mathcal{C}}_l^{(2)}\right)^c.$$

From Algorithm 1, $\ddot{\mathcal{C}}_l^{(1)}$ and $\ddot{\mathcal{C}}_l^{(2)}$ are both D-dimensional boxes since both of them are Cartesian product of D intervals. Let

$$\ddot{\mathcal{C}}_l^{(h)} = [\breve{a}_1^{(h)}, \breve{b}_1^{(h)}] \times \cdots \times [\breve{a}_D^{(h)}, \breve{b}_D^{(h)}] \subseteq [0,1]^D; \breve{a}_d^{(h)} < \breve{b}_d^{(h)}; \forall d \in \{1, 2, \cdots, D\}; h = 1, 2.$$

Then

$$|\mathbf{s}_k^{(1)} - \mathbf{s}_k^{(2)}| \leqslant \delta \text{ implies } |\breve{a}_d^{(1)} - \breve{a}_d^{(2)}| \leqslant \delta; |\breve{b}_d^{(1)} - \breve{b}_d^{(2)}| \leqslant \delta; \forall d \in \{1, 2, \cdots, D\}.$$

Without loss of generality, we assume $Vol(\ddot{\mathcal{C}}_l^{(1)}) \leqslant Vol(\ddot{\mathcal{C}}_l^{(2)})$. One scenario where $Vol(\ddot{\mathcal{C}}_l^{(1)} \triangle \ddot{\mathcal{C}}_l^{(2)})$ is maximised is

$$\breve{a}_d^{(2)} = \breve{a}_d^{(1)} - \delta; \;\; \breve{b}_d^{(2)} = \breve{b}_d^{(1)} + \delta; \;\; \forall d \in \{1, 2, \cdots, D\}.$$

Hence we have

$$\begin{aligned} Vol(\ddot{\mathcal{C}}_l^{(1)} \triangle \ddot{\mathcal{C}}_l^{(2)}) &\leqslant \prod_{d=1}^{D} |\breve{b}_d^{(2)} - \breve{a}_d^{(2)}| - \prod_{d=1}^{D} |\breve{b}_d^{(2)} - \breve{a}_d^{(2)} - 2\delta| \\ &\leqslant \sum_{d=1}^{D} \binom{D}{d} (2\delta)^d \\ &\leqslant \delta 2^D \left(\sum_{d=1}^{D} \binom{D}{d}\right); [\text{ as } \delta < 1] \\ &\leqslant 2^{2D}\delta \end{aligned}$$

By (S5) on Ω_n, $\frac{1}{n}\sum_i |\ddot{\mathbf{D}}_{i,l}| \leqslant Vol(\ddot{\mathcal{C}}_l^{(1)} \triangle \ddot{\mathcal{C}}_l^{(2)}) + \delta^2 \leqslant 2^{2D}\delta + \delta^2 = O(\delta)$. Therefore $\frac{1}{n}\lambda_{\max}(\ddot{\mathbf{D}}^2) = \frac{1}{n}\|\mathbf{Z}_1 - \mathbf{Z}_2\|_2^2 = O_p(\delta)$. Plugging this into (S8), we have

$$\|\mathbf{Q}^{\frac{\top}{2}}\mathbf{Z}_1 - \mathbf{Q}^{\frac{\top}{2}}\mathbf{Z}_2\|_2 = O_p(\sqrt{n\delta}). \tag{S9a}$$

For the other component in (S8), i.e. $\min\{\|\mathbf{H}_{\mathbf{Z}_1}\|_2, \|\mathbf{H}_{\mathbf{Z}_2}\|_2\}$, for $\mathbf{Z} \in \{\mathbf{Z}_1, \mathbf{Z}_2\}$, we have

$$\begin{aligned}\|\mathbf{H}_{\mathbf{Z}}\|_2 &= \sqrt{\lambda_{max}((\mathbf{Q}^{\frac{T}{2}}\mathbf{Z}\,[\mathbf{Z}^\top\mathbf{Q}\mathbf{Z}]^{-1})(\mathbf{Q}^{\frac{T}{2}}\mathbf{Z}\,[\mathbf{Z}^\top\mathbf{Q}\mathbf{Z}]^{-1})^\top)} \\ &= \sqrt{\lambda_{\max}([\mathbf{Z}^\top\mathbf{Q}\mathbf{Z}]^{-1})} = \Big(\lambda_{\min}(\mathbf{Z}^\top\mathbf{Q}\mathbf{Z})\Big)^{-1/2}.\end{aligned}$$

From Gershgorin circle theorem Loeve (1977), we have

$$\lambda_{\min}(\Big[\mathbf{Z}^\top\mathbf{Q}\mathbf{Z}\Big]) \geqslant \min_{1\leqslant l_1\leqslant g^{(k)}+1}\Big\{(\mathbf{Z}^\top\mathbf{Q}\mathbf{Z})_{l_1,l_1} - \sum_{l_2\neq l_1}|(\mathbf{Z}^\top\mathbf{Q}\mathbf{Z})_{l_1,l_2}|\Big\}.$$

Using Lemma S3.3 and the diagonal dominance from Assumptions 2 and 3,

$$\lambda_{\min}(\Big[\mathbf{Z}^\top\mathbf{Q}\mathbf{Z}\Big]) \geq \xi\sum_i \mathbf{Z}_{i,l1} \text{ where } \xi = \min_{1\leq i\leq q+1}(\mathbf{Q}_{ii} - \sum_{j\neq i}|\mathbf{Q}_{ij}|) \tag{S9b}$$

is just a positive constant only dependent on $\boldsymbol{\rho}$ and $\mathbf{L}$. Hence we have

$$\lambda^{-1}_{\min}(\Big[\mathbf{Z}_h^\top\mathbf{Q}\mathbf{Z}_h\Big]) \leqslant \frac{1}{\xi}\frac{1}{\min_{1\leqslant l\leqslant g^{(k)}+1}|\ddot{\mathcal{C}}_l^{(h)}|}, h = 1, 2. \tag{S9c}$$

By (S5) on Ω_n, $|\ddot{\mathcal{C}}_l^{(h)}|/n \geqslant Vol(\ddot{\mathcal{C}}_l^{(h)}) - \delta^2$. As $\mathbf{s}_k^{(1)}, \mathbf{s}_k^{(2)} \in \bar{\mathfrak{S}}_k^{\varepsilon}(\mathbf{x})$, by definition of $\bar{\mathfrak{S}}_k^{\varepsilon}(\mathbf{x})$, $\tilde{\mathbf{s}}_{k-1}^{(1)}, \tilde{\mathbf{s}}_{k-1}^{(2)} \in \tilde{\mathfrak{S}}_{k-1}^{\varepsilon}$. As each node of $\mathfrak{C}^{(k)}$ corresponding to both these splits contains a hypercube of edge length ε, we have

$$Vol(\ddot{\mathcal{C}}_l)^{(h)} \geqslant \varepsilon^D;\ \forall i = 1, 2, \cdots, g^{(k)} - 1; Vol(\ddot{\mathcal{C}}_{g^{(k)}}^{(h)}) + Vol(\ddot{\mathcal{C}}_{g^{(k)}+1}^{(h)}) \geqslant \varepsilon^D; h = 1, 2.$$

Since **R1** is satisfied, without loss of generality we can assume

$$Vol(\ddot{\mathcal{C}}_{g^{(k)}}^{(h)}) \geqslant \varepsilon^D\sqrt{\delta}; Vol(\ddot{\mathcal{C}}_{g^{(k)}+1}^{(h)}) \geqslant \varepsilon^D\sqrt{\delta};\ \text{for } h = 1.$$

This implies for sufficiently large n,

$$\sqrt{n}\min\{\|\mathbf{H}_{\mathbf{Z}_1}\|_2, \|\mathbf{H}_{\mathbf{Z}_2}\|_2\} \leqslant \sqrt{\frac{\Big(\min_{1\leqslant l\leqslant g^{(k)}+1}\frac{|\ddot{\mathcal{C}}_l^{(h)}|}{n}\Big)^{-1}}{\xi}} \leqslant \sqrt{\frac{\Big(\varepsilon^D\sqrt{\delta} - \delta^2\Big)^{-1}}{\xi}}. \tag{S9d}$$

Finally, using Lemma S3.2 and boundedness of m and $\lambda_{\max}(\mathbf{Q})$, we have $\frac{1}{n}\mathbf{Y}^\top\mathbf{Q}\mathbf{Y} = O(1)$ on Ω_n. Next, combining Eqs. (S9a) and (S9d), we have under **R1**, on Ω_n

$$\begin{aligned}
&\frac{1}{n}|\mathbf{Y}^\top\mathbf{Q}^{\frac{1}{2}}\left[\dot{\mathbf{P}}_{\mathbf{Z}_1} - \dot{\mathbf{P}}_{\mathbf{Z}_2}\right]\mathbf{Q}^{\frac{\top}{2}}\mathbf{Y}| \\
&\leqslant \left(\frac{1}{n}\mathbf{Y}^\top\mathbf{Q}\mathbf{Y}\right)\left(\sqrt{n}\min\{\|\mathbf{H}_{\mathbf{Z}_1}\|_2, \|\mathbf{H}_{\mathbf{Z}_2}\|_2\}\right)\left(\frac{1}{\sqrt{n}}\|\mathbf{Q}^{\frac{\top}{2}}\mathbf{Z}_1 - \mathbf{Q}^{\frac{\top}{2}}\mathbf{Z}_2\|_2\right) \\
&= O(\delta^{\frac{1}{4}}).
\end{aligned}$$

To bound the second term $\frac{1}{n}|\mathbf{Y}^\top\mathbf{Q}\mathbf{Z}_1^{(0)}\hat{\boldsymbol{\beta}}(\mathbf{Z}_1^{(0)}) - \mathbf{Y}^\top\mathbf{Q}\mathbf{Z}_2^{(0)}\hat{\boldsymbol{\beta}}(\mathbf{Z}_2^{(0)})|$ of (S6), we observe that the leaf nodes corresponding to $\mathbf{Z}_1^{(0)}$ are given by $\ddot{\mathcal{C}}_1^{(h)}, \cdots, \ddot{\mathcal{C}}_{g^{(k)}-1}^{(h)}, \ddot{\mathcal{C}}_{g^{(k)}}^{(h)} \cup \ddot{\mathcal{C}}_{g^{(k)}+1}^{(h)}; h = 1, 2$. From definition of $\bar{\mathfrak{S}}_k^{\varepsilon}(\mathbf{x})$,

$$Vol(\ddot{\mathcal{C}}_l)^{(h)} \geqslant \varepsilon^D \ \forall i = 1, 2, \cdots, g^{(k)} - 1; Vol(\ddot{\mathcal{C}}_{g^{(k)}}^{(h)}) + Vol(\ddot{\mathcal{C}}_{g^{(k)}+1}^{(h)}) \geqslant \varepsilon^D; h = 1, 2.$$

Hence, using similar perturbation bounds as in (S7), (S8), we can conclude that on Ω_n

$$\frac{1}{n}|\mathbf{Y}^\top\mathbf{Q}\mathbf{Z}_1^{(0)}\hat{\boldsymbol{\beta}}(\mathbf{Z}_1^{(0)}) - \mathbf{Y}^\top\mathbf{Q}\mathbf{Z}_2^{(0)}\hat{\boldsymbol{\beta}}(\mathbf{Z}_2^{(0)})| \leqslant C_6\sqrt{\frac{(2^{2D}\delta + \delta^2)}{(\varepsilon^D - \delta^2)}} = O(\delta^{\frac{1}{2}}).$$

Next, combining under **R1**, we have on Ω_n,

$$|v_{n,k,\mathbf{Q}}(\mathbf{x}, \mathbf{s}_k^{(1)}) - v_{n,k,\mathbf{Q}}(\mathbf{x}, \mathbf{s}_k^{(2)})| \leqslant O(\delta^{\frac{1}{4}}) + O(\delta^{\frac{1}{2}}).$$

This quantity goes to 0 uniformly in δ (as $\delta^{\frac{1}{4}}$ is a uniformly continuous function of δ on [0,1]). This completes the proof under **R1**. The proof under scenario **R2** is more technical and is available in the Additional results section (Section S2). □

S1.2 Estimation Error

(C.3.iid) (Analog of ULLN **C.3** for i.i.d. processes): There exists a function class $\mathcal{F}_n \ni m_n(\cdot)$ such that for all arbitrary $L > 0$

(a) Let $\{\epsilon_i^*\}$ denote an i.i.d. process, independent of $\mathcal{D}_n$ and $\dot{\mathcal{D}}_n$, such ϵ_i^* is identically distributed

as ϵ_i. Let $Y_i^* = m(X_i) + \epsilon_i^*; Y_{i,L}^* = T_L Y_i^*$. Then,

$$\lim_{n\to\infty} \mathbb{E}\left[\sup_{f\in T_{\zeta_n}\mathcal{F}_n} |\frac{1}{n}\sum_i |f(X_i) - Y_{i,L}^*|^2 - \mathbb{E}_{\dot{\mathcal{D}}_n}|f(\dot{X}_1) - \dot{Y}_{1,L}|^2|\right] = 0$$

(b) For all $1 \leqslant j \leqslant q$, let $\{(\tilde{X}_i, \ddot{X}_{i-j})\}_{i\geq j+1}$ and $\{(\tilde{\epsilon}_i, \ddot{\epsilon}_{i-j})\}_{i\geq j+1}$ be bivariate i.i.d. processes, independent of the data and of $\dot{G}$, such that $(\tilde{X}_i, \ddot{X}_{i-j})$ is identically distributed as (X_i, X_{i-j}); $(\tilde{\epsilon}_i, \ddot{\epsilon}_{i-j})$ is identically distributed as $(\epsilon_i, \epsilon_{i-j})$ for all i. Define $\tilde{Y}_i = m(\tilde{X}_i) + \tilde{\epsilon}_i$, $\tilde{Y}_{i,L} = T_L\tilde{Y}_i$, $\ddot{Y}_i = m(\ddot{X}_i) + \ddot{\epsilon}_i$, and $\ddot{Y}_{i,L} = T_L\ddot{Y}_i \forall i$. Then the following holds:

$$\lim_{n\to\infty} \mathbb{E}\Bigg[\sup_{f\in T_{\zeta_n}\mathcal{F}_n} \Big|\frac{1}{n}\sum_i (f(\tilde{X}_i) - \tilde{Y}_{i,L})(f(\ddot{X}_{i-j}) - \ddot{Y}_{i-j,L}) - \mathbb{E}_{\dot{\mathcal{D}}_n}(f(\dot{X}_{1+j}) - \dot{Y}_{1+j,L})(f(\dot{X}_1) - \dot{Y}_{1,L})\Big|\Bigg] = 0$$

To prove Condition **C.3.iid**(b) for RF-GLS trees, we will first prove a general result on the concentration rates of cross-product function classes. We recall the definition of random $\mathbb{L}_p$ norm entropy numbers. For a sequence of i.i.d random variable $\{R_i\}_1^n = (R_1, R_2, \cdots, R_n); R_i \in \mathbb{R}^D$, $\varepsilon > 0$, $1 \leqslant p < \infty$, let $\mathcal{W}_n$ be a set of functions $\mathbb{R}^D \mapsto \mathbb{R}$, and for a function $w : \mathbb{R}^d \to \mathbb{R}$, $\|w\|^p_{\{R_i\}_1^n} = \left\{\frac{1}{n}\sum_{i=1}^n |w(R_i)|^p\right\}^{1/p}$. Then, $\mathcal{N}_p(\varepsilon, \mathcal{W}_n, \{R_i\}_1^n)$, the ε-covering number of $\mathcal{W}_n$ w.r.t the random $\mathbb{L}_p$-norm $\|.\|^p_{\{R_i\}_1^n}$ is the minimal $C \in \mathbb{N}$, such that there exists functions $w_1, w_2, \cdots, w_C : \mathbb{R}^D \to \mathbb{R}$ with the property that for every $w \in \mathcal{W}_n$, there is a $j = j(w) \in \{1, 2, \cdots, C\}$ such that $\{\frac{1}{n}\sum_{i=1}^n |w(R_i) - w_j(R_i)|^p\}^{\frac{1}{p}} < \varepsilon$.

Proposition S1.1. (ULLN for cross-product terms for i.i.d. data) Let D_X denote a distribution on $\mathbb{R}^D$ and $D_{\mathbf{Y}} = (D_{Y_1}, D_{Y_2})$ denote a bivariate distribution. Let $\mathbf{X}_i$ and $\mathbf{Y}_i$ denote bivariate i.i.d. processes such that $\{\mathbf{X}_i = (X_{1i}, X_{2i})\}_{i\geq 1} \overset{\text{i.i.d.}}{\sim} D_X \times D_X$ (product measure) and $\{\mathbf{Y}_i = (Y_{1i}, Y_{2i})\}_{i\geq 1} \overset{\text{i.i.d.}}{\sim} D_{\mathbf{Y}}$. Let $Y_{hi,L} = T_L Y_{hi}; h = 1, 2$ for any $L > 0$. Let $(\dot{X}_1, \dot{X}_2) \sim D_X \times D_X$ and $(\dot{Y}_1, \dot{Y}_2) \sim D_{\mathbf{Y}}$, independent of $\{\mathbf{X}_i\}$ and $\{\mathbf{Y}_i\}$. Let $\mathcal{F}_n$ denote some class of real-valued

functions on $\mathbb{R}^D$ class and $\zeta_n \to \infty$. Then for all $\varepsilon > 0$, we have

$$\begin{aligned}\mathbb{P}\Bigg[\sup_{f\in T_{\zeta_n}\mathcal{F}_n}\Big|\frac{1}{n}\sum_i (f(X_{1i})-Y_{1i,L})(f(X_{2i})-Y_{2i,L}) \\ -\mathbb{E}(f(\dot{X}_1)-\dot{Y}_{1,L})(f(\dot{X}_2)-\dot{Y}_{2,L})\Big| > \varepsilon\Bigg] \\ \leqslant 8\mathbb{E}\mathcal{N}_1\left(\frac{\varepsilon}{32\zeta_n}, T_{\zeta_n}\mathcal{F}_n, \{X_i^*\}_{i=1}^{2n}\right)\exp\left(-\frac{n\varepsilon^2}{2048\zeta_n^4}\right).\end{aligned}$$

where $X^*_{2i-1} = X_{1i}$ and $X^*_{2i} = X_{2i}$ for $i = 1, \ldots, n$.

Note that the class $\mathcal{F}_n$ is dependent on the data $\{\mathbf{X}_i | 1 \leq i \leq n\}$ and hence generally we cannot say $\mathbb{E}f(X_{1i}) = \mathbb{E}f(\dot{X}_1)$ for the new sample $\dot{X}_1$. So, the cross-term is not a sample covariance and cannot be bounded by direct application of Cauchy-Schwartz inequality on the corresponding ULLN for the squared terms.

Proof of Proposition S1.1. For convenience, we denote $H_i = (X_{1i}, X_{2i}, \epsilon_{1i}, \epsilon_{2i})$ and $\dot{H} = (\dot{X}_1, \dot{X}_2, \dot{\epsilon}_1, \dot{\epsilon}_2)$, and $\mathcal{W}_n = \{w \,|\, w(x_1, x_2, y_1, y_2) = (f(x_1) - y_1)(f(x_2) - y_2), f \in T_{\zeta_n}\mathcal{F}_n\}$. For large enough n, $\zeta_n > L$, and $|w| \leq 4\zeta_n^2$ for all $w \in \mathcal{W}_n$. By Theorem 9.1 of Györfi et al. (2006),

$$\mathbb{P}\Big[\sup_{w\in\mathcal{W}_n}\Big|\frac{1}{n}\sum_i w(H_i) - \mathbb{E}w(\dot{H})\Big| > \varepsilon\Big] \leq 8\mathcal{N}_1\left(\frac{\varepsilon}{8}, \mathcal{H}_n, \{W_i\}_{i=1}^n\right)\exp\left(-\frac{n\varepsilon^2}{128(4\zeta_n^2)^2}\right).$$

Let $w_j(H_i) = (f_j(X_{1i}) - Y_{1i,L})(f_j(X_{2i}) - Y_{2i,L})$, for some $f_j \in T_{\zeta_n}\mathcal{F}_n$. Then for large enough n,

$\zeta_n > L$ and we have,

$$
\begin{aligned}
&\frac{1}{n}\sum_i |w_j(H_i) - w_{j'}(H_i)| \\
&= \frac{1}{n}\sum_i |(f_j(X_{1i}) - Y_{1i,L})(f_j(X_{2i}) - Y_{2i,L}) \\
&\quad - (f_{j'}(X_{1i}) - Y_{1i,L})(f_{j'}(X_{2i}) - Y_{2i,L})| \\
&= \frac{1}{n}\sum_i \left| \left[f_j(X_{1i})f_j(X_{2i}) - f_{j'}(X_{1i})f_{j'}(X_{2i})\right] \right| \\
&+ \frac{1}{n}\sum_i \left| \left[Y_{1i,L}(f_{j'}(X_{2i}) - f_j(X_{2i})) + Y_{2i,L}(f_{j'}(X_{1i}) - f_j(X_{1i}))\right] \right| \\
&= \frac{1}{n}\sum_i \left| \left[f_j(X_{1i})(f_j(X_{2i}) - f_{j'}(X_{2i})) + f_{j'}(X_{2i})(f_j(X_{1i}) - f_{j'}(X_{1i})\right] \right| \\
&+ \frac{1}{n}\sum_i \left| \left[Y_{1i,L}(f_{j'}(X_{2i}) - f_j(X_{2i})) + Y_{2i,L}(f_{j'}(X_{1i}) - f_j(X_{1i}))\right] \right| \\
&\leqslant \frac{1}{n}\sum_i \left| (|f_j(X_{1i})| + |Y_{1i,L}|)\Big|f_j(X_{2i}) - f_{j'}(X_{2i})\Big| \right| \\
&+ \frac{1}{n}\sum_i \left| (|f_{j'}(X_{2i})| + |Y_{2i,L}|)\Big|f_j(X_{1i}) - f_{j'}(X_{1i})\Big| \right| \\
&\leqslant 2\zeta_n \frac{1}{n}\sum_i |f_j(X_{2i}) - f_{j'}(X_{2i})| + 2\zeta_n \frac{1}{n}\sum_i |f_j(X_{1i}) - f_{j'}(X_{1i})| \\
&\leqslant 4\zeta_n \frac{1}{2n}\sum_{i=1}^{2n} |f_j(X_i^*) - f_{j'}(X_i^*)|.
\end{aligned}
$$

Hence

$$
\mathcal{N}_1\left(\frac{\varepsilon}{8}, \mathcal{W}_n, \{H_i\}_{i=1}^n\right) \leqslant \mathcal{N}_1\left(\frac{\varepsilon}{32\zeta_n}, T_{\zeta_n}\mathcal{F}_n, \{X_i^*\}_{i=1}^{2n}\right).
$$

□

Proof of Proposition A2.3. We consider the following two cases:

Case 1: Functions in $\mathcal{G}_n$ are uniformly bounded by $C_6 \in \mathbb{R}, C_6 < \infty$.

In this scenario, due to the fact that the constant function C_6 works as a bounded envelope for $\mathcal{G}_n, \forall n$, the result of Nobel and Dembo (1993) ensures almost sure convergence for $\{U_i\}$ even under the assumption of $\mathbb{L}_1$ convergence of the independent counterpart $\{U_i\}$ [1] Since $\mathcal{G}_n$ is bounded,

[1] The statement of Theorem 1 of Nobel and Dembo (1993) assumes almost strong (almost sure) ULLN for the

almost sure convergence is enough to ensure convergence in $\mathbb{L}_1$. This completes the proof for this case.

Case 2: No uniformly bounded envelope. For every $g \in \mathcal{G}_n$, we define $g_1 = g(\mathbb{I}(G_n \leqslant C_6))$ and $g_2 = g(\mathbb{I}(G_n > C_6))$. So we have

$$\sup_{g\in\mathcal{G}_n}\left|\frac{1}{n}\sum_i g(U_i^{(C)})\right| = \sup_{g_1,\, g\in\mathcal{G}_n}\left|\frac{1}{n}\sum_i (g_1(U_i) - \mathbb{E}g_1(U_i))\right| + \sup_{g_2,\, g\in\mathcal{G}_n}\left|\frac{1}{n}\sum_i (g_2(U_i) - \mathbb{E}g_2(U_i))\right|.$$

The first term converges to 0 in $\mathbb{L}_1$, using Case 1. For the second term, we have

$$\mathbb{E}\sup_{g_2,\, g\in\mathcal{G}_n}\left|\frac{1}{n}\sum_i (g_2(U_i) - \mathbb{E}g_2(U_i))\right| \leqslant \frac{2}{n}\sum_i \mathbb{E}|G_n(U_i)|\mathbb{I}(|G_n(U_i)| > C_6).$$

From (A6), this goes to zero as $C_6 \to \infty$. □

Proposition S1.2. (ULLN for squared and cross-product classes for dependent data) If $\mathcal{F}_n$ denote a class of functions satisfying **(C3.iid)** for i.i.d. processes, and Condition **C.4**, then $\mathcal{F}_n$ satisfies Condition **(C3)** for a β-mixing process $\{\epsilon_i\}$ under Assumption 2.

Condition **(C.4)** on uniformly bounded $(2+\delta)^{th}$ moment for the class $\mathcal{F}_n$ will be sufficient to ensure the moment-bound of (A6) for squared and cross-product function classes allowing use of the ULLN in Proposition A2.3 to prove Proposition S1.2.

Proof of Proposition S1.2. To establish the ULLN in condition **(C.3)** for the dependent error process, it is enough to establish the following two conditions:

(a)
$$\lim_{n\to\infty}\mathbb{E}\left[\sup_{f\in T_{\zeta_n}\mathcal{F}_n}\left|\frac{1}{n}\sum_i |f(X_i) - Y_{i,L}|^2 - \mathbb{E}|f(\dot{X}_1) - \dot{Y}_{1,L}|^2\right|\right] = 0.$$

i.i.d. process, but the proof only requires weak (convergence in probability) ULLN. As we assume an $\mathbb{L}_1$ ULLN for the i.i.d. process, by Markov inequality weak ULLN holds here as well, and the same result is obtained.

(b) For all $1 \le j \le q$,

$$\lim_{n\to\infty} \mathbb{E}\Bigg[\sup_{f\in T_{\zeta_n}\mathcal{F}_n} \Big| \frac{1}{n}\sum_i (f(X_i) - Y_{i,L})(f(X_{i-j}) - Y_{i-j,L}) \\ - \mathbb{E}(f(\dot{X}_{1+j}) - \dot{Y}_{1+j,L})(f(\dot{X}_1) - \dot{Y}_{1,L}) \Big| \Bigg] = 0.$$

Since $\{X_i\}$ is an i.i.d. process independent of the β-mixing process $\{\epsilon_i\}$, using the property (S20) of β-mixing coefficients presented in Lemma S3.1, we have $U_i = (X_i, \epsilon_i)$ is also a β-mixing process. Let $\mathcal{G}_n$ be the class of functions $g : \mathbb{R}^{D+1} \to \mathbb{R}$ such that $g(U_i) = (Y_{i,L} - f(X_i))^2$ for some $f \in T_{\zeta_n}\mathcal{F}_n$. Clearly, under condition **(C.4)**, $G_n = 2L^2 + F_n^2$ is an envelope for $\mathcal{G}_n$. Choosing $C > 2L^2 + 1$ we have

$$\frac{1}{n}\sum_i \mathbb{E}|G_n(Y_i, X_i)|\mathbb{I}(|G_n(Y_i, X_i)| > C) = \frac{1}{n}\sum_i \mathbb{E}2(L^2 + F_n^2(X_i))\mathbb{I}(|F_n(X_i)| > \sqrt{C - 2L^2}) \\ \le \frac{2(L^2+1)}{(C-2L^2)^{\frac{\delta}{2}}} \frac{1}{n}\sum_i \mathbb{E}|F_n X_i)|^{2+\delta}$$

Here, the inequality uses the fact that for any random variable X, $\mathbb{E}|X|\mathbb{I}(|X| > C) \le \mathbb{E}|X|^{1+\delta}/C^{\delta}$. Clearly, the quantity above goes to 0 by Condition **(C.4)** by first taking $n \to \infty$ and then $C \to \infty$. Hence (A6) is satisfied by the class $\mathcal{G}_n$, and ULLN **(C.3)**(a) is established using Proposition A2.3.

Next, define $H_i = (X_i, X_{i-j}, \epsilon_i, \epsilon_{i-j})$. Since X_i's are i.i.d., (X_i, X_{i-j}) is a bivariate m-dependent process with lag at most j. Also, since $\{\epsilon_i\}$ is a β-mixing process, so is $(\epsilon_i, \epsilon_{i-j})$ with mixing coefficient at lag a given by $\beta_{(\epsilon_i,\epsilon_{i-j})}(a) \le \beta_{\epsilon_i}(a-j)$. Once again, since $\{(X_i, X_{i-j})\} \perp\!\!\!\perp \{(\epsilon_i, \epsilon_{i-j})\}$ and both are β-mixing, using (S20) established in the proof of Lemma S3.1, $\{H_i\}$ is a β-mixing process. Define $\mathcal{G}_n^{(j)}$ to be the class of functions $g^{(j)}$ of the form $g^{(j)}(H_i) = (Y_{i,L} - f(X_i))(Y_{i-j,L} - f(X_{i-j}))$ for $f \in T_{\zeta_n}\mathcal{F}_n$. Like $\mathcal{G}_n$, $\mathcal{G}_n^{(j)}$ admits an envelope $G_n^{(j)} \le 2L^2 + F_n^2$ which satisfies the mean uniform integrability condition (A6). Hence, using condition **(C.3.iid)**(b) and **(C.4)**, and applying Proposition A2.3, we have **(C.3)**(b). □

Proof of Proposition A2.4. For a GLS-style regression tree $m_n(., \Theta)$ in RF-GLS, built with data $\mathcal{D}_n$ and randomness Θ, let the partition obtained from $\mathcal{D}_n$ and $\boldsymbol{\Theta}$ be denoted by $\mathcal{P}_{GLS,n}(\Theta) = \mathcal{P}_n(\Theta)$. Then $\mathcal{F}_n = \mathcal{F}_n(\Theta)$ is the set of all functions $f : [0,1]^D \to \mathbb{R}$ piece-wise constant on each cell of the

partition $\mathcal{P}_n(\Theta)$.

We will show that $T_{\zeta_n}\mathcal{F}_n$ satisfies **(C.3.iid)**(a) and (b) which would imply that the smaller class $T_{\zeta_n}\tilde{\mathcal{F}}_n$ also satisfy these. As the class $\tilde{F}_n$ satisfies **C.4**, the result is then proved by Proposition S1.2.

Condition **(C.3.iid)**(a) is proved in Scornet et al. (2015) (page 1731) (and more generally in Theorem 13.1 of Györfi et al. (2006)) for the class $T_{\zeta_n}\dot{\mathcal{F}}_n$, the set of all functions $f : [0,1]^D \to \mathbb{R}$ piece-wise constant on each cell of the partition $\mathcal{P}_{OLS,n}(\Theta)$ generated by a RF-tree. The result only relies on the number of RF trees t_n, the number of samples n and the number of features D, and hence the proof holds also for $T_{\zeta_n}F_n$ for RF-GLS using same number of trees.

To prove the cross-product loss estimation error in Condition **(C.3.iid)**(b), we let $X_{1i} = \tilde{X}_i$, $X_{2i} = \ddot{X}_{i-j}$, $(Y_{1i}, Y_{2i}) = (m(X_{1i}), m(X_{2i})) + (\tilde{\epsilon}_i, \ddot{\epsilon}_{i-j})$ and use Proposition S1.1. The bounding tail probability for the cross-product term in that proposition is almost identical to that for the squared error term used in Scornet et al. (2015). Only difference is that that the entropy number $\mathcal{N}_1\left(\frac{\varepsilon}{32\zeta_n}, T_{\zeta_n}\mathcal{F}_n, \{X_i^*\}_{i=1}^{2n}\right)$ is based on the empirical measure on the samples $\{X_i^*\}$ of size $2n$ as opposed to n samples for the square term. However, by Theorem 9.4 of Györfi et al. (2006), the bound on the entropy number $\mathcal{N}_1\left(\frac{\varepsilon}{32\zeta_n}, T_{\zeta_n}\mathcal{F}_n, \nu\right)$ is free of the choice of the measure ν and only depends on the choice of the class $\mathcal{F}_n$. Hence, like the squared error ULLN, Condition **(C.3.iid)**(b) holds as long as the scaling of t_n and tail moments in Assumption 4(a) are satisfied with ζ_n. □

S1.3 Consistency of data-driven-partitioning-based GLS estimates under dependent errors

Proof of Theorem A2.1. We first show that to prove $\mathbb{L}_2$ consistency of m_n, it is enough to show that $\mathbb{E}\Big[\mathbb{E}_{\dot{G}}[\boldsymbol{\rho}^\top(m_n(\dot{X}^{(q+1)}) - m(\dot{X}^{(q+1)}))]^2\Big] \to 0$, where $\dot{X}^{(q+1)} = (\dot{X}_{q+1}, \dot{X}_q, \cdots, \dot{X}_1)$, $\dot{\epsilon}^{(q+1)} = (\dot{\epsilon}_{q+1}, \dot{\epsilon}_q, \cdots, \dot{\epsilon}_1)$, $\dot{G} = (\dot{X}^{(q+1)}, \dot{\epsilon}^{(q+1)})$ and for any function $f : \mathbb{R}^d \to \mathbb{R}$, $f(\dot{X}^{(q+1)}) = (f(\dot{X}_{q+1}), \ldots, f(\dot{X}_1))^\top$. Note that from Assumption 2, for any $q < i \leqslant n-q$, $\mathbf{Q}_{ii} = \alpha$, $\mathbf{Q}_{ij} = 0$ for $|j-i| > q$ and $\mathbf{Q}_{ij} = \sum_{j'=|j-i|}^{q} \rho_{j'}\rho_{j'-|i-j|}$ for $|i-j| \leq q$. Hence, by Assumption 3,

$$\alpha - 2\sum_{j=1}^{q} |\sum_{j'=j}^{q} \rho_{j'}\rho_{j'-j}| > 0. \tag{S10}$$

Since $\dot{X}_i$'s are i.i.d., using (S10) and Jensen's inequality,

$$\begin{aligned}\mathbb{E}_X\left(\boldsymbol{\rho}^\top f(\dot{X}^{(q+1)})\right)^2 =& \alpha\mathbb{E}f(\dot{X}_1)^2 + 2(\mathbb{E}f(\dot{X}_1))^2\sum_{j=1}^{q}\sum_{j'=j}^{q}\rho_{j'}\rho_{j'-j} \\ \geqslant& \mathbb{E}f(\dot{X}_1)^2(\alpha - 2\sum_{j=1}^{q}|\sum_{j'=j}^{q}\rho_{j'}\rho_{j'-j}|)\end{aligned} \tag{S11}$$

Choosing $f = m_n - m$ proves the result showing that it is enough to work with $\mathbb{E}\Big[\mathbb{E}[\boldsymbol{\rho}^\top(m_n(\dot{X}^{(q+1)}) - m(\dot{X}^{(q+1)}))]^2\Big]$.

The next part rest of the proof showing consistency of the truncated estimator $T_{\zeta_n}m_n$ largely emulates the technique from Györfi et al. (2006) (Theorem 10.2). Throughout, careful adjustments need to be made to account for use of the quadratic form $\mathbf{Q}$. The result is summarized in Lemma S1.1.

Lemma S1.1. Under the conditions of Theorem A2.1, we have

$$\lim_{n\to\infty}\mathbb{E}\left[\mathbb{E}_X\left[\boldsymbol{\rho}^\top\left(T_{\zeta_n}m_n(\dot{X}^{(q+1)},\Theta) - m(\dot{X}^{(q+1)})\right)\right]^2\right] = 0.$$

Proof of Lemma S1.1. Let $\tilde{m}_n = T_{\zeta_n}m_n$.

$$\begin{aligned}&\mathbb{E}_{\dot{G}}\left[\boldsymbol{\rho}^\top\left(\tilde{m}_n(\dot{X}^{(q+1)}) - \dot{Y}^{(q+1)}\right)\right]^2 \\ &= \mathbb{E}_{\dot{G}}\left[\boldsymbol{\rho}^\top\left(\tilde{m}_n(\dot{X}^{(q+1)}) - m(\dot{X}^{(q+1)})\right) + \boldsymbol{\rho}^\top\left(m(\dot{X}^{(q+1)}) - \dot{Y}^{(q+1)}\right)\right]^2 \\ &= \mathbb{E}_{\dot{G}}\left[\boldsymbol{\rho}^\top\left(\tilde{m}_n(\dot{X}^{(q+1)}) - m(\dot{X}^{(q+1)})\right)\right]^2 + \mathbb{E}_{\dot{G}}\left[\boldsymbol{\rho}^\top\left(m(\dot{X}^{(q+1)}) - \dot{Y}^{(q+1)}\right)\right]^2 \\ &+ 2\mathbb{E}_{\dot{G}}\left[\boldsymbol{\rho}^\top\left(\tilde{m}_n(\dot{X}^{(q+1)}) - m(\dot{X}^{(q+1)})\right)\boldsymbol{\rho}^\top\left(m(\dot{X}^{(q+1)}) - \dot{Y}^{(q+1)}\right)\right]\end{aligned}$$

Now,

$$\begin{aligned}&\mathbb{E}_{\dot{G}}\left[\boldsymbol{\rho}^\top\left(\tilde{m}_n(\dot{X}^{(q+1)}) - m(\dot{X}^{(q+1)})\right)\boldsymbol{\rho}^\top\left(m(\dot{X}^{(q+1)}) - \dot{Y}^{(q+1)}\right)\right] \\ &= \mathbb{E}_{\dot{X}^{(q+1)}}\left[\boldsymbol{\rho}^\top\left(\tilde{m}_n(\dot{X}^{(q+1)}) - m(\dot{X}^{(q+1)})\right)\boldsymbol{\rho}^\top\mathbb{E}_{\dot{Y}^{(q+1)}}\left[\left(m(\dot{X}^{(q+1)}) - \dot{Y}^{(q+1)}\right)\right]\right] \\ &= 0\end{aligned}$$

Hence, we have,

$$
\begin{aligned}
&\mathbb{E}_{\dot{G}}\left[\boldsymbol{\rho}^\top\left(\tilde{m}_n(\dot{X}^{(q+1)}) - m(\dot{X}^{(q+1)})\right)\right]^2 \\
&= \mathbb{E}_{\dot{G}}\left[\boldsymbol{\rho}^\top\left(\tilde{m}_n(\dot{X}^{(q+1)}) - \dot{Y}^{(q+1)}\right)\right]^2 - \mathbb{E}_{\dot{G}}\left[\boldsymbol{\rho}^\top\left(m(\dot{X}^{(q+1)}) - \dot{Y}^{(q+1)}\right)\right]^2 \\
&= A\left(A + 2\left(\mathbb{E}_{\dot{G}}\left[\boldsymbol{\rho}^\top\left(m(\dot{X}^{(q+1)} - \dot{Y}^{(q+1)}\right)\right]^2\right)^{\frac{1}{2}}\right)
\end{aligned}
$$

where

$$
A := \left(\mathbb{E}\left[\boldsymbol{\rho}^\top\left(\tilde{m}_n(\dot{X}^{(q+1)}) - \dot{Y}^{(q+1)}\right)\right]^2\right)^{\frac{1}{2}} - \left(\mathbb{E}\left[\boldsymbol{\rho}^\top\left(m(\dot{X}^{(q+1)}) - \dot{Y}^{(q+1)}\right)\right]^2\right)^{\frac{1}{2}}
$$

As the term $\mathbb{E}_{\dot{G}}\left[\boldsymbol{\rho}^\top\left(m(\dot{X}^{(q+1)} - \dot{Y}^{(q+1)}\right)\right]^2$ is non-random and $O(1)$ ($\boldsymbol{\rho}$ being of fixed-dimension), using Cauchy-Schwartz inequality, it is enough to show $\mathbb{E}A^2 \to 0$. Applying $(a+b)^2 \leq 2(a^2+b^2)$, we have

$$
\begin{aligned}
\mathbb{E}A^2 \leqslant 2\mathbb{E}\Bigg[&\left(\mathbb{E}_{\dot{G}}\left[\boldsymbol{\rho}^\top\left(\tilde{m}_n(\dot{X}^{(q+1)}) - \dot{Y}^{(q+1)}\right)\right]^2\right)^{\frac{1}{2}} \\
&- \inf_{f\in T_{\zeta_n}\mathcal{F}_n}\left(\mathbb{E}_{\dot{G}}\left[\boldsymbol{\rho}^\top\left(f(\dot{X}^{(q+1)}) - \dot{Y}^{(q+1)}\right)\right]^2\right)^{\frac{1}{2}}\Bigg]^2 \\
+ 2\mathbb{E}\Bigg[&\inf_{f\in T_{\zeta_n}\mathcal{F}_n}\left(\mathbb{E}_{\dot{G}}\left[\boldsymbol{\rho}^\top\left(f(\dot{X}^{(q+1)}) - \dot{Y}^{(q+1)}\right)\right]^2\right)^{\frac{1}{2}} \\
&- \left(\mathbb{E}_{\dot{G}}\left[\boldsymbol{\rho}^\top\left(m(\dot{X}^{(q+1)}) - \dot{Y}^{(q+1)}\right)\right]^2\right)^{\frac{1}{2}}\Bigg]^2
\end{aligned}
$$

Using triangular inequality for the second quantity, we have

$$
\begin{aligned}
&\mathbb{E}\Bigg[\inf_{f\in T_{\zeta_n}\mathcal{F}_n}\left(\mathbb{E}_{\dot{G}}\left[\boldsymbol{\rho}^\top\left(f(\dot{X}^{(q+1)}) - \dot{Y}^{(q+1)}\right)\right]^2\right)^{\frac{1}{2}} \\
&\qquad - \left(\mathbb{E}_{\dot{G}}\left[\boldsymbol{\rho}^\top\left(m(\dot{X}^{(q+1)}) - \dot{Y}^{(q+1)}\right)\right]^2\right)^{\frac{1}{2}}\Bigg]^2 \\
&\leqslant \mathbb{E}\inf_{f\in T_{\zeta_n}\mathcal{F}_n}\left(\mathbb{E}_{\dot{G}}\left[\boldsymbol{\rho}^\top\left(f(\dot{X}^{(q+1)}) - m(\dot{X}^{(q+1)})\right)\right]^2\right) \\
&\leqslant \left(\alpha + \sum_{j\neq j'}|\rho_j\rho_{j'}|\right)\mathbb{E}\left[\inf_{f\in T_{\zeta_n}\mathcal{F}_n}\mathbb{E}_{\dot{X}_1}|f(\dot{X}_1) - m(\dot{X}_1)|^2\right]
\end{aligned}
$$

This vanishes asymptotically, due to the approximation error condition **(C.2)**. Hence, we focus on

the other term $\mathbb{E}A_1^2$ in the expression of $\mathbb{E}A^2$, where

$$A_1 := \left(\mathbb{E}_{\dot{G}}\left[\boldsymbol{\rho}^\top\left(\tilde{m}_n(\dot{X}^{(q+1)}) - \dot{Y}^{(q+1)}\right)\right]^2\right)^{\frac{1}{2}} - \inf_{f\in T_{\zeta_n}\mathcal{F}_n}\left(\mathbb{E}_{\dot{G}}\left[\boldsymbol{\rho}^\top\left(f(\dot{X}^{(q+1)}) - \dot{Y}^{(q+1)}\right)\right]^2\right)^{\frac{1}{2}}.$$

$A1$ can be decomposed and bounded as follows by sum of ten terms:

$$\begin{aligned}
A_1 \leqslant \sup_{f\in T_{\zeta_n}\mathcal{F}_n}\Bigg\{ & \left(\mathbb{E}_{\dot{G}}\left[\boldsymbol{\rho}^\top\left(\tilde{m}_n(\dot{X}^{(q+1)}) - \dot{Y}^{(q+1)}\right)\right]^2\right)^{\frac{1}{2}} - \left(\mathbb{E}_{\dot{G}}\left[\boldsymbol{\rho}^\top\left(\tilde{m}_n(\dot{X}^{(q+1)}) - \dot{Y}_L^{(q+1)}\right)\right]^2\right)^{\frac{1}{2}} \\
& + \left(\mathbb{E}_{\dot{G}}\left[\boldsymbol{\rho}^\top\left(\tilde{m}_n(\dot{X}^{(q+1)}) - \dot{Y}_L^{(q+1)}\right)\right]^2\right)^{\frac{1}{2}} - \left(\frac{1}{n}\sum_i\left[\boldsymbol{\rho}^\top(\tilde{m}_n(\mathbf{X}^{(i)}) - \mathbf{Y}_L^{(i)})\right]^2\right)^{\frac{1}{2}} \\
& + \left(\frac{1}{n}\sum_i\left[\boldsymbol{\rho}^\top(\tilde{m}_n(\mathbf{X}^{(i)}) - \mathbf{Y}_L^{(i)})\right]^2\right)^{\frac{1}{2}} - \left(\frac{1}{n}\sum_i\left[\boldsymbol{\rho}^\top(m_n(\mathbf{X}^{(i)}) - \mathbf{Y}_L^{(i)})\right]^2\right)^{\frac{1}{2}} \\
& + \left(\frac{1}{n}\sum_i\left[\boldsymbol{\rho}^\top(m_n(\mathbf{X}^{(i)}) - \mathbf{Y}_L^{(i)})\right]^2\right)^{\frac{1}{2}} - \left(\frac{1}{n}\sum_i\left[\boldsymbol{\rho}^\top(m_n(\mathbf{X}^{(i)}) - \mathbf{Y}^{(i)})\right]^2\right)^{\frac{1}{2}} \\
& + \left(\frac{1}{n}\sum_i\left[\boldsymbol{\rho}^\top(m_n(\mathbf{X}^{(i)}) - \mathbf{Y}^{(i)})\right]^2\right)^{\frac{1}{2}} - \left(\frac{1}{n}\left(m_n(\mathbf{X}) - \mathbf{Y}\right)^\top\mathbf{Q}\left(m_n(\mathbf{X}) - \mathbf{Y}\right)\right)^{\frac{1}{2}} \\
& + \left(\frac{1}{n}\left(m_n(\mathbf{X}) - \mathbf{Y}\right)^\top\mathbf{Q}\left(m_n(\mathbf{X}) - \mathbf{Y}\right)\right)^{\frac{1}{2}} - \left(\frac{1}{n}\left(f(\mathbf{X}) - \mathbf{Y}\right)^\top\mathbf{Q}\left(f(\mathbf{X}) - \mathbf{Y}\right)\right)^{\frac{1}{2}} \\
& + \left(\frac{1}{n}\left(f(\mathbf{X}) - \mathbf{Y}\right)^\top\mathbf{Q}\left(f(\mathbf{X}) - \mathbf{Y}\right)\right)^{\frac{1}{2}} - \left(\frac{1}{n}\sum_i\left[\boldsymbol{\rho}^\top(f(\mathbf{X}^{(i)}) - \mathbf{Y}^{(i)})\right]^2\right)^{\frac{1}{2}} \\
& + \left(\frac{1}{n}\sum_i\left[\boldsymbol{\rho}^\top(f(\mathbf{X}^{(i)}) - \mathbf{Y}^{(i)})\right]^2\right)^{\frac{1}{2}} - \left(\frac{1}{n}\sum_i\left[\boldsymbol{\rho}^\top(f(\mathbf{X}^{(i)}) - \mathbf{Y}_L^{(i)})\right]^2\right)^{\frac{1}{2}} \\
& + \left(\frac{1}{n}\sum_i\left[\boldsymbol{\rho}^\top(f(\mathbf{X}^{(i)}) - \mathbf{Y}_L^{(i)})\right]^2\right)^{\frac{1}{2}} - \left(\mathbb{E}_{\dot{G}}\left[\boldsymbol{\rho}^\top\left(f(\dot{X}^{(q+1)}) - \dot{Y}_L^{(q+1)}\right)\right]^2\right)^{\frac{1}{2}} \\
& + \left(\mathbb{E}_{\dot{G}}\left[\boldsymbol{\rho}^\top\left(f(\dot{X}^{(q+1)}) - \dot{Y}_L^{(q+1)}\right)\right]^2\right)^{\frac{1}{2}} - \left(\mathbb{E}_{\dot{G}}\left[\boldsymbol{\rho}^\top\left(f(\dot{X}^{(q+1)}) - \dot{Y}^{(q+1)}\right)\right]^2\right)^{\frac{1}{2}} \Bigg\}
\end{aligned}$$

Here $\mathbf{X}^{(i)} = (X_i, X_{i-1}, \ldots, X_{i-q})^\top$. Let the 10 terms in the above inequality be denoted by $b_1, \ldots, b_{10}$. The 6^{th} term is negative by definition (Equation A5). Hence, $A_1 \leq \sum_{t\in\{1,\ldots,10\}\setminus\{6\}} b_t$.

On the other hand,

$$\begin{aligned} A_1 &\geqslant \left(\mathbb{E}_{\dot{G}}\left[\boldsymbol{\rho}^\top\left(m(\dot{X}^{(q+1)}) - \dot{Y}^{(q+1)}\right)\right]^2\right)^{\frac{1}{2}} \\ &\qquad - \inf_{f\in T_{\zeta_n}\mathcal{F}_n}\left(\mathbb{E}_{\dot{G}}\left[\boldsymbol{\rho}^\top\left(f(\dot{X}^{(q+1)}) - \dot{Y}^{(q+1)}\right)\right]^2\right)^{\frac{1}{2}} \\ &\geqslant -\inf_{f\in T_{\zeta_n}\mathcal{F}_n}\left(\mathbb{E}_{\dot{G}}\left[\boldsymbol{\rho}^\top\left(f(\dot{X}^{(q+1)}) - m(\dot{X}^{(q+1)})\right)\right]^2\right)^{\frac{1}{2}} \end{aligned}$$

Denoting the right-hand side of the above equation by a, and using $(\sum_{t\in\{1,\ldots,10\}\setminus\{6\}} b_t)^2 \leqslant 9\sum_{t\in\{1,\ldots,10\}\setminus\{6\}} b_t^2$, we have $A_1^2 \leq a^2 + 9\sum_{t\in\{1,\ldots,10\}\setminus\{6\}} b_t^2$. Hence, to show $\mathbb{E}(A_1^2)$ vanishes it is enough to show the terms $\mathbb{E}(a^2)$ and $\mathbb{E}(b_t^2)$ vanishes.

The term $\mathbb{E}(a^2)$ directly goes to 0 using the approximation error condition **(C.2)**.

Using triangular inequality, the 1^{st} and 10^{th} $\mathbb{E}(b_t^2)$ terms are bounded above by

$$(q+1)\alpha\left(\mathbb{E}_{\dot{G}}\left[\dot{Y} - \dot{Y}_L\right]^2\right)$$

Similarly, the 4^{th} and 8^{th} term is bounded above by the following:

$$(q+1)\alpha\mathbb{E}\left(\frac{1}{n}\sum_i [Y_i - Y_{i,L}]^2\right).$$

The 3^{rd} term is bounded above by the following which vanishes by Condition **(C.1)**.

$$(q+1)\alpha\mathbb{E}\left(\frac{1}{n}\sum_i (m_n(X_i) - \tilde{m}_n(X_i))^2.\right) \leq (q+1)\alpha\mathbb{E}\max_i \left(m_n(X_i) - \tilde{m}_n(X_i)\right)^2.$$

The 5^{th} and 7^{th} $\mathbb{E}(b_t^2)$ terms only consists of the q^2 residual terms arising from the first q rows of the Cholesky factorization in Assumption 2. Hence, they are bounded by:

$$\begin{aligned} &\left(\sum_{1\leq i,j\leq q} |(\mathbf{L}^\top\mathbf{L})_{ij}|\right)\frac{1}{n}\mathbb{E}\max_{1\leqslant i\leq q}\sup_{f\in\{m_n\}\cup T_{\zeta_n}\mathcal{F}_n}(f(X_i) - Y_i)^2 \\ \leq\ & 4\left(\sum_{1\leq i,j\leq q} |(\mathbf{L}^\top\mathbf{L})_{ij}|\right)\frac{1}{n}\mathbb{E}\left(\zeta_n^2 + \|m\|_\infty^2 + \max_{1\leqslant i\leq q}\left[m_n(X_i)^2\mathbb{I}(|m_n(X_i)\geqslant\zeta_n) + \epsilon_i^2\right]\right) \end{aligned}$$

Using $\zeta_n^2/n \to 0$, boundedness of m, and Condition **(C.1)**, this goes to zero.

The 2^{nd} and 9^{th} term are bounded by the following:

$$\begin{aligned}
&\mathbb{E}\left[\sup_{f\in T_{\zeta_n}\mathcal{F}_n}\left|\frac{1}{n}\sum_i\left[\boldsymbol{\rho}^\top\left(f(X^{(i)})-Y^{(i)}\right)\right]^2-\mathbb{E}_{\dot{G}}\left[\boldsymbol{\rho}^\top\left(f(\dot{X}^{(q+1)})-\dot{Y}_L^{(q+1)}\right)\right]^2\right|\right.\\
&\leqslant\mathbb{E}\left[\sup_{f\in T_{\zeta_n}\mathcal{F}_n}\left|\alpha\left(\frac{1}{n}\sum_i(f(X_i)-Y_{i,L})^2-\mathbb{E}_{\dot{G}}(f(\dot{X}_1)-\dot{Y}_{1,L})^2\right)\right.\right.\\
&+2\sum_{j=1}^{q}\sum_{j'\neq j}^{q}\rho_{j'}\rho_{j'-j}\left(\frac{1}{n}\sum_i(f(X_i)-Y_{i,L})(f(X_{i-j})-Y_{i-j,L})\right.\\
&\left.\left.\left.-\mathbb{E}_{\dot{G}}(f(\dot{X}_i)-\dot{Y}_{i,L})(f(\dot{X}_{i-j})-\dot{Y}_{i-j,L})\right)\right|\right]
\end{aligned}$$

Direct application of Assumption 2 (Equation A1) and the ULLN **C.3** sends this to zero.

Combining all of this, as their are 9 b_t's, we have

$$\lim_{n\to\infty}\mathbb{E}A_1^2\leq 18(q+1)\alpha\left(\mathbb{E}_{\dot{G}}\left[\dot{Y}-\dot{Y}_L\right]^2\right)+18(q+1)\alpha\lim_{n\to\infty}\mathbb{E}\left(\frac{1}{n}\sum_i\left[Y_i-Y_{i,L}\right]^2\right).$$

The first term above goes to 0 as $L\to\infty$. As ϵ_i is β-mixing, it can be proved similar to Lemma S3.1 part 2, that $\frac{1}{n}\sum_i\mathbb{E}(\epsilon_i^2 I(|\epsilon_i|>L))\to\mathbb{E}(\epsilon_1^2 I(|\epsilon_1|>L))$ a.s. Hence, for $L>M_0+1$, that the limit in the second term is bounded by $(M_0^2+1)\mathbb{E}\epsilon_1^2\mathbb{I}(|\epsilon_1|>L-M_0)$ which also goes to 0 as $L\to\infty$ due to the finite $(2+\delta)^{th}$ moment from Assumption 1. □

Returning to the proof of Theorem A2.1, Lemma S1.1 and (S11) implies that $\lim_{n\to\infty}\mathbb{E}[\mathbb{E}_{\dot{X}_1}(T_{\zeta_n}m_n(\dot{X}_1)-m(\dot{X}_1))^2)=0$. The RF-GLS estimate $m_n(\mathbf{x})$ for any $\mathbf{x}$ can only take one of the possible leaf node values. Hence $|m_n(x)|\leq\max_i|m_n(X_i)|$.

$$\begin{aligned}
\lim_{n\to\infty}\mathbb{E}\left[\mathbb{E}_{\dot{X}_1}(m_n(\dot{X}_1)-m(\dot{X}_1))^2)\right]&\leq\lim_{n\to\infty}\mathbb{E}\left[\mathbb{E}_{\dot{X}_1}(T_{\zeta_n}m_n(\dot{X}_1)-m(\dot{X}_1))^2)\right]\\
&+\lim_{n\to\infty}\mathbb{E}\max_i m_n^2(X_i)\mathbb{I}(|m_n(X_i)|\geqslant\zeta_n).
\end{aligned}$$

The last term is zero by Assumption **(C.1)**, completing the $\mathbb{L}_2$ consistency result for the tree estimates $m_n(\cdot,\Theta)$. To get the result for the average estimate $\bar{m}_n=\mathbb{E}_\Theta m_n$, by Jensen's inequality

and Fubini's theorem, we have

$$\begin{aligned}\mathbb{E}_{\mathcal{D}_n}\left[\mathbb{E}_{\dot{X}_1}(\mathbb{E}_{\Theta}(m_n(\dot{X}_1,\Theta))-m(\dot{X}_1)^2)\right] &\leqslant \mathbb{E}_{\mathcal{D}_n}\left[\mathbb{E}_{\dot{X}_1}\mathbb{E}_{\Theta}(m_n(\dot{X}_1,\Theta)-m(\dot{X}_1))^2)\right] \\ &= \mathbb{E}_{\mathcal{D}_n,\Theta}\left[\mathbb{E}_{\dot{X}_1}(m_n(X_1,\Theta)-m(\dot{X}_1))^2)\right] \to 0\end{aligned}$$

□

S1.4 Choice of Truncation Threshold

Proposition S1.3. Under Assumptions 2, 3 and 4(a), the GLS tree estimator m_n satisfies the truncation threshold condition **(C.1)**.

Proof of Proposition S1.3. For any n, let the values corresponding to the t_n leaf nodes be denoted by $\mathbf{r} = \left(\mathbf{Z}^\top\mathbf{Q}\mathbf{Z}\right)^{-1}\mathbf{Z}^\top\mathbf{Q}\mathbf{Y}$, where i^{th} row of $\mathbf{Z}_{n\times t_n}$ denotes the membership of i^{th} data in any of the t_n leaf nodes. Let us define

$$\mathbf{B} := \left(\mathbf{Z}^\top\mathbf{Q}\mathbf{Z}\right);\ \mathbf{u} := \mathbf{Z}^\top\mathbf{Q}\mathbf{Y};\ \text{and } l_n := \underset{l\in\{1,\cdots,t_n\}}{\arg\max}\,|r_l|;\ \text{i.e., } \|\mathbf{r}\|_\infty = |\mathbf{r}_{l_n}|.$$

Then

$$\begin{aligned}\mathbf{B}_{l_n,l_n}\mathbf{r}_{l_n} = \mathbf{u}_{l_n} - \sum_{l\neq l_n}\mathbf{B}_{l_n,l}\mathbf{r}_l \implies \mathbf{B}_{l_n,l_n}|\mathbf{r}_{l_n}| &\leqslant |\mathbf{u}_{l_n}| + \sum_{l\neq l_n}|\mathbf{B}_{l_n,l}||\mathbf{r}_l| \\ \implies \|\mathbf{r}\|_\infty = |\mathbf{r}_{l_n}| &\leqslant \frac{|u_{l_n}|}{\mathbf{B}_{l_n,l_n} - \sum_{l\neq l_n}|\mathbf{B}_{l_n,l}|},\end{aligned}$$

since Lemma S3.3 and the Assumptions 2 and 3 implies

$$\mathbf{B}_{l_n,l_n} - \sum_{l\neq l_n}|\mathbf{B}_{l_n,l}| \geqslant \xi|\mathcal{C}_{l_n}| \text{ where } \xi = \min_{i=1,\ldots,q+1}(\mathbf{Q}_{ii} - \sum_{j\neq i}|\mathbf{Q}_{ij}|).$$

Using (A1),

$$
\begin{aligned}
|u_{l_n}| &= |\mathbf{Z}_{,l_n}^{\top}\mathbf{Q}\mathbf{Y}| \\
&= \alpha \sum_i \mathbf{Z}_{i,l_n}\mathbf{Y}_i + \sum_{j\neq j'=0}^{q} \rho_j\rho_{j'} \sum_i \mathbf{Z}_{i-j,l_n}\mathbf{Y}_{i-j'} + \sum_{i\in\tilde{\mathcal{A}}_1}\sum_{i'\in\tilde{\mathcal{A}}_2} \tilde{\gamma}_{i,i'}\mathbf{Z}_{i,l_n}\mathbf{Y}_{i'} \\
&\leqslant \max_i |y_i| \left[\alpha \sum_i \mathbf{Z}_{i,l_n} + \left[\sum_{j\neq j'=0}^{q} |\rho_j\rho_{j'}| \sum_i \mathbf{Z}_{i-j,l_n} + \sum_{i\in\tilde{\mathcal{A}}_1}\sum_{i'\in\tilde{\mathcal{A}}_2} |\tilde{\gamma}_{i,i'}|\mathbf{Z}_{i,l_n}\right]\right] \\
&\leqslant \max_i |y_i| \left(\alpha + \sum_{j\neq j'=0}^{q} |\rho_j\rho_{j'}| + \sum_{i\in\tilde{\mathcal{A}}_1}\sum_{i'\in\tilde{\mathcal{A}}_2} |\tilde{\gamma}_{i,i'}|\right) |\mathcal{C}_{l_n}|.
\end{aligned}
$$

Hence

$$
\begin{aligned}
\|\mathbf{r}\|_\infty &\leqslant \frac{\left(\alpha + \sum_{j\neq j'=0}^{q} |\rho_j\rho_{j'}| + \sum_{i\in\tilde{\mathcal{A}}_1}\sum_{i'\in\tilde{\mathcal{A}}_2} |\tilde{\gamma}_{i,i'}|\right)}{\xi} \max_i |y_i| \\
&\leqslant C\left[\|m\|_\infty + \max_i |\epsilon_i|\right],
\end{aligned} \tag{S12}
$$

and

$$
\begin{aligned}
\max_i [m_n(X_i) - T_{\zeta_n} m_n(X_i)]^2 &\leqslant \left[\|\mathbf{r}\|_\infty^2 \mathbb{I}\left(\|\mathbf{r}\|_\infty \geqslant \zeta_n\right)\right] \\
&\leqslant \left[C^2\left[\|m\|_\infty^2 + \max_i |\epsilon_i|^2\right] \mathbb{I}\left(C\left[\|m\|_\infty + \max_i |\epsilon_i|\right] \geqslant \zeta_n\right)\right] \\
&\leqslant \left[C^2\left[\|m\|_\infty^2 + \max_i |\epsilon_i|^2\right] \mathbb{I}\left(\max_i |\epsilon_i| \geqslant \tilde{C}\zeta_n\right)\right],
\end{aligned}
$$

where as $\zeta_n \to \infty$ we can choose a constant $\tilde{C}$ such that $\tilde{C}\zeta_n \leq \zeta_n/C - \|m\|_\infty$ for large n. Choosing ζ_n to be $\zeta_n/\bar{C}$ from Assumption 4(a), Condition **(C.1)** is satisfied. □

S1.5 Proof of corollaries

Proof of Corollary A1.1. To apply Theorem A1.1 to prove this corollary, we only need to show that moment condition $\lim_{n\to\infty} \frac{1}{n}\sum_i \mathbb{E}|m_n(X_i)|^{2+\delta} < \infty$ is satisfied.

For bounded errors, direct application of (S12) implies $|m_n(x)|$ is uniformly bounded and hence will satisfy the $(2+\delta)^{th}$ moment condition. Hence, part (a) is proved.

For part (b), let $D = \min_i \mathbf{Q}_{ii}$ and $O = \max_{i,j}\sum_i \sum_{j\neq i} |\mathbf{Q}_{ij}|$. By the condition for part 2, $D > \sqrt{2}O$. As we defined earlier, let

$$\mathbf{B} = \left(\mathbf{Z}^\top \mathbf{Q} \mathbf{Z}\right);\ \mathbf{Z}^\top \mathbf{Q} \mathbf{Y} = \mathbf{u};\implies \mathbf{B}_{l_1,l_1}\mathbf{r}_{l_1} = \mathbf{u}_{l_1} - \sum_{l_2 \neq l_1} \mathbf{B}_{l_1,l_2}\mathbf{r}_{l_1}$$

With these notations,

$$\sum_i |m_n(\mathbf{X}_i)|^{2+\dot{\delta}} = \sum_{l_1=1}^{t_n} |\mathcal{C}_{l_1}||r_{l_1}|^{2+\dot{\delta}}$$

Hence, $\mathbf{B}_{l_1,l_1}\mathbf{r}_{l_1} = \mathbf{u}_{l_1} - \sum_{l_2 \neq l_1} \mathbf{B}_{l_1,l_2}\mathbf{r}_{l_1}$ implies the following:

$$\begin{aligned}
\mathbf{r}_{l_1}\left(\sum_i \mathbf{Z}_{il_1}\mathbf{Q}_{ii} + \sum_i \sum_{j\neq i} \mathbf{Z}_{il_1}\mathbf{Z}_{jl_1}\mathbf{Q}_{ij}\right) &= \mathbf{u}_{l_1} - \sum_{l\neq l_1} \mathbf{r}_l \left(\sum_i \sum_{j\neq i} \mathbf{Z}_{il_1}\mathbf{Z}_{jl}\mathbf{Q}_{ij}\right) \\
\implies \mathbf{r}_{l_1}\sum_i \mathbf{Z}_{il_1}\mathbf{Q}_{ii} &= \mathbf{u}_{l_1} - \sum_{l} \mathbf{r}_l \left(\sum_i \sum_{j\neq i} \mathbf{Z}_{il_1}\mathbf{Z}_{jl}\mathbf{Q}_{ij}\right) \\
\implies |\mathbf{r}_{l_1}|\sum_i \mathbf{Z}_{il_1}\mathbf{Q}_{ii} &\leqslant |\mathbf{u}_{l_1}| + \sum_l w_l^{(l_1)}|\mathbf{r}_l| \\
\implies D|\mathcal{C}_{l_1}||\mathbf{r}_{l_1}| &\leqslant |\mathbf{u}_{l_1}| + \sum_l w_l^{(l_1)}|\mathbf{r}_l|
\end{aligned}$$

where

$$w_l^{(l_1)} = \sum_i \sum_{j\neq i} \mathbf{Z}_{il_1}\mathbf{Z}_{jl}|\mathbf{Q}_{ij}|$$

satisfies $\sum_l w_l^{(l_1)} = \sum_i \sum_{j\neq i} \mathbf{Z}_{il_1}|\mathbf{Q}_{ij}| \leq O|\mathcal{C}_{l_1}|$ and similarly, $\sum_{l_1} w_l^{(l_1)} \leq O|\mathcal{C}_l|$. Using Jensen's inequality twice we have

$$\begin{aligned}
|\mathbf{r}_{l_1}|^{2+\delta}(D|\mathcal{C}_{l_1}|)^{2+\delta} &\leqslant 2^{1+\delta}\left[|\mathbf{u}_{l_1}|^{2+\delta} + \left(\sum_l w_l^{(l_1)}\right)^{1+\delta}\sum_l w_l^{(l_1)}|\mathbf{r}_l|^{2+\delta}\right] \\
\implies \sum_{l_1=1}^{t_n}|\mathcal{C}_{l_1}||\mathbf{r}_{l_1}|^{2+\delta} &\leq \frac{2^{1+\delta}}{D^{2+\delta}}\sum_{l_1=1}^{t_n}\left[\frac{|\mathbf{u}_{l_1}|^{2+\delta}}{|\mathcal{C}_{l_1}|^{1+\delta}} + O^{1+\delta}\sum_l w_l^{(l_1)}|\mathbf{r}_l|^{2+\delta}\right] \\
\implies \sum_{l_1=1}^{t_n}|\mathcal{C}_{l_1}||\mathbf{r}_{l_1}|^{2+\delta} &\leq \frac{2^{1+\delta}}{D^{2+\delta}}\sum_{l_1=1}^{t_n}\frac{|\mathbf{u}_{l_1}|^{2+\delta}}{|\mathcal{C}_{l_1}|^{1+\delta}} + \frac{(2O)^{1+\delta}}{D^{2+\delta}}\sum_l |\mathbf{r}_l|^{2+\delta}\sum_{l_1=1}^{t_n} w_l^{(l_1)} \\
\implies \sum_{l_1=1}^{t_n}|\mathcal{C}_{l_1}||\mathbf{r}_{l_1}|^{2+\delta} &\leq \frac{2^{1+\delta}}{D^{2+\delta}}\sum_{l_1=1}^{t_n}\frac{|\mathbf{u}_{l_1}|^{2+\delta}}{|\mathcal{C}_{l_1}|^{1+\delta}} + \frac{2^{1+\delta}O^{2+\delta}}{D^{2+\delta}}\sum_l |\mathbf{r}_l|^{2+\delta}|\mathcal{C}_l|
\end{aligned}$$

Bring over the second term from the right hand side to the left, we have

$$\left(1 - \frac{2^{1+\delta}O^{2+\delta}}{D^{2+\delta}}\right) \frac{1}{n} \sum_{l_1=1}^{t_n} |\mathcal{C}_{l_1}||\mathbf{r}_{l_1}|^{2+\delta} \leq \frac{1}{n} \frac{2^{1+\delta}}{D^{2+\delta}} \sum_{l_1=1}^{t_n} \frac{|\mathbf{u}_{l_1}|^{2+\delta}}{|\mathcal{C}_{l_1}|^{1+\delta}}$$

As $D > \sqrt{2}O$, the term $1 - \frac{2^{1+\delta}O^{2+\delta}}{D^{2+\delta}}$ is positive and bounded away from 0 for small enough δ. Hence, we only need to show the right hand side has finite expectation.

$$\begin{aligned}
\frac{1}{n} \sum_{l=1}^{t_n} |\mathcal{C}_l| \left(\frac{|\mathbf{u}_l|}{|\mathcal{C}_l|}\right)^{2+\delta} \leq & \frac{1}{n} \sum_{l=1}^{t_n} |\mathcal{C}_l| \left(\frac{1}{|\mathcal{C}_l|} \sum_i \sum_{j \neq i} \mathbf{Z}_{il} |\mathbf{Q}_{ij}||Y_j|\right)^{2+\delta} \\
= & \frac{1}{n} \sum_{l=1}^{t_n} |\mathcal{C}_l| \left(\frac{1}{|\mathcal{C}_l|} \sum_{i \in \mathcal{C}_l} \sum_{j=-q}^{q} |\mathbf{Q}_{i,i+j}| Y_{i+j}|\right)^{2+\delta} \quad \text{[Assumption 2]} \\
\leq & \frac{1}{n} \sum_{l=1}^{t_n} |\mathcal{C}_l| \frac{1}{|\mathcal{C}_l|} \sum_{i \in \mathcal{C}_l} \left(\sum_{j=-q}^{q} |\mathbf{Q}_{i,i+j}||Y_{i+j}|\right)^{2+\delta} \quad \text{[Jensen's Inequality]} \\
= & \frac{1}{n} \sum_{i=1}^{n} \left(\sum_{j=-q}^{q} |\mathbf{Q}_{i,i+j}||Y_{i+j}|\right)^{2+\delta} \\
\leq & \frac{1}{n} K_1 \sum_{i=1}^{n} \sum_{j=-q}^{q} |Y_{i+j}|^{2+\delta}. \quad \text{[Jensen's Inequality]}
\end{aligned}$$

The last inequality also uses the fact that $\mathbf{Q}$ has only $O(q^2)$ unique entries whose maximum is hence bounded by some K_1. As ϵ_i's have finite $(2+\delta)^{th}$ moment, the expectation of this term is finite. □

Proof of Corollary A1.2. RF is RF-GLS with $\mathbf{Q} = \mathbf{I}$. Hence, the condition $\min_i \mathbf{Q}_{ii} > \sqrt{2} \max_i \sum_{j \neq I} |\mathbf{Q}_{ij}|$ is trivially satisfied. So, the proof follows from Corollary A1.1. □

S1.6 Examples

Proof of Proposition A1.2. We will directly apply Corollary A1.1 to prove the result and hence only need to prove that all assumptions are satisfied.

Assumption 1 is satisfied as AR processes have been shown to be β-mixing Mokkadem (1988), and sub-Gaussianity of the errors ensures all moments are finite (Vershynin (2010), Lemma 5.5). Next we verify Assumptions 2 and 3 on the working covariance matrix $\mathbf{\Sigma}$. We can write $\mathbf{\Sigma} = Cov(\tilde{\boldsymbol{\epsilon}})$

where $\tilde{\boldsymbol{\epsilon}} = (\tilde{\epsilon}_1, \ldots, \tilde{\epsilon}_n)^\top$ generated from an $AR(q)$ process with coefficients $\tilde{a}_i$'s. We note from (A3) that we can write $\mathbf{A}\tilde{\boldsymbol{\epsilon}} = \tilde{\boldsymbol{\eta}}$ where $\tilde{\boldsymbol{\eta}} = (\tilde{\eta}_{q+1}, \ldots, \tilde{\eta}_n)^\top$ is the white noise process used to generate the $\tilde{\epsilon}_i$'s.

$$\mathbf{A}_{(n-q)\times n} = \begin{pmatrix} -\tilde{a}_q & -\tilde{a}_{q-1} & \ldots & -\tilde{a}_1 & 1 & 0 & 0 & \ldots & 0 & 0 \\ 0 & -\tilde{a}_q & -\tilde{a}_{q-1} & \ldots & -\tilde{a}_1 & 1 & 0 & \ldots & 0 & 0 \\ \ldots & \ldots & \ldots & \ldots & \ldots & \ldots & \ldots & \ldots & \ldots & \ldots \\ 0 & 0 & \ldots & \ldots & \ldots & 0 & -\tilde{a}_q & \ldots & -\tilde{a}_1 & 1 \end{pmatrix},$$

Let $\tilde{\boldsymbol{\epsilon}}_{1:q} = (\tilde{\epsilon}_1, \ldots, \tilde{\epsilon}_q)^\top$ and let $\mathbf{L}_{q\times q}$ be the lower triangular Cholesky factor of the inverse of $\mathbf{M} = Cov(\tilde{\boldsymbol{\epsilon}}_{1:q})$ i.e. $\mathbf{L}^\top\mathbf{L} = \mathbf{M}^{-1}$. Then $\mathbb{V}(\mathbf{L}\tilde{\boldsymbol{\epsilon}}_{1:q}) = \mathbf{I}_{q\times q}$ and $Cov(\mathbf{L}\tilde{\boldsymbol{\epsilon}}_{1:q}, \mathbf{A}\tilde{\boldsymbol{\epsilon}}) = Cov(\mathbf{L}\tilde{\boldsymbol{\epsilon}}_{1:q}, \tilde{\boldsymbol{\eta}}) = \mathbf{O}$ as $\tilde{\boldsymbol{\eta}}$ is independent of $\boldsymbol{\epsilon}_{1:q}$. Defining $\sigma^2 = \mathbb{V}(\tilde{\eta}_i)$,

$$\mathbf{B} = \begin{pmatrix} \mathbf{L}_{q\times q} & \mathbf{O}_{q\times n-q} \\ \frac{1}{\sigma}\mathbf{A}_{n-q\times n} \end{pmatrix}, \tag{S13}$$

we have $Cov(\mathbf{B}\tilde{\boldsymbol{\epsilon}}) = \mathbf{I}$. Hence, $\mathbf{B}$ is the Cholesky factor $\boldsymbol{\Sigma}^{-1/2}$ making it clear that $\boldsymbol{\Sigma}^{-1/2}$ satisfies (2) with $\boldsymbol{\rho} = \frac{1}{\sigma}(-\tilde{a}_q, \ldots, -\tilde{a}_1, 1)^\top$.

To check Assumption 3, we first consider $q = 1$. Then $\mathbf{Q}$ is simply the autoregressive covariance matrix with parameter ρ, i.e., $\mathbf{Q}_{11} = \mathbf{Q}_{nn} = 1$, $\mathbf{Q}_{ii} = 1+\rho^2$ for $2 \le i \le n-1$, $\mathbf{Q}_{i,i+1} = \mathbf{Q}_{i,i-1} = -\rho$, $\mathbf{Q}_{ij} = 0$ for $|i-j| \ge 2$. Hence Assumption 3 is always satisfied as $1+\rho^2 > 2|\rho|$, i.e., for any $|\rho| < 1$. For unbounded errors, additionally $\mathbf{Q}$ needs to satisfy

$$\min_i \mathbf{Q}_{ii} > \sqrt{2}\max_i \sum_{j\neq i} |\mathbf{Q}_{ij}|. \tag{S14}$$

This reduces to $1 > 2\sqrt{2}|\rho|$, i.e., $|\rho| < 1/2\sqrt{2}$. For $q \ge 2$, by the statement of Proposition A1.2 part 2, Assumption 3 or condition (S14) (for unbounded errors) is directly satisfied.

Finally, we need to verify the tail bounds of Assumption 4. Proof of part (a) is same as that in Scornet et al. (2015) (p. 1733) with $\zeta_n = O(\log n)^2$ as maximum of n sub-Gaussian and correlated random variables satisfy the same tail bound (Software (Software), Theorem 1.14).

The same bound can be used to prove part (b). As ϵ_i's are identically distributed being a

stationary process, once again using Software (Software) (Theorem 1.14), we have

$$\mathbb{P}(\max_{i\in\mathcal{I}_n}|\epsilon_i| > C_\pi\sqrt{\log|\mathcal{I}_n|}) \leqslant |\mathcal{I}_n|^{\left(1-C_\pi^2/(2\sigma_\epsilon^2)\right)}$$

where σ_ϵ^2 denote the sub-Gaussian parameter of ϵ_i's. For any choice of C_π such that $C_\pi^2 > 2\sigma_\epsilon^2$, this goes to zero as $n \to \infty$, proving part (b).

For part (c), we make the observation that if $\mathbf{\Sigma}_0$ denote the true autoregressive covariance matrix of the errors $\boldsymbol{\epsilon}$, then following the argument above, we can write $\boldsymbol{\epsilon} = \mathbf{\Sigma}_0^{1/2}\boldsymbol{\eta}$ where $\boldsymbol{\eta} = (\eta_1, \ldots, \eta_n)$ is the collection of i.i.d. sub-Gaussian random variables. If $\mathbf{a}$ denotes the $n \times 1$ binary vector corresponding to the selection $\mathcal{I}_n$, we have

$$\begin{aligned}
\mathbb{P}(\frac{1}{|\mathcal{I}_n|}|\sum_{i\in\mathcal{I}_n}\epsilon_i| > \delta) =& \mathbb{P}(\frac{1}{|\mathcal{I}_n|}|\mathbf{a}^\top\mathbf{\Sigma}_0^{1/2}\boldsymbol{\eta}| > \delta) \\
\leq& C\exp(-c|\mathcal{I}_n|^2/\mathbf{a}^\top\mathbf{\Sigma}_0\mathbf{a}) \\
\leq& C\exp(-c|\mathcal{I}_n|/\lambda_{\max}(\mathbf{\Sigma}_0)).
\end{aligned}$$

The first inequality follows from Vershynin (2010) Proposition 5.10. Hence it is enough to show that $\lambda_{\max}(\mathbf{\Sigma}_0)$ is bounded. As ϵ_i's are generated from a stable $AR(q_0)$ process, the roots of the characteristic polynomial lie outside zero and the spectral density is bounded from above (Basu et al. (2015) Eqn. 2.6) which in turn bounds the spectral norm of $\mathbf{\Sigma}_0 = Cov(\boldsymbol{\epsilon})$ (Basu et al. (2015) Prop. 2.3), proving the first part of Assumption 4(c).

For the second part, as η_i are sub-Gaussian, η_i^2 are sub-exponential and $\|\boldsymbol{\epsilon}\|_2^2 \leq \lambda_{\max}(\mathbf{\Sigma}_0)\|\boldsymbol{\eta}\|_2^2$. We have

$$\begin{aligned}
\mathbb{P}(\frac{1}{n}\|\boldsymbol{\epsilon}\|_2^2 > 1.1\lambda_{\max}(\mathbf{\Sigma}_0)\mathbb{E}\eta_1^2) &\leq \mathbb{P}(\frac{1}{n}\|\boldsymbol{\eta}\|_2^2 > 1.1\mathbb{E}\eta_1^2) \\
&= \mathbb{P}(\frac{1}{n}\sum_i \eta_i^2 - \mathbb{E}\eta_1^2 > .1\mathbb{E}\eta_1^2) \\
&\leq C\exp(-cn).
\end{aligned}$$

The last inequality is from Vershynin (2010) Corollary 5.17. □

Proof of Proposition A1.1. We first prove the result without the nugget process $\epsilon^*(\ell)$. Once we prove that the Matérn GP is β-mixing, the extension to include the nugget is trivial as sum of a β-mixing and an i.i.d process by (as proved from S20). Let $\epsilon(\ell)$ denote a Matérn process with

$\nu = q_0 - 1/2$ on the real line, q_0 being a positive integer. From Hartikainen and Särkkä (2010), $\epsilon(\ell)$ admits a state-space representation:

$$a_q \frac{\partial^{q_0}\epsilon(\ell)}{\partial \ell^{q_0}} + a_{q_0-1}\frac{\partial^{q_0-1}\epsilon(\ell)}{\partial \ell^{q_0-1}} + \ldots a_1 \frac{\partial \epsilon(\ell)}{\partial \ell} + \ldots a_0 \epsilon(\ell) = b_0 z(\ell)$$

where $z(\ell)$ is a white-noise process with covariance $\mathbb{I}(\ell - \ell' = 0)$. Processes satisfying such q_0-order stochastic differential equations are continuous-domain AR(q_0) processes (Rasmussen (2003) Eqn. B.2) which when sampled on a discrete integer lattice become an ARMA(q_0, q_0') process ($q_0' < q_0$) (Ihara (1993) Theorem 2.7.1). As ARMA processes are β-mixing Mokkadem (1988), the process $\epsilon(\ell_i)$ sampled on the integer lattice is proved to be β-mixing. Since they are also Gaussian, they have a bounded $(2+\delta)^{th}$ moment. Hence, Assumption 1 is satisfied.

Assumption (2) is directly satisfied by a NNGP working covariance matrix $\mathbf{\Sigma}$ as explained in Section A1.3.1. Since we are considering Gaussian (unbounded) errors, to apply Theorem A1.1, the working precision matrix $\mathbf{Q}$ needs to satisfy (S14) which will also ensure Assumption 3 holds. Let $d_i = \mathbf{Q}_{ii}$ and $o_i = \sum_{j \neq i} |\mathbf{Q}_{ij}|$. For an NNGP using q nearest neighbors, there are only $q+1$ unique combinations of (o_i, d_i). For each combination, as $o_i \to 1$ and $d_i \to 0$ as $\phi \to \infty$, there exists ϕ_i such that $o_i > max_{i=1}^{q+1}\, d_i$, choosing $\phi > K = \max_{i=1}^{q+1} \phi_i$, (S14) holds.

For the tail-bounds in Assumption 4, we note that the proofs for showing Assumptions 4(a) and (b) hold remain identical to those proofs in Proposition A1.2 as they only require result on maximal inequalities of sub-Gaussian variables. For part(c), the proof will once again emulate that from Proposition A1.2 and we only need to show that the spectral norm of $\mathbf{\Sigma}_0 = Cov(\boldsymbol{\epsilon}, \boldsymbol{\epsilon})$ is bounded. Let $f_{\mathbb{Z}}(\omega)$ the spectral density of a Matern process when sampled on the integer lattice $\mathbb{Z}$. Then, $f_{\mathbb{Z}}(\omega) = \sum_{k=\infty}^{\infty} C(k) \exp(ik\omega)$. Hence, $\sup_\omega |f_{\mathbb{Z}}(\omega)| \leq \sum_{k=\infty}^{\infty} C(k)$. Using Abramowitz and Stegun (1948) (Eqn. 9.7.2) for large k, $C(k)$ is equivalent to $k^{q_0-1}\exp(-k)$. So the above series is summable and $\sup_\omega |f_{\mathbb{Z}}(\omega)| < \infty$. This is sufficient to uniformly bound the spectral norm $\|\mathbf{\Sigma}_0\|_2$ (Basu et al. (2015) Proposition 2.3). □

S2 Additional Proofs

*Proof of Proposition A2.1 under Scenario **R2**.* In Section S1, we proved the lemma under **R1**. The equicontinuity of the term $\frac{1}{n}|\mathbf{Y}^\top\mathbf{Q}\mathbf{Z}_1^{(0)}\hat{\boldsymbol{\beta}}(\mathbf{Z}_1^{(0)}) - \mathbf{Y}^\top\mathbf{Q}\mathbf{Z}_2^{(0)}\hat{\boldsymbol{\beta}}(\mathbf{Z}_2^{(0)})|$ established in that proof also holds under **R2** since the term does not involve the children nodes. So, it is enough to prove equicontinuity of $\frac{1}{n}|\mathbf{Y}^\top\mathbf{Q}\mathbf{Z}_1\hat{\boldsymbol{\beta}}(\mathbf{Z}_1) - \mathbf{Y}^\top\mathbf{Q}\mathbf{Z}_2\hat{\boldsymbol{\beta}}(\mathbf{Z}_2)|$ under conditions of the proposition and **R2**.

For $h = 1, 2$, $\forall i \in \{1, 2, \cdots, n\}$, define matrices $\tilde{\mathbf{Z}}_h^{(0)}$ such that $\tilde{\mathbf{Z}}_{h_{i,l}}^{(0)} = \mathbb{I}\left(\mathbf{x}_i \in \ddot{\mathcal{C}}_l^{(h)}\right)$; $l = 1, 2, \cdots, g^{(k)} - 1$, and

$$\tilde{\mathbf{Z}}_{h_{i,l}}^{(0)} = \mathbb{I}\left(\frac{Vol(\ddot{\mathcal{C}}_l^{(h)})}{Vol(\ddot{\mathcal{C}}_{g^{(k)}}^{(h)}) + Vol(\ddot{\mathcal{C}}_{g^{(k)}+1}^{(h)})} \geqslant \sqrt{\delta}\right)\mathbb{I}\left(\mathbf{x}_i \in \{\ddot{\mathcal{C}}_{g^{(k)}}^{(h)} \cup \ddot{\mathcal{C}}_{g^{(k)}+1}^{(h)}\}\right); \; l = g^{(k)}, g^{(k)} + 1.$$

Basically, $\tilde{\mathbf{Z}}_h^{(0)}$ adds a column of zeroes to $\mathbf{Z}_h^{(0)}$ to match the dimensions of $\mathbf{Z}_h$ and rearranges the columns so that the column of zeroes aligns with the column in $\mathbf{Z}_h$ corresponding to the child node with few members (which is posited under **R2**).

$$\begin{aligned}
&\frac{1}{n}|\mathbf{Y}^\top\mathbf{Q}\mathbf{Z}_1\hat{\boldsymbol{\beta}}(\mathbf{Z}_1) - \mathbf{Y}^\top\mathbf{Q}\mathbf{Z}_2\hat{\boldsymbol{\beta}}(\mathbf{Z}_2)| \\
&= |\frac{1}{n}\mathbf{Y}^\top\mathbf{Q}\mathbf{Z}_1\left[\mathbf{Z}_1^\top\mathbf{Q}\mathbf{Z}_1\right]^{-1}\mathbf{Z}_1^\top\mathbf{Q}\mathbf{y} - \frac{1}{n}\mathbf{Y}^\top\mathbf{Q}\mathbf{Z}_2\left[\mathbf{Z}_2^\top\mathbf{Q}\mathbf{Z}_2\right]^{-1}\mathbf{Z}_2^\top\mathbf{Q}\mathbf{y}| \\
&\leqslant J_1 + J_2 + J_3 + J_4 + J_5
\end{aligned}$$

where,

$$J_1 = \frac{1}{n}|\mathbf{Y}^\top\mathbf{Q}\mathbf{Z}_1\left(\left[\mathbf{Z}_1^\top\mathbf{Q}\mathbf{Z}_1\right]^{-1} - \left[\tilde{\mathbf{Z}}_1^{(0)^\top}\mathbf{Q}\tilde{\mathbf{Z}}_1^{(0)}\right]^+\right)\mathbf{Z}_1^\top\mathbf{Q}\mathbf{y}|,$$

$$J_2 = \frac{1}{n}|\mathbf{Y}^\top\mathbf{Q}\mathbf{Z}_1\left[\tilde{\mathbf{Z}}_1^{(0)^\top}\mathbf{Q}\tilde{\mathbf{Z}}_1^{(0)}\right]^+\mathbf{Z}_1^\top\mathbf{Q}\mathbf{y} - \mathbf{Y}^\top\mathbf{Q}\tilde{\mathbf{Z}}_1^{(0)}\left[\tilde{\mathbf{Z}}_1^{(0)^\top}\mathbf{Q}\tilde{\mathbf{Z}}_1^{(0)}\right]^+\tilde{\mathbf{Z}}_1^{(0)^\top}\mathbf{Q}\mathbf{y}|,$$

$$J_3 = \frac{1}{n}|\mathbf{Y}^\top\mathbf{Q}\mathbf{Z}_2\left(\left[\mathbf{Z}_2^\top\mathbf{Q}\mathbf{Z}_2\right]^{-1} - \left[\tilde{\mathbf{Z}}_2^{(0)^\top}\mathbf{Q}\tilde{\mathbf{Z}}_2^{(0)}\right]^+\right)\mathbf{Z}_2^\top\mathbf{Q}\mathbf{y}|,$$

$$J_4 = \frac{1}{n}|\mathbf{Y}^\top\mathbf{Q}\mathbf{Z}_2\left[\tilde{\mathbf{Z}}_2^{(0)^\top}\mathbf{Q}\tilde{\mathbf{Z}}_2^{(0)}\right]^+\mathbf{Z}_2^\top\mathbf{Q}\mathbf{y} - \mathbf{Y}^\top\mathbf{Q}\tilde{\mathbf{Z}}_2^{(0)}\left[\tilde{\mathbf{Z}}_2^{(0)^\top}\mathbf{Q}\tilde{\mathbf{Z}}_2^{(0)}\right]^+\tilde{\mathbf{Z}}_2^{(0)^\top}\mathbf{Q}\mathbf{y}|,$$

$$J_5 = \frac{1}{n}|\mathbf{Y}^\top\mathbf{Q}\tilde{\mathbf{Z}}_1^{(0)}\left[\tilde{\mathbf{Z}}_1^{(0)^\top}\mathbf{Q}\tilde{\mathbf{Z}}_1^{(0)}\right]^+\tilde{\mathbf{Z}}_1^{(0)^\top}\mathbf{Q}\mathbf{y} - \mathbf{Y}^\top\mathbf{Q}\tilde{\mathbf{Z}}_2^{(0)}\left[\tilde{\mathbf{Z}}_2^{(0)^\top}\mathbf{Q}\tilde{\mathbf{Z}}_2^{(0)}\right]^+\tilde{\mathbf{Z}}_2^{(0)^\top}\mathbf{Q}\mathbf{y}|.$$

First we focus on J_5. The terms are of the form

$$\mathbf{Y}^\top \mathbf{Q}\tilde{\mathbf{Z}}_h^{(0)} \left[\tilde{\mathbf{Z}}_h^{(0)^\top} \mathbf{Q}\tilde{\mathbf{Z}}_h^{(0)}\right]^+ \tilde{\mathbf{Z}}_h^{(0)^\top} \mathbf{Q}\mathbf{y} = \mathbf{Y}^\top \mathbf{Q}^{\frac{1}{2}} \dot{\mathbf{P}}_{\mathbf{Q}^{\frac{\top}{2}}\tilde{\mathbf{Z}}_h^{(0)}} \mathbf{Q}^{\frac{\top}{2}} \mathbf{y}$$

where $\dot{\mathbf{P}}_{\mathbf{X}}$ denotes the projection operator for a matrix $\mathbf{X}$. By construction of $\tilde{\mathbf{Z}}_h^{(0)}$, column space of $\mathbf{Q}^{\frac{\top}{2}}\tilde{\mathbf{Z}}_h^{(0)}$ is same as that of $\mathbf{Q}^{\frac{\top}{2}}\mathbf{Z}_h^{(0)}$. Hence, $\mathbf{P}_{\mathbf{Q}^{\frac{\top}{2}}\tilde{\mathbf{Z}}_h^{(0)}} = \mathbf{P}_{\mathbf{Q}^{\frac{\top}{2}}\mathbf{Z}_h^{(0)}}$, and,

$$\mathbf{Y}^\top \mathbf{Q}\tilde{\mathbf{Z}}_h^{(0)} \left[\tilde{\mathbf{Z}}_h^{(0)^\top} \mathbf{Q}\tilde{\mathbf{Z}}_h^{(0)}\right]^+ \tilde{\mathbf{Z}}_h^{(0)^\top} \mathbf{Q}\mathbf{y} = \mathbf{Y}^\top \mathbf{Q}\mathbf{Z}_h^{(0)} \left[\mathbf{Z}_h^{(0)^\top} \mathbf{Q}\mathbf{Z}_h^{(0)}\right]^+ \mathbf{Z}_h^{(0)^\top} \mathbf{Q}\mathbf{y}.$$

Thus, J_5 simply becomes $\frac{1}{n}|\mathbf{Y}^\top \mathbf{Q}\mathbf{Z}_1^{(0)}\hat{\boldsymbol{\beta}}(\mathbf{Z}_1^{(0)}) - \mathbf{Y}^\top \mathbf{Q}\mathbf{Z}_2^{(0)}\hat{\boldsymbol{\beta}}(\mathbf{Z}_2^{(0)})|$ which has been shown to be small earlier in the proof of this Lemma in Section S1, irrespective of **R1** or **R2**.

To simplify J_2, let for $h = 1, 2$, $\tilde{\mathbf{Z}}_h^L$ and $\tilde{\mathbf{Z}}_h^R$ respectively denote the columns of $\mathbf{Z_h}$ corresponding to the bigger and smaller child nodes. As the zero column of $\tilde{\mathbf{Z}}_h^{(0)}$ aligns with the column $\tilde{\mathbf{Z}}_h^R$ of $\mathbf{Z_h}$, we have

$$\mathbf{Z}_h \left[\tilde{\mathbf{Z}}_h^{(0)^\top} \mathbf{Q}\tilde{\mathbf{Z}}_h^{(0)}\right]^+ \mathbf{Z}_h^\top = \tilde{\mathbf{Z}}_h \left[\mathbf{Z}_h^{(0)^\top} \mathbf{Q}\mathbf{Z}_h^{(0)}\right]^{-1} \tilde{\mathbf{Z}}_h^\top \text{ where } \tilde{\mathbf{Z}}_h = \left(\mathbf{Z}_{h\cdot,1:g^{(k)}-1} \;\; \tilde{\mathbf{Z}}_h^L\right). \tag{S15}$$

Writing, $\mathbf{Z}_1^{(0)} = \tilde{\mathbf{Z}}_1 + \left[\mathbf{0}_{n\times g^{(k)}-1} \;\; \tilde{\mathbf{Z}}_1^R\right]$, we have

$$\begin{aligned}
J_2 &= \frac{1}{n}\left|\mathbf{Y}^\top \mathbf{Q}\mathbf{Z}_1^{(0)} \left[\mathbf{Z}_1^{(0)^\top} \mathbf{Q}\mathbf{Z}_1^{(0)}\right]^{-1} \mathbf{Z}_1^{(0)^\top} \mathbf{Q}\mathbf{y} - \mathbf{Y}^\top \mathbf{Q}\tilde{\mathbf{Z}}_1 \left[\mathbf{Z}_1^{(0)^\top} \mathbf{Q}\mathbf{Z}_1^{(0)}\right]^{-1} \tilde{\mathbf{Z}}_1^\top \mathbf{Q}\mathbf{y}\right| \\
&\leq \frac{1}{n}\left| \left[\mathbf{0}_{1\times g^{(k)}-1} \;\; \mathbf{Y}^\top \mathbf{Q}\tilde{\mathbf{Z}}_1^R\right] \left[\mathbf{Z}_1^{(0)^\top} \mathbf{Q}\mathbf{Z}_1^{(0)}\right]^{-1} \left[\mathbf{0}_{1\times g^{(k)}-1} \;\; \mathbf{Y}^\top \mathbf{Q}\tilde{\mathbf{Z}}_1^R\right]^\top \right| \\
&\quad + \frac{2}{n}\left| \left[\mathbf{0}_{1\times g^{(k)}-1} \;\; \mathbf{Y}^\top \mathbf{Q}\tilde{\mathbf{Z}}_1^R\right] \left[\mathbf{Z}_1^{(0)^\top} \mathbf{Q}\mathbf{Z}_1^{(0)}\right]^{-1} (\mathbf{Y}^\top \mathbf{Q}\tilde{\mathbf{Z}}_1)^\top \right|
\end{aligned}$$

Denote these two terms on the right hand side respectively as J_{21} and J_{22}. For J_{21}, we have on Ω_n

(set with probability at least $1-3\pi/4$),

$$
\begin{aligned}
J_{21} &= \frac{1}{n}\Big|\begin{bmatrix}\mathbf{0}_{1\times g^{(k)}-1} & \mathbf{Y}^\top\mathbf{Q}\tilde{\mathbf{Z}}_1^R\end{bmatrix}\left[\mathbf{Z}_1^{(0)^\top}\mathbf{Q}\mathbf{Z}_1^{(0)}\right]^{-1}\begin{bmatrix}\mathbf{0}_{1\times g^{(k)}-1} & \mathbf{Y}^\top\mathbf{Q}\tilde{\mathbf{Z}}_1^R\end{bmatrix}^\top\Big| \\
&\leqslant \frac{1}{n}\lambda_{\max}\left[\mathbf{Z}_1^{(0)^\top}\mathbf{Q}\mathbf{Z}_1^{(0)}\right]^{-1}\| \left(\mathbf{0}_{1\times g^{(k)}-1}, \mathbf{Y}^\top\mathbf{Q}\tilde{\mathbf{Z}}_1^R\right)\|^2 \\
&= \left(\frac{\mathbf{Y}^\top\mathbf{Q}\tilde{\mathbf{Z}}_1^R}{n}\right)^2\frac{n}{|\lambda_{min}(\mathbf{Z}_1^{(0)^\top}\mathbf{Q}\mathbf{Z}_1^{(0)})|} \\
&\leqslant \left(\frac{\mathbf{Y}^\top\mathbf{Q}\tilde{\mathbf{Z}}_1^R}{n}\right)^2\frac{1}{\xi}\left[\left(\min_{l\in 1,2,\cdots,g^{(k)}-1}|\ddot{\mathcal{C}}_l^{(1)}|, |\ddot{\mathcal{C}}_{g^{(k)}}^{(1)}| + |\ddot{\mathcal{C}}_{g^{(k)}+1}^{(1)}|\right)\right]^{-1} n \\
&\leqslant \left(\frac{\mathbf{Y}^\top\mathbf{Q}\tilde{\mathbf{Z}}_1^R}{n}\right)^2\frac{1}{\xi}\frac{1}{\varepsilon^D-\delta^2}; [\text{ for sufficiently large } n]
\end{aligned}
$$

Here the last two inequalities are obtained similar to the derivation of (S9d), exploiting the lower bound of the eigenvalue in (S9b) and the fact that $Vol(\ddot{\mathcal{C}}_l^{(1)}) \geqslant \varepsilon^D, \forall l = 1, 2, \cdots, g^{(k)}-1$, and $Vol(\ddot{\mathcal{C}}_{g^{(k)}}^{(1)} \cup \ddot{\mathcal{C}}_{g^{(k)}+1}^{(1)}) \geqslant \varepsilon^D$, in conjunction with the Glivenko-Cantelli result of (S5) . We now focus on $\mathbf{Y}^\top\mathbf{Q}\tilde{\mathbf{Z}}_h^R$. Let $l_0 \in \{g^{(k)}, g^{(k)}+1\}$ denote the column of $\mathbf{Z}_1$ corresponding to the smaller child node. Using (A1), we have

$$
\begin{aligned}
\frac{1}{n}\mathbf{Y}^\top\mathbf{Q}\tilde{\mathbf{Z}}_1^R &= \alpha\frac{\sum_i \mathbf{Z}_{1_{i,l_0}} m(X_i)}{n} + \sum_{j\neq j'=0}^{q}\rho_j\rho_{j'}\frac{\sum_i \mathbf{Z}_{1_{i-j,l_0}} m(X_{i-j'})}{n} \\
&\quad + \alpha\frac{\sum_i \mathbf{Z}_{1i,l_0}\epsilon_i}{n} + \sum_{j\neq j'=0}^{q}\rho_j\rho_{j'}\frac{\sum_i \mathbf{Z}_{1_{i-j,l_0}}\epsilon_{i-j'}}{n} \\
&\quad + \frac{1}{n}\sum_{i\in\tilde{\mathcal{A}}_1}\sum_{i'\in\tilde{\mathcal{A}}_2}\tilde{\gamma}_{i,i'}\mathbf{Z}_{1_{h_1,l_0}} m(X_{h_2}) + \frac{1}{n}\sum_{i\in\tilde{\mathcal{A}}_1}\sum_{i'\in\tilde{\mathcal{A}}_2}\tilde{\gamma}_{i,i'}\mathbf{Z}_{1_{h_1,l_0}}\epsilon_{h_2}
\end{aligned}
$$

On Ω_n, using the tail bounds from Assumptions 4(b) and (c) and Lemma S3.2, and the fact that $|m(\mathbf{x})| \leqslant M_0; \mathbf{x} \in [0,1]^D$, we have for large enough n,

$$
\begin{aligned}
\frac{1}{n}|\mathbf{Y}^\top\mathbf{Q}\tilde{\mathbf{Z}}_1^R| &\leqslant \frac{1}{n}M_0\left[\alpha\sum_i \mathbf{Z}_{i,l_0} + \left[\sum_{j\neq j'=0}^{q}|\rho_j\rho_{j'}|\sum_i \mathbf{Z}_{i-j,l_0} + \sum_{i\in\tilde{\mathcal{A}}_1}\sum_{i'\in\tilde{\mathcal{A}}_2}|\tilde{\gamma}_{i,i'}|\mathbf{Z}_{h_1,l_0}\right]\right] \\
&\quad + \frac{\alpha}{n}\Big|\sum_i \mathbf{Z}_{i,l_0}\epsilon_i\Big| + \sum_{j\neq j'=0}^{q}\frac{|\rho_j\rho_{j'}|}{n}|\sum_i \mathbf{Z}_{i-j,l_0}\epsilon_{i-j'}| + \frac{C_\pi\sqrt{\ln n}}{n}\sum_{i\in\tilde{\mathcal{A}}_1}\sum_{i'\in\tilde{\mathcal{A}}_2}|\tilde{\gamma}_{i,i'}|\mathbf{Z}_{h_1,l_0} \\
&\leqslant (\alpha + \sum_{j\neq j'=0}^{q}|\rho_j\rho_{j'}|)\left(M_0\frac{\sum_i \mathbf{Z}_{i,l_0}}{n} + \phi\right) + (C_\pi\sqrt{\log n} + M_0)\frac{1}{n}\sum_{i\in\tilde{\mathcal{A}}_1}\sum_{i'\in\tilde{\mathcal{A}}_2}|\tilde{\gamma}_{i,i'}|.
\end{aligned}
$$

The terms involving $\tilde{\gamma}_{i,i'}$ are $o(1)$ as there are only $O(q^2)$ of them and the factor $\log n/n \to 0$. The terms involving ϵ_i's were bounded using Lemma S3.2.

Under **R2** on Ω_n, using (S5), for large enough n

$$\frac{1}{n}|\mathbf{Y}^\top \mathbf{Q}\tilde{\mathbf{Z}}_1^R| \leqslant (\alpha + \sum_{j\neq j'=0}^{q} |\rho_j \rho_{j'}|) M_0(\sqrt{\delta} + \delta^2) + C_7\phi. \tag{S16}$$

As ϕ can be chosen arbitrarily small, we have $J_{21} \to 0$ uniformly as $\delta \to 0$. Similarly, on Ω_n for large enough n, we have

$$\begin{aligned}
J_{23} &:= \frac{1}{n}|\mathbf{Y}^\top \mathbf{Q}\tilde{\mathbf{Z}}_1 \left[\mathbf{Z}_1^{(0)\top} \mathbf{Q}\mathbf{Z}_1^{(0)}\right]^{-1} (\mathbf{Y}^\top \mathbf{Q}\tilde{\mathbf{Z}}_1)^\top| \\
&\leqslant \frac{1}{n}\lambda_{\max}\left(\left[\mathbf{Z}_1^{(0)\top} \mathbf{Q}\mathbf{Z}_1^{(0)}\right]^{-1}\right) \|\mathbf{Y}^\top \mathbf{Q}\tilde{\mathbf{Z}}_1\|_2^2 \\
&\leqslant \frac{\|\mathbf{Y}^\top \mathbf{Q}\tilde{\mathbf{Z}}_1\|_2^2/n^2}{\lambda_{\min}\left(\left[\mathbf{Z}_1^{(0)\top} \mathbf{Q}\mathbf{Z}_1^{(0)}\right]\right)/n} \\
&\leqslant \frac{C_8}{\varepsilon^D - \delta^2}
\end{aligned}$$

The last inequality follows from (S9c) as the minimum volume of the parent nodes are ϵ^D, and the following bound for $\frac{1}{n^2}\|\mathbf{Y}^\top \mathbf{Q}\tilde{\mathbf{Z}}_1\|_2^2$ on Ω_n. Using sub-multiplicative property of $\mathbb{L}_2$ norm, we write $\frac{1}{n}\|\mathbf{Y}^\top \mathbf{Q}\tilde{\mathbf{Z}}_1\|_2 \leq \frac{1}{n}\|\mathbf{Q}\|_2\|\widetilde{\mathbf{Z}}_1\|_2\|\mathbf{Y}\|$. Following assumption 4 (c) and its corollary Lemma S3.2, $y^\top y/n$ is $O(1)$. From Assumption 2, $\|\mathbf{Q}\|_2$ is bounded. Finally, as $\mathbf{Z}_1^\top \mathbf{Z}_1/n$ is diagonal with entries $\leqslant 1$, we have $\frac{1}{n^2}\|\mathbf{Y}^\top \mathbf{Q}\tilde{\mathbf{Z}}_1\|_2^2$ to be $O(1)$.

By Cauchy-Schwartz inequality, we have $J_{22} \leq \sqrt{J_{21}J_{23}}$ thereby making the J_2 term $O(1)$. By symmetry of **R2**, J_4 is also $O(1)$.

Next we focus on J_1. Noting that $\tilde{\mathbf{Z}}_1^{(0)}$ has a column of zeros aligned with the column $\tilde{\mathbf{Z}}_1^R$ of $\mathbf{Z}_1$, without loss of generality we can write

$$J_1 = \frac{1}{n}\left|\boldsymbol{\psi}^\top \left(\begin{bmatrix} \tilde{\mathbf{Z}}_1^\top \mathbf{Q}\tilde{\mathbf{Z}}_1 & \tilde{\mathbf{Z}}_1^\top \mathbf{Q}\tilde{\mathbf{Z}}_1^R; \\ \tilde{\mathbf{Z}}_1^{R\top} \mathbf{Q}\tilde{\mathbf{Z}}_1 & \tilde{\mathbf{Z}}_1^{R\top} \mathbf{Q}\tilde{\mathbf{Z}}_1^R \end{bmatrix}^{-1} - \begin{bmatrix} \left(\mathbf{Z}_1^{(0)\top} \mathbf{Q}\mathbf{Z}_1^{(0)}\right)^{-1} & 0 \\ 0 & 0 \end{bmatrix}\right)\boldsymbol{\psi}\right|$$

where, $\boldsymbol{\psi}^\top = \boldsymbol{\psi}_1^\top + \boldsymbol{\psi}_2^\top$; $\boldsymbol{\psi}_1^\top = \begin{bmatrix} \mathbf{Y}^\top \mathbf{Q} \tilde{\mathbf{Z}}_1 & 0 \end{bmatrix}$; $\boldsymbol{\psi}_2^\top = \begin{bmatrix} \mathbf{0}_{1\times g^{(k)}} & \mathbf{Y}^\top \mathbf{Q} \tilde{\mathbf{Z}}_1^R \end{bmatrix}$. Let

$$\mathbf{U}_{11} = \tilde{\mathbf{Z}}_1^\top \mathbf{Q} \tilde{\mathbf{Z}}_1; \quad \mathbf{U}_{12} = \tilde{\mathbf{Z}}_1^\top \mathbf{Q} \tilde{\mathbf{Z}}_1^R; \quad \mathbf{U}_{22} = \tilde{\mathbf{Z}}_1^{R^\top} \mathbf{Q} \tilde{\mathbf{Z}}_1^R; \quad \mathbf{U}_{11}^{(0)} = \mathbf{Z}_1^{(0)^\top} \mathbf{Q} \mathbf{Z}_1^{(0)};$$

Hence, we have

$$J_1 \leqslant J_{11} + J_{12} + J_{13}$$

where,

$$J_{11} = \frac{1}{n} \left| \boldsymbol{\psi}_1^\top \left(\begin{bmatrix} \mathbf{U}_{11} & \mathbf{U}_{12}; \\ \mathbf{U}_{12}^\top & \mathbf{U}_{22} \end{bmatrix}^{-1} - \begin{bmatrix} \mathbf{U}_{11}^{(0)^{-1}} & 0 \\ 0 & 0 \end{bmatrix} \right) \boldsymbol{\psi}_1 \right|$$

$$J_{12} = \frac{2}{n} \left| \boldsymbol{\psi}_2^\top \left(\begin{bmatrix} \mathbf{U}_{11} & \mathbf{U}_{12}; \\ \mathbf{U}_{12}^\top & \mathbf{U}_{22} \end{bmatrix}^{-1} - \begin{bmatrix} \mathbf{U}_{11}^{(0)^{-1}} & 0 \\ 0 & 0 \end{bmatrix} \right) \boldsymbol{\psi}_1 \right|$$

$$J_{13} = \frac{1}{n} \left| \boldsymbol{\psi}_2^\top \left(\begin{bmatrix} \mathbf{U}_{11} & \mathbf{U}_{12}; \\ \mathbf{U}_{12}^\top & \mathbf{U}_{22} \end{bmatrix}^{-1} - \begin{bmatrix} \mathbf{U}_{11}^{(0)^{-1}} & 0 \\ 0 & 0 \end{bmatrix} \right) \boldsymbol{\psi}_2 \right|$$

$$\begin{aligned} J_{11} =& \frac{1}{n} \left| \boldsymbol{\psi}_1^\top \left(\begin{bmatrix} \mathbf{U}_{11} & \mathbf{U}_{12}; \\ \mathbf{U}_{12}^\top & \mathbf{U}_{22} \end{bmatrix}^{-1} - \begin{bmatrix} \mathbf{U}_{11}^{(0)^{-1}} & 0 \\ 0 & 0 \end{bmatrix} \right) \boldsymbol{\psi}_1 \right| \\ =& \frac{1}{n} |\mathbf{Y}^\top \mathbf{Q} \tilde{\mathbf{Z}}_1 \left(\left(\mathbf{U}_{11} - \mathbf{U}_{12} \mathbf{U}_{22}^{-1} \mathbf{U}_{12}^\top \right)^{-1} - \mathbf{U}_{11}^{(0)^{-1}} \right) (\mathbf{Y}^\top \mathbf{Q} \tilde{\mathbf{Z}}_1)^\top | \\ \leqslant& \frac{\mathbf{Y}^\top \mathbf{Q} \tilde{\mathbf{Z}}_1 (\mathbf{Y}^\top \mathbf{Q} \tilde{\mathbf{Z}}_1)^\top}{n} \| \left(\left(\mathbf{U}_{11} - \mathbf{U}_{12} \mathbf{U}_{22}^{-1} \mathbf{U}_{12}^\top \right)^{-1} - \mathbf{U}_{11}^{(0)^{-1}} \right) \|_2 \\ \leqslant& \frac{\mathbf{Y}^\top \mathbf{Q} \tilde{\mathbf{Z}}_1 (\mathbf{Y}^\top \mathbf{Q} \tilde{\mathbf{Z}}_1)^\top}{n} \frac{\| \mathbf{U}_{11} - \mathbf{U}_{12} \mathbf{U}_{22}^{-1} \mathbf{U}_{12}^\top - \mathbf{U}_{11}^{(0)} \|_2}{\lambda_{min} \left(\mathbf{U}_{11} - \mathbf{U}_{12} \mathbf{U}_{22}^{-1} \mathbf{U}_{12}^\top \right) \lambda_{min} \left(\mathbf{U}_{11}^{(0)} \right)} \end{aligned}$$

where the second inequality follows from taking $\mathbf{A} = \mathbf{U}_{11} - \mathbf{U}_{12} \mathbf{U}_{22}^{-1} \mathbf{U}_{12}^\top$ and $\mathbf{B} = \mathbf{U}_{11}^{(0)}$ in the following identity.

$$\begin{aligned} \|\mathbf{A}^{-1} - \mathbf{B}^{-1}\|_2 &= \|\mathbf{A}^{-1}(\mathbf{B} - \mathbf{A})\mathbf{B}^{-1}\|_2 \\ &\leqslant \|\mathbf{A}^{-1}\|_2 \|\mathbf{A} - \mathbf{B}\|_2 \|\mathbf{B}^{-1}\|_2; [\|.\|_2 \text{ is submultiplicative}] \\ &= \frac{\|\mathbf{A} - \mathbf{B}\|_2}{\lambda_{\min}(\mathbf{A}) \lambda_{\min}(\mathbf{B})}. \end{aligned}$$

We have already established the following bound on $|\lambda_{min}^{-1}\left(\mathbf{U}_{11}^{(0)}\right)| = |\lambda_{\min}^{-1}(\mathbf{Z}_1^{(0)^\top}\mathbf{Q}\mathbf{Z}_1^{(0)})|$ in (S9c),

$$|\lambda_{min}^{-1}\left(\mathbf{U}_{11}^{(0)}\right)| \leqslant \frac{1}{\xi}\left[\min\left(\min_{l\in 1,2,\cdots,g^{(k)}-1}|\ddot{\mathcal{C}}_l^{(1)}|, |\ddot{\mathcal{C}}_{g^{(k)}}^{(1)}| + |\ddot{\mathcal{C}}_{g^{(k)}+1}^{(1)}|\right)\right]^{-1} \tag{S17a}$$

Next, we use Weyl's inequality Horn et al. (1994). Let $\mathbf{A}, \mathbf{B}, \mathbf{C}$ be any $j \times j$ Hermitian matrices, with eigenvalues $\lambda_{\max}(\mathbf{A}) = a_{(1)} \geqslant a_{(2)} \geqslant \cdots \geqslant a_{(j)} = \lambda_{\min}(\mathbf{A})$; $\lambda_{\max}(\mathbf{B}) = b_{(1)} \geqslant b_{(2)} \geqslant \cdots \geqslant b_{(j)} = \lambda_{\min}(\mathbf{B})$; and $\lambda_{\max}(\mathbf{C}) = c_{(1)} \geqslant c_{(2)} \geqslant \cdots \geqslant c_{(j)} = \lambda_{\min}(\mathbf{C})$ respectively, with $\mathbf{A} = \mathbf{B} + \mathbf{C}$. Then, we have:

$$b_{(j_0)} + c_{(j)} \leqslant a_{(j_0)} \leqslant b_{(j_0)} + c_{(1)}; \forall j_0 = 1, 2, \cdots, j.$$

Specifically, with $j_0 = j$, we have,

$$\lambda_{\min}(\mathbf{A} - \mathbf{B}) \geqslant \lambda_{\min}(\mathbf{A}) + \lambda_{\min}(-\mathbf{B}) = \lambda_{\min}(\mathbf{A}) - \lambda_{\max}(\mathbf{B}) \tag{S17b}$$

Applying the aforementioned inequality with $\mathbf{A} = \mathbf{U}_{11}$ and $\mathbf{B} = \mathbf{U}_{12}\mathbf{U}_{22}^{-1}\mathbf{U}_{12}^\top$, we have

$$\lambda_{min}\left(\mathbf{U}_{11} - \mathbf{U}_{12}\mathbf{U}_{22}^{-1}\mathbf{U}_{12}^\top\right) \geqslant \lambda_{min}(\mathbf{U}_{11}) - \lambda_{max}(\mathbf{U}_{12}\mathbf{U}_{22}^{-1}\mathbf{U}_{12}^\top)$$

Using (S9c) as before, we can derive

$$|\lambda_{min}^{-1}(\mathbf{U}_{11})| \leqslant \frac{1}{\xi}\min\left(\min_{l\in 1,2,\cdots,g^{(k)}-1}|\ddot{\mathcal{C}}_l^{(1)}|, \mathbf{1}^\top\widetilde{\mathbf{Z}}_1^L\right)^{-1} = \frac{1}{\xi}\left(\min_{l\in 1,2,\cdots,g^{(k)}}|\ddot{\mathcal{C}}_l^{(1)}|\right)^{-1} \tag{S17c}$$

where the last equality follows from letting, without loss of generality, $\widetilde{\mathbf{Z}}_1^L$ and $\widetilde{\mathbf{Z}}_1^R$ to be respectively the $(g^{(k)})^{th}$ and $(g^{(k)}+1)^{th}$ column of $\mathbf{Z}_1$.

Using $\mathbf{U}_{22}$ is 1×1 and $\mathbf{U}_{12}$ is $g^{(k)} \times 1$,

$$\lambda_{max}(\mathbf{U}_{12}\mathbf{U}_{22}^{-1}\mathbf{U}_{12}^\top) = \frac{\left(\tilde{\mathbf{Z}}_1^{R^\top}\mathbf{Q}\tilde{\mathbf{Z}}_1\right)\left(\tilde{\mathbf{Z}}_1^{R^\top}\mathbf{Q}\tilde{\mathbf{Z}}_1\right)^\top}{\tilde{\mathbf{Z}}_1^{R^\top}\mathbf{Q}\tilde{\mathbf{Z}}_1^R}$$

Using Assumption 2, we have

$$
\begin{aligned}
&\left(\tilde{\mathbf{Z}}_1^{R^\top}\mathbf{Q}\tilde{\mathbf{Z}}_1\right)\left(\tilde{\mathbf{Z}}_1^{R^\top}\mathbf{Q}\tilde{\mathbf{Z}}_1\right)^\top \\
&= \sum_{l=1}^{g^{(k)}}\left(\tilde{\mathbf{Z}}_1^{R^\top}\mathbf{Q}\tilde{\mathbf{Z}}_{1_{\cdot,l}}\right)^2 \\
&= \sum_{l=1}^{g^{(k)}}\Bigg(\alpha\sum_i \mathbf{Z}_{\mathbf{1}i,g^{(k)}+1}\mathbf{Z}_{\mathbf{1}i,l} + \sum_{j\neq j'=0}^{q}\rho_j\rho_{j'}\sum_i \mathbf{Z}_{\mathbf{1}i-j,g^{(k)}+1}\mathbf{Z}_{\mathbf{1}i-j',l} \\
&\qquad + \sum_{i\in\tilde{\mathcal{A}}_1}\sum_{i'\in\tilde{\mathcal{A}}_2}\tilde{\gamma}_{i,i'}\mathbf{Z}_{\mathbf{1}i,g^{(k)}+1}\mathbf{Z}_{\mathbf{1}i',l}\Bigg)^2 \\
&= \sum_{l=1}^{g^{(k)}}\left(\sum_{j\neq j'=0}^{q}\rho_j\rho_{j'}\sum_i \mathbf{Z}_{\mathbf{1}i-j,g^{(k)}+1}\mathbf{Z}_{\mathbf{1}i-j',l} + \sum_{i\in\tilde{\mathcal{A}}_1}\sum_{i'\in\tilde{\mathcal{A}}_2}\tilde{\gamma}_{i,i'}\mathbf{Z}_{\mathbf{1}i,g^{(k)}+1}\mathbf{Z}_{\mathbf{1}i',l}\right)^2 ;[\ l < g^{(k)}+1] \\
&\leqslant \sum_{l=1}^{g^{(k)}}\left(\sum_{j\neq j'=0}^{q}|\rho_j\rho_{j'}|\sum_i \mathbf{Z}_{\mathbf{1}i-j,g^{(k)}+1}\mathbf{Z}_{\mathbf{1}i-j',l} + \sum_{i\in\tilde{\mathcal{A}}_1}\sum_{i'\in\tilde{\mathcal{A}}_2}|\tilde{\gamma}_{i,i'}|\mathbf{Z}_{\mathbf{1}i,g^{(k)}+1}\mathbf{Z}_{\mathbf{1}i',l}\right)^2 \\
&\leqslant \left(\sum_{j\neq j'=0}^{q}|\rho_j\rho_{j'}|\sum_i \mathbf{Z}_{\mathbf{1}i-j,g^{(k)}+1}\sum_{l=1}^{g^{(k)}}\mathbf{Z}_{\mathbf{1}i-j',l} + \sum_{i\in\tilde{\mathcal{A}}_1}\sum_{i'\in\tilde{\mathcal{A}}_2}|\tilde{\gamma}_{i,i'}|\mathbf{Z}_{\mathbf{1}i,g^{(k)}+1}\sum_{l=1}^{g^{(k)}}\mathbf{Z}_{\mathbf{1}i',l}\right)^2 \\
&\leqslant \left(\sum_{j\neq j'=0}^{q}|\rho_j\rho_{j'}|\sum_i \mathbf{Z}_{\mathbf{1}i-j,g^{(k)}+1} + \sum_{i\in\tilde{\mathcal{A}}_1}\sum_{i'\in\tilde{\mathcal{A}}_2}|\tilde{\gamma}_{i,i'}|\mathbf{Z}_{\mathbf{1}i,g^{(k)}+1}\right)^2
\end{aligned}
$$

The last expression above can be written as $(\sum_i c_i\mathbf{Z}_{i,g^{(k)}+1})^2$ where the constants $c_i \geq 0$ can only take values from a set of positive values free of n and of $O(q^2)$ cardinality. Hence, replacing c_i's with their maximum, we have $\left(\tilde{\mathbf{Z}}_1^{R^\top}\mathbf{Q}\tilde{\mathbf{Z}}_1\right)\left(\tilde{\mathbf{Z}}_1^{R^\top}\mathbf{Q}\tilde{\mathbf{Z}}_1\right)^\top \leq c(|\ddot{\mathcal{C}}^{(1)}_{g^{(k)}+1}|^2)$ for some constant c.

Similarly, we can prove,

$$\tilde{\mathbf{Z}}_1^{R^\top}\mathbf{Q}\tilde{\mathbf{Z}}_1^R \geqslant \xi\left(|\ddot{\mathcal{C}}^{(1)}_{g^{(k)}+1}|\right)$$

Combining, we have

$$\lambda_{max}(\mathbf{U}_{12}\mathbf{U}_{22}^{-1}\mathbf{U}_{12}^\top) \leq \frac{c}{\xi}(|\ddot{\mathcal{C}}^{(1)}_{g^{(k)}+1}|) \tag{S17d}$$

From Eqs. (S17c) and (S17d), we have

$$\lambda_{min}\left(\mathbf{U}_{11}-\mathbf{U}_{12}\mathbf{U}_{22}^{-1}\mathbf{U}_{12}^{\top}\right)\geqslant\xi\min_{l\in 1,2,\cdots,g^{(k)}}|\ddot{\mathcal{C}}_{l}^{(1)}|-\frac{c}{\xi}|\ddot{\mathcal{C}}_{g^{(k)}+1}^{(1)}|. \tag{S17e}$$

Next, to control $\|\mathbf{U}_{11}-\mathbf{U}_{12}\mathbf{U}_{22}^{-1}\mathbf{U}_{12}^{\top}-\mathbf{U}_{11}^{(0)}\|_2$, let $\mathbf{A}=\mathbf{U}_{11}-\mathbf{U}_{11}^{(0)}$ and $\mathbf{B}=\mathbf{U}_{12}\mathbf{U}_{22}^{-1}\mathbf{U}_{12}^{\top}$. We note that for any symmetric matrix $\mathbf{A}-\mathbf{B}$, $\|\mathbf{A}-\mathbf{B}\|_2=\max\{\lambda_{\max}(\mathbf{A}-\mathbf{B}),-\lambda_{\min}(\mathbf{A}-\mathbf{B})\}$. By Weyl's inequality Horn et al. (1994), with $j_0=1$, we have

$$\lambda_{\max}(\mathbf{A}-\mathbf{B})\leqslant\lambda_{\max}(\mathbf{A})+\lambda_{\max}(-\mathbf{B})=\lambda_{\max}(\mathbf{A})-\lambda_{\min}(\mathbf{B})$$

Since $\mathbf{U}_{12}\mathbf{U}_{22}^{-1}\mathbf{U}_{12}^{\top}$ is rank-deficient, $\lambda_{\min}(\mathbf{B})=0$, and we have

$$\lambda_{max}\left(\mathbf{U}_{11}-\mathbf{U}_{12}\mathbf{U}_{22}^{-1}\mathbf{U}_{12}^{\top}-\mathbf{U}_{11}^{(0)}\right)\leqslant\lambda_{max}\left(\mathbf{U}_{11}-\mathbf{U}_{11}^{(0)}\right).$$

$$\begin{aligned}
&\mathbf{U}_{11}-\mathbf{U}_{11}^{(0)}\\
&=\tilde{\mathbf{Z}}_1^{\top}\mathbf{Q}\tilde{\mathbf{Z}}_1-{\mathbf{Z}_1^{(0)}}^{\top}\mathbf{Q}\mathbf{Z}_1^{(0)}\\
&=\tilde{\mathbf{Z}}_1^{\top}\mathbf{Q}\tilde{\mathbf{Z}}_1-\left(\tilde{\mathbf{Z}}_1+\begin{bmatrix}\mathbf{0}_{n\times g^{(k)}-1} & \tilde{\mathbf{Z}}_1^R\end{bmatrix}\right)^{\top}\mathbf{Q}\left(\tilde{\mathbf{Z}}_1+\begin{bmatrix}\mathbf{0}_{n\times g^{(k)}-1} & \tilde{\mathbf{Z}}_1^R\end{bmatrix}\right)\\
&=-\tilde{\mathbf{Z}}_1^{\top}\mathbf{Q}\begin{bmatrix}\mathbf{0}_{n\times g^{(k)}-1} & \tilde{\mathbf{Z}}_1^R\end{bmatrix}-\begin{bmatrix}\mathbf{0}_{n\times g^{(k)}-1} & \tilde{\mathbf{Z}}_1^R\end{bmatrix}^{\top}\mathbf{Q}\tilde{\mathbf{Z}}_1\\
&\qquad\qquad-\begin{bmatrix}\mathbf{0}_{n\times g^{(k)}-1} & \tilde{\mathbf{Z}}_1^R\end{bmatrix}^{\top}\mathbf{Q}\begin{bmatrix}\mathbf{0}_{n\times g^{(k)}-1} & \tilde{\mathbf{Z}}_1^R\end{bmatrix}\\
&=-\begin{bmatrix}\mathbf{0}_{g^{(k)}\times g^{(k)}-1} & \tilde{\mathbf{Z}}_1^{\top}\mathbf{Q}\tilde{\mathbf{Z}}_1^R\end{bmatrix}-\begin{bmatrix}\mathbf{0}_{g^{(k)}\times g^{(k)}-1} & \tilde{\mathbf{Z}}_1^{\top}\mathbf{Q}\tilde{\mathbf{Z}}_1^R\end{bmatrix}^{\top}-\begin{bmatrix}\mathbf{0}_{g^{(k)}-1\times g^{(k)}-1} & \mathbf{0}_{g^{(k)}-1\times 1}\\ \mathbf{0}_{1\times g^{(k)}-1} & \tilde{\mathbf{Z}}_1^{R^{\top}}\mathbf{Q}\tilde{\mathbf{Z}}_1^R\end{bmatrix}
\end{aligned}$$

Since $\left(\mathbf{U}_{11}-\mathbf{U}_{11}^{(0)}\right)$ is symmetric and only has a non-zero last row and column, by Gershgorin circle theorem, all its eigen-values lie in $[-u,u]$ where u is the sum of absolute values of its last row, i.e., $u\leqslant\|\tilde{\mathbf{Z}}_1^{R^{\top}}\mathbf{Q}\tilde{\mathbf{Z}}_1\|_1+\tilde{\mathbf{Z}}_1^{R^{\top}}\mathbf{Q}\tilde{\mathbf{Z}}_1^R$. Similar to the derivation of (S17d), we can establish that each of the terms in u are of the order $|\ddot{\mathcal{C}}_{g^{(k)}+1}^{(1)}|$. Hence,

$$\lambda_{max}\left(\mathbf{U}_{11}-\mathbf{U}_{12}\mathbf{U}_{22}^{-1}\mathbf{U}_{12}^{\top}-\mathbf{U}_{11}^{(0)}\right)\leqslant C_9\left(|\ddot{\mathcal{C}}_{g^{(k)}+1}^{(1)}|\right) \tag{S17f}$$

Similarly, by Weyl's inequality and applying (S17e) we have

$$\begin{aligned}\lambda_{min}\left(\mathbf{U}_{11}-\mathbf{U}_{12}\mathbf{U}_{22}^{-1}\mathbf{U}_{12}^{\top}-\mathbf{U}_{11}^{(0)}\right) &\geqslant -\lambda_{\max}(\mathbf{U}_{12}\mathbf{U}_{22}^{-1}\mathbf{U}_{12}^{\top})+\lambda_{\min}\left(\mathbf{U}_{11}-\mathbf{U}_{11}^{(0)}\right)\\ &\geqslant -C_{10}\left(|\ddot{\mathcal{C}}^{(1)}_{g^{(k)}+1}|\right)\end{aligned} \tag{S17g}$$

As proved in the case of J_2, for large n, we have on Ω_n,

$$\frac{\mathbf{Y}^{\top}\mathbf{Q}\tilde{\mathbf{Z}}_1(\mathbf{Y}^{\top}\mathbf{Q}\tilde{\mathbf{Z}}_1)^{\top}}{n^2}\leqslant C_8 \tag{S17h}$$

Combining Eqs. (S17a) and (S17e) to (S17h) for large n, we have

$$\begin{aligned}J_{11} &\leqslant \frac{\mathbf{Y}^{\top}\mathbf{Q}\tilde{\mathbf{Z}}_1(\mathbf{Y}^{\top}\mathbf{Q}\tilde{\mathbf{Z}}_1)^{\top}}{n^2}\frac{\|\left(\mathbf{U}_{11}-\mathbf{U}_{12}\mathbf{U}_{22}^{-1}\mathbf{U}_{12}^{\top}-\mathbf{U}_{11}^{(0)}\right)\|_2/n}{(|\lambda_{min}\left(\mathbf{U}_{11}-\mathbf{U}_{12}\mathbf{U}_{22}^{-1}\mathbf{U}_{12}^{\top}\right)|/n)|\lambda_{min}\left(\mathbf{U}_{11}^{(0)}\right)/n|}\\ &= C_{11}\frac{|\ddot{\mathcal{C}}^{(1)}_{g^{(k)}+1}|/n}{(\xi\min_{l\in 1,2,\cdots,g^{(k)}}|\ddot{\mathcal{C}}^{(1)}_{l}|/n-\frac{c}{\xi}|\ddot{\mathcal{C}}^{(1)}_{g^{(k)}+1}|/n)}\times\\ &\qquad\frac{1}{\left[\min\left(\min_{l\in 1,2,\cdots,g^{(k)}-1}|\ddot{\mathcal{C}}^{(1)}_{l}|,|\ddot{\mathcal{C}}^{(1)}_{g^{(k)}}|+|\ddot{\mathcal{C}}^{(1)}_{g^{(k)}+1}|\right)\right]/n}\\ &\leqslant C_{11}\frac{(\sqrt{\delta}+\delta^2)}{(\xi(\varepsilon^D(1-\sqrt{\delta})-\delta^2)-\frac{c}{\xi}(\sqrt{\delta}+\delta^2)}\frac{1}{\varepsilon^D-\delta^2}\end{aligned}$$

The aforementioned inequality used the following:

1. $Vol(|\ddot{\mathcal{C}}^{(1)}_{l}|)\geqslant\varepsilon^D;\forall l=1,\cdots,g^{(k)}-1;$
2. $Vol(|\ddot{\mathcal{C}}^{(1)}_{g^{(k)}}|)\geqslant\varepsilon^D(1-\sqrt{\delta});$
3. $Vol(|\ddot{\mathcal{C}}^{(1)}_{g^{(k)}+1}|)\leqslant\sqrt{\delta}$
4. $Vol(|\ddot{\mathcal{C}}^{(1)}_{g^{(k)}+1}|)+Vol(|\ddot{\mathcal{C}}^{(1)}_{g^{(k)}}|)\geqslant\varepsilon^D$

Hence J_{11} converges to 0 in probability uniformly as $\delta\downarrow 0$. Next,

$$
\begin{aligned}
J_{13} &= \frac{1}{n}\left|\begin{bmatrix}\mathbf{0}_{1\times g^{(k)}} & \mathbf{Y}^\top\mathbf{Q}\tilde{\mathbf{Z}}_1^R\end{bmatrix}\left(\begin{bmatrix}\mathbf{U}_{11} & \mathbf{U}_{12};\\ \mathbf{U}_{12}^\top & \mathbf{U}_{22}\end{bmatrix}^{-1} - \begin{bmatrix}\mathbf{U}_{11}^{(0)^{-1}} & 0\\ 0 & 0\end{bmatrix}\right)\begin{bmatrix}\mathbf{0}_{1\times g^{(k)}} & \mathbf{Y}^\top\mathbf{Q}\tilde{\mathbf{Z}}_1^R\end{bmatrix}^\top\right| \\
&= \frac{1}{n}\left|\begin{bmatrix}\mathbf{0}_{1\times g^{(k)}} & \mathbf{Y}^\top\mathbf{Q}\tilde{\mathbf{Z}}_1^R\end{bmatrix}\begin{bmatrix}\mathbf{U}_{11} & \mathbf{U}_{12};\\ \mathbf{U}_{12}^\top & \mathbf{U}_{22}\end{bmatrix}^{-1}\begin{bmatrix}\mathbf{0}_{1\times g^{(k)}} & \mathbf{Y}^\top\mathbf{Q}\tilde{\mathbf{Z}}_1^R\end{bmatrix}^\top\right| \\
&= \frac{1}{n}\left|\begin{bmatrix}\mathbf{0}_{1\times g^{(k)}} & \mathbf{Y}^\top\mathbf{Q}\tilde{\mathbf{Z}}_1^R\end{bmatrix}\left[\begin{bmatrix}\tilde{\mathbf{Z}}_1 & \tilde{\mathbf{Z}}_1^R\end{bmatrix}^\top\mathbf{Q}\begin{bmatrix}\tilde{\mathbf{Z}}_1 & \tilde{\mathbf{Z}}_1^R\end{bmatrix}\right]^{-1}\begin{bmatrix}\mathbf{0}_{1\times g^{(k)}} & \mathbf{Y}^\top\mathbf{Q}\tilde{\mathbf{Z}}_1^R\end{bmatrix}^\top\right| \\
&\leqslant \frac{1}{n}\lambda_{\max}\left[\begin{bmatrix}\tilde{\mathbf{Z}}_1 & \tilde{\mathbf{Z}}_1^R\end{bmatrix}^\top\mathbf{Q}\begin{bmatrix}\tilde{\mathbf{Z}}_1 & \tilde{\mathbf{Z}}_1^R\end{bmatrix}\right]^{-1}\|(\mathbf{0}_{1\times g^{(k)}}, \mathbf{Y}^\top\mathbf{Q}\tilde{\mathbf{Z}}_1^R)\|_2^2 \\
&\leqslant \frac{\mathbf{Y}^\top\mathbf{Q}\tilde{\mathbf{Z}}_1^R(\mathbf{Y}^\top\mathbf{Q}\tilde{\mathbf{Z}}_1^R)^\top}{n^2}\frac{n}{\xi}\left[\min_{l\in 1,2,\cdots,g^{(k)}+1}|\ddot{\mathcal{C}}_l^{(1)}|\right]^{-1}
\end{aligned}
$$

Here the last inequality follows from (S9c). Additionally, in (S16) we have established for large enough n, we have

$$
\frac{1}{n}|\mathbf{Y}^\top\mathbf{Q}\tilde{\mathbf{Z}}_1^R| \leqslant (\alpha + \sum_{j\neq j'=0}^{q}|\rho_j\rho_{j'}|)M_0(\sqrt{\delta}+\delta^2) + C_7\phi.
$$

Hence, we have the following:

$$
J_{13} \leqslant \left((\alpha + \sum_{j\neq j'=0}^{q}|\rho_j\rho_{j'}|)M_0(\sqrt{\delta}+\delta^2) + C_7\phi\right)^2\frac{1}{\xi}\frac{1}{\sqrt{\delta}-\delta^2}
$$

As ϕ can be chosen arbirtrarily small, this converges to 0 in probability uniformly as $\delta \downarrow 0$. By the Cauchy-Schwarz inequality, $J_{12} \leqslant 2\sqrt{|J_{11}||J_{13}|}$. Hence J_1 (and similarly J_3) converges to 0 in probability uniformly as $\delta \downarrow 0$, completing the proof of the lemma under **R2**.

□

Proof of Proposition A2.2. For $\zeta_n > M_0$,

$$\begin{aligned}
&\lim_{n\to\infty} \mathbb{E}\left[\inf_{f\in T_{\zeta_n}\mathcal{F}_n} \mathbb{E}_X|f(X)-m(X)|^2\right] \\
&\leqslant \lim_{n\to\infty} \mathbb{E}\left[\inf_{f\in T_{\zeta_n}(\{\sum_{l=1}^{t_n} m(\mathbf{x}_l)I(\mathbf{x}\in\mathcal{B}_l)|\mathbf{x}_l\in\mathcal{B}_l,l=1,\ldots,t_n\})} \mathbb{E}_X|f(X)-m(X)|^2\right] \\
&\leqslant \lim_{n\to\infty} \mathbb{E}\left[\mathbb{E}_X\Delta^2(m,\mathcal{B}_n(X,\Theta))\right]
\end{aligned}$$

where the variation of m in node $\mathcal{B}$ is given by $\Delta(m,\mathcal{B}) = \sup_{\mathbf{x}_1,\mathbf{x}_2\in\mathcal{B}}|m(\mathbf{x}_1)-m(\mathbf{x}_2)|$. Hence, it is enough to show that the variation of m in leaf nodes $\mathcal{B}_n$ of GLS-style tree vanishes asymptotically.

Theorem 2.1 shows the existence of a theoretical DART-split criterion. Let $\mathcal{B}_k^*(X,\Theta)$ be the leaf node of a theoretical RF-GLS (i.e. in $(k+1)^{th}$ level) built with randomness Θ and containing X and let the set of optimal theoretical splits to build $\mathcal{B}_k^*(X,\Theta)$ be denoted by $\mathfrak{S}_k^*(X,\Theta)$. Since Theorem 2.1 also shows that the theoretical DART-split criterion is simply a constant multiplier of the theoretical CART-split criterion, under Assumption 5, we immediately have from Lemma 1 of Scornet et al. (2015) that

$$\Delta(m,\mathcal{B}_k^*(X,\Theta)) \xrightarrow{a.s.} 0 \text{ as } k\to\infty. \tag{S18}$$

Next, using (S18) and Proposition A2.1 on stochastic equicontinuity of the empirical DART split criterion, we are now ready to prove that optimal theoretical and empirical splits become identical asymptotically as follows. Let $\hat{\mathbf{s}}_{k,n}(X,\Theta)\in\mathfrak{S}_k$ be the set of empirical optimal splits used to build the node containing X in level $k+1$. For fixed $\varepsilon,\tau>0$; $k\in\mathbb{N}$; $\exists\, n_4\in\mathbb{N}$, such that

$$\mathbb{P}\left[dist\left(\hat{\mathbf{s}}_{k,n}(X,\Theta),\mathfrak{S}_k^*(X,\Theta)\right)\leqslant\varepsilon\right]\geqslant 1-\tau,\ \forall n>n_4. \tag{S19}$$

Equation (S19) is established identical to the proof of the analogous result (Lemma 3) in Scornet et al. (2015). Only change is that for a single split at level k, we now condition on all the splits in levels $1,2,\cdots,k-1$ i.e. $\tilde{\mathbf{s}}_{k-1}\in\tilde{\mathfrak{S}}_{k-1}$ with $\tilde{\mathfrak{S}}_{k-1}^{\varepsilon}$ playing the role of conditioning set.

The result in (S18) lets us control the variation of regression function m in theoretical GLS-style tree leaf nodes, and (S19) establishes that the difference between the theoretical optimal splits and the empirical optimal splits vanishes asymptotically. Combining these two results, we obtain, identical to the proof of Proposition 2 in Scornet et al. (2015) that $\forall\varepsilon,\gamma>0$, $\exists n_5\in\mathbb{N}$ such that

$\forall n > n_5$,

$$\mathbb{P}_{X,\Theta}[\Delta(m, \mathcal{B}_n(X, \Theta)) \leqslant \varepsilon] \geqslant 1 - \gamma.$$

Since Δ is bounded, this ensures $\mathbb{E}\Delta^2 \leq \varepsilon + M_0^2 \mathbb{P}_{X,\Theta}[\Delta(m, \mathcal{B}_n(X, \Theta)) \geqslant \varepsilon]$ and letting $n \to \infty$ and then $\epsilon \to 0$ yields the result. □

S3 Technical Results

Lemma S3.1. Under Assumption 1, we have

1. $\frac{1}{n}\sum_i B_i\epsilon_i \overset{a.s.}{\to} 0; B_i \overset{i.i.d}{\sim} Bernoulli(p); \forall p > 0; B_i \perp\!\!\!\perp \epsilon_i$;
2. $\exists n_1$ such that $\forall n > n_1$, we have $\frac{1}{n}\sum_i \epsilon_i^2 \leqslant C_0$ a.s.

Proof. For any β-mixing process $\{\chi_i\}$, the β-mixing coefficients can be written as

$$\beta_\chi(a) := \sup_t \|\mathbb{P}_{-\infty}^t \otimes \mathbb{P}_{t+a}^\infty - \mathbb{P}_{t,a}\|_{TV}$$

where, $\mathbb{P}_a^b$ is the joint-distribution of $\{\chi_i\}_{\{a<i\leq b\}}$, $\mathbb{P}_{t,a}$ is the joint-distribution of $\{\chi_i\}_{\{i\leq t\}\cup\{i>t+a\}}$, and $\|\cdot\|_{TV}$ is the total variation norm for measures (Definition 2.1 of Mcdonald et al. (2011)). Hence, we can write

$$\begin{aligned}
\beta_{(\epsilon,B)}(a) &:= \sup_t \|\mathbb{P}_{-\infty}^t(\epsilon, B) \otimes \mathbb{P}_{t+a}^\infty(\epsilon, B) - \mathbb{P}_{t,a}(\epsilon, B)\|_{TV} \\
&= \sup_t \|\mathbb{P}_{-\infty}^t(\epsilon) \otimes \mathbb{P}_{-\infty}^t(B) \otimes \mathbb{P}_{t+a}^\infty(\epsilon) \otimes \mathbb{P}_{t+a}^\infty(B) - \mathbb{P}_{t,a}(\epsilon) \otimes \mathbb{P}_{t,a}(B)\|_{TV} \\
&\leq \sup_t \|\mathbb{P}_{-\infty}^t(\epsilon) \otimes \mathbb{P}_{t+a}^\infty(\epsilon) - \mathbb{P}_{t,a}(\epsilon)\|_{TV} \\
&\qquad + \sup_t \|\mathbb{P}_{-\infty}^t(B) \otimes \mathbb{P}_{t+a}^\infty(B) - \mathbb{P}_{t,a}(B)\|_{TV}
\end{aligned}$$

Here the first equality follows from $\{B_i\}$ being independent of $\{\epsilon_i\}$, and the inequality is from Lemma 1 of Eberlein (1984). Hence,

$$\beta_{(\epsilon,B)}(a) \leq \beta_{(\epsilon)}(a) + \beta_{(B)}(a). \tag{S20}$$

This implies that as both $\{B_i\}$ (i.i.d. process) and $\{\epsilon_i\}$ β-mixing, so is $\{\epsilon_i, B_i\}$. Now, using Theorem

1 of Nobel and Dembo (1993) with the singleton class comprising of the function $(B_i, \epsilon_i) \mapsto B_i\epsilon_i$ (as $B_i\epsilon_i \leqslant |\epsilon_i|$ which is uniformly integrable by Assumption 1), we have the strong-law of part 1.

For part 2, as $\{\epsilon_i\}_{i=1}^n$ is stationary and absolutely-regular mixing, and $\mathbb{E}\epsilon_1^2 < \infty$ (by Assumption 1), once again using Theorem 1 of Nobel and Dembo (1993) on $\{\epsilon_i\}_{i=1}^n$ now with the singleton class $\{f(x) = x^2\}$, we have the result. □

Lemma S3.2. Let $X_i \overset{i.i.d}{\sim} X, \{X_i\} \perp\!\!\!\perp \{\epsilon_i\}$. For $\mathcal{C} \in \sigma(X)$, let $B_i^{(\mathcal{C})} = \mathbb{I}(X_i \in \mathcal{C}); i = 1, \cdots, n$. Let $\mathcal{P}$ denote a collection of polynomial-in-n number of such sets $\mathcal{C}$. Then under Assumptions 4(b) and 4(c), for any $\pi, \phi > 0, \exists\ n_2 \in \mathbb{N}^*$ such that with probability $1 - \pi$, $\forall n > n_2$,

$$\cap_{\mathcal{C}\in\mathcal{P}} \left\{ \left| \frac{1}{n} \sum_i \epsilon_i B_i^{(\mathcal{C})} \right| \leqslant \phi \right\} \cap \left\{ \left| \frac{1}{n} \sum_i \epsilon_i^2 \right| \leqslant \sigma_0^2 \right\}.$$

Proof. Let $U_{\mathcal{C}} = \{\left|\frac{1}{n}\sum_i \epsilon_i B_i^{(\mathcal{C})}\right| \geqslant \phi\}$, $V_{\mathcal{C}} = \{\sum_i B_i^{(\mathcal{C})} \geqslant \sqrt{n}\}$, $W = \{\max_i |\epsilon_i| \leqslant C_\pi \sqrt{\log n}\}$. Then, on $V_{\mathcal{C}}^c \cap W$ for all $\mathcal{C} \in \sigma(X)$, we have, for large enough n,

$$\left| \frac{1}{n} \sum_i \epsilon_i B_i^{(\mathcal{C})} \right| \leqslant C_\pi \frac{1}{n} \sqrt{\log n} \sum_i B_i^{(\mathcal{C})} \leqslant C_\pi \frac{1}{\sqrt{n}} \sqrt{\log n} < \phi.$$

Hence, for large enough n, we have $\cup_{\mathcal{C}} \{U_{\mathcal{C}} \cap V_{\mathcal{C}}^c \cap W\} = \{\}$. Using $\mathbb{P}(W) \geq 1 - \pi/4$ for large enough n from Assumption 4(b), we write

$$\begin{aligned}
\mathbb{P}\left(\cup_{\mathcal{C}\in\mathcal{P}} \left\{ \left| \frac{1}{n} \sum_i \epsilon_i B_i^{(\mathcal{C})} \right| \geqslant \phi \right\} \right) &\leqslant \pi/4 + \mathbb{P}\left(\cup_{\mathcal{C}\in\mathcal{P}} U_{\mathcal{C}} \cap W \right) \\
&= \pi/4 + \mathbb{P}\left(\cup_{\mathcal{C}\in\mathcal{P}} U_{\mathcal{C}} \cap V_{\mathcal{C}} \cap W \right) \\
&\leqslant \pi/4 + \mathbb{P}\left(\cup_{\mathcal{C}\in\mathcal{P}} U_{\mathcal{C}} \cap V_{\mathcal{C}} \right).
\end{aligned}$$

We can write $\left|\frac{1}{n}\sum_i \epsilon_i B_i^{(\mathcal{C})}\right| \leq \left|\frac{1}{\sum_i B_i^{(\mathcal{C})}} \sum_i \epsilon_i B_i^{(\mathcal{C})}\right|$. Let $b_i, i = 1, \ldots, n$ denote a realization of $B_i^{(\mathcal{C})}$ for some $\mathcal{C}$. Denoting $\mathcal{I}_n = \{i \in \{1, \ldots, n\} \,|\, b_i > 0\}$ and using independence of $\{X_i\}$'s and $\{\epsilon_i\}$'s, we can write

$$\mathbb{P}\left(\left| \frac{1}{\sum_i B_i^{(\mathcal{C})}} \sum_i \epsilon_i B_i^{(\mathcal{C})} \right| > \phi \right) = \sum_{\{b_i\}} \mathbb{P}\left(\frac{1}{|\mathcal{I}_n|} \sum_{i\in\mathcal{I}_n} \epsilon_i > \phi \right) \mathbb{P}\left(\cap_{i\in\mathcal{I}_n} \{B_i = b_i\} \right)$$

The sub-Gaussian tail of Assumption 4(c) implies that $P(\frac{1}{|\mathcal{I}_n|}|\sum_{i\in\mathcal{I}_n}\epsilon_i| > \phi) \leq C\exp(-c|\mathcal{I}_n|)$ for all choices of the sub-sequence $\mathcal{I}_n$. Hence, we have

$$P(|\frac{1}{\sum_i B_i^{(\mathcal{C})}}\sum_i \epsilon_i B_i^{(\mathcal{C})}| > \phi) \leq C\exp(-c\sqrt{n}) \text{ on } \{\sum_i B_i^{(\mathcal{C})} \geqslant \sqrt{n}\}.$$

This proves $\mathbb{P}(U_{\mathcal{C}} \cap V_{\mathcal{C}}) \leqslant C\exp(-c\sqrt{n})$. As there are a polynomial in n number of $\mathcal{C}$'s, taking union bounds yields the result simultaneously for all $\mathcal{C}$. The adjustment in probability due to the ϵ_i^2 term follows directly from the second tail-bound assumption in 4(c). □

Lemma S3.3. For any matrix $\mathbf{Q}$ and a binary matrix $\mathbf{Z}$ with columns $\mathbf{Z}_{\cdot l}$ and row-sums 1, $\mathbf{Z}_{\cdot l}^\top\mathbf{Q}\mathbf{Z}_{\cdot l} - \sum_{l'\neq l}|\mathbf{Z}_{\cdot l}^\top\mathbf{Q}\mathbf{Z}_{\cdot l'}| \geqslant \sum_i \mathbf{Z}_{il}(\mathbf{Q}_{ii} - \sum_{j\neq i}|\mathbf{Q}_{ij}|)$.

Proof. Using the statement of the Lemma, $\mathbf{Z}_{il}\mathbf{Z}_{il'} = \mathbf{Z}_{il}\mathbb{I}(l = l')$. Hence,

$$\begin{aligned}
\mathbf{Z}_{\cdot l}^\top\mathbf{Q}\mathbf{Z}_{\cdot l} - \sum_{l'\neq l}|\mathbf{Z}_{\cdot l}^\top\mathbf{Q}\mathbf{Z}_{\cdot l'}| &= \sum_i \mathbf{Q}_{ii}\mathbf{Z}_{il} + \sum_{j\neq i}\mathbf{Q}_{ij}\mathbf{Z}_{il}\mathbf{Z}_{jl} - \sum_{l\neq l'}|\sum_{j\neq i}\mathbf{Q}_{ij}\mathbf{Z}_{il}\mathbf{Z}_{jl'}| \\
&\geqslant \sum_i \mathbf{Q}_{ii}\mathbf{Z}_{il} + \sum_{j\neq i}\mathbf{Q}_{ij}\mathbf{Z}_{il}\mathbf{Z}_{jl} - \sum_{l\neq l'}\sum_{j\neq i}|\mathbf{Q}_{ij}|\mathbf{Z}_{il}\mathbf{Z}_{jl'} \\
&\geqslant \sum_i \mathbf{Q}_{ii}\mathbf{Z}_{il} + \sum_{j\neq i}\mathbf{Q}_{ij}\mathbf{Z}_{il}\mathbf{Z}_{jl} - \sum_{j\neq i}|\mathbf{Q}_{ij}|\mathbf{Z}_{il}(1-\mathbf{Z}_{jl}) \\
&\geqslant \sum_i \mathbf{Q}_{ii}\mathbf{Z}_{il} - \sum_{j\neq i}|\mathbf{Q}_{ij}|\mathbf{Z}_{il}\mathbf{Z}_{jl} - \sum_{j\neq i}|\mathbf{Q}_{ij}|\mathbf{Z}_{il}(1-\mathbf{Z}_{jl}) \\
&= \sum_i \mathbf{Q}_{ii}\mathbf{Z}_{il} - \sum_{j\neq i}|\mathbf{Q}_{ij}|\mathbf{Z}_{il}
\end{aligned}$$

□

Appendix & Supplementary Material References

Abramowitz, M. and Stegun, I. A. (1948). *Handbook of mathematical functions with formulas, graphs, and mathematical tables*, volume 55. US Government printing office.

Basu, S., Michailidis, G., et al. (2015). Regularized estimation in sparse high-dimensional time series models. *The Annals of Statistics* **43,** 1535–1567.

Bradley, R. C. (2005). Basic properties of strong mixing conditions. a survey and some open questions. *arXiv preprint math/0511078* .

Breiman, L. (2001). Random forests. *Machine learning* **45,** 5–32.

Carrasco, M. and Chen, X. (2002). Mixing and moment properties of various garch and stochastic volatility models. *Econometric Theory* pages 17–39.

Chen, Y. M., Chen, X. S., and Li, W. (2016). On perturbation bounds for orthogonal projections. *Numerical Algorithms* **73,** 433–444.

Datta, A., Banerjee, S., Finley, A. O., and Gelfand, A. E. (2016). Hierarchical Nearest-Neighbor Gaussian Process Models for Large Geostatistical Datasets. *Journal of the American Statistical Association* **111,** 800–812.

Dehling, H. and Philipp, W. (2002). Empirical process techniques for dependent data. In *Empirical process techniques for dependent data*, pages 3–113. Springer.

Doukhan, P. (2012). *Mixing: properties and examples*, volume 85. Springer Science & Business Media.

Du, J., Zhang, H., Mandrekar, V., et al. (2009). Fixed-domain asymptotic properties of tapered maximum likelihood estimators. *the Annals of Statistics* **37,** 3330–3361.

Eberlein, E. (1984). Weak convergence of partial sums of absolutely regular sequences. *Statistics & probability letters* **2,** 291–293.

Finley, A. O., Datta, A., Cook, B. D., Morton, D. C., Andersen, H. E., and Banerjee, S. (2019). Efficient algorithms for bayesian nearest neighbor gaussian processes. *Journal of Computational and Graphical Statistics* **28,** 401–414.

Friedman, J., Hastie, T., and Tibshirani, R. (2001). *The elements of statistical learning*, volume 1. Springer series in statistics New York.

Friedman, J. H. (2006). Recent advances in predictive (machine) learning. *Journal of classification* **23,** 175–197.

Györfi, L., Kohler, M., Krzyzak, A., and Walk, H. (2006). *A distribution-free theory of nonparametric regression*. Springer Science & Business Media.

Hartikainen, J. and Särkkä, S. (2010). Kalman filtering and smoothing solutions to temporal gaussian process regression models. In *2010 IEEE international workshop on machine learning for signal processing*, pages 379–384. IEEE.

Hoeffding, W., Robbins, H., et al. (1948). The central limit theorem for dependent random variables. *Duke Mathematical Journal* **15,** 773–780.

Horn, R. A., Horn, R. A., and Johnson, C. R. (1994). *Topics in matrix analysis*. Cambridge university press.

Ihara, S. (1993). *Information theory for continuous systems*, volume 2. World Scientific.

Loeve, M. (1977). Elementary probability theory. In *Probability theory i*, pages 1–52. Springer.

Mcdonald, D., Shalizi, C., and Schervish, M. (2011). Estimating beta-mixing coefficients. In *Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics*, pages 516–524.

Mokkadem, A. (1988). Mixing properties of arma processes. *Stochastic processes and their applications* **29,** 309–315.

Nobel, A. and Dembo, A. (1993). A note on uniform laws of averages for dependent processes. *Statistics & Probability Letters* **17,** 169–172.

Nobel, A. et al. (1996). Histogram regression estimation using data-dependent partitions. *The Annals of Statistics* **24,** 1084–1105.

Peligrad, M. (2001). A note on the uniform laws for dependent processes via coupling. *Journal of Theoretical Probability* **14,** 979–988.

Rasmussen, C. E. (2003). Gaussian processes in machine learning. In *Summer School on Machine Learning*, pages 63–71. Springer.

Scornet, E., Biau, G., Vert, J.-P., et al. (2015). Consistency of random forests. *The Annals of Statistics* **43,** 1716–1741.

Software, M. O. C. Sub-gaussian random variables. *https://ocw.mit.edu/courses/mathematics/18-s997-high-dimensional-statistics-spring-2015/lecture-notes/MIT18_S997S15_Chapter1.pdf* .

Stein, M. L. (2012). *Interpolation of spatial data: some theory for kriging.* Springer Science & Business Media.

Stein, M. L. et al. (2002). The screening effect in kriging. *The Annals of Statistics* **30,** 298–323.

Vershynin, R. (2010). Introduction to the non-asymptotic analysis of random matrices. *arXiv preprint arXiv:1011.3027* .

Zhang, H. (2004). Inconsistent estimation and asymptotically equal interpolations in model-based geostatistics. *Journal of the American Statistical Association* **99,** 250–261.